January 1999

A Computational Analysis of Information Structure Using Parallel Expository Texts in English and Japanese

Nobo N. Komagata
University of Pennsylvania

University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-99-07.

A Computational Analysis of Information Structure Using Parallel Expository Texts in English and Japanese

Abstract
This thesis concerns the notion of 'information structure': informally, organization of information in an utterance with respect to the context. Information structure has been recognized as a critical element in a number of computer applications: e.g., selection of contextually appropriate forms in machine translation and speech generation, and analysis of text readability in computer-assisted writing systems.

One of the problems involved in these applications is how to identify information structure in extended texts. This problem is often ignored, assumed to be trivial, or reduced to a sub-problem that does not correspond to the complexity of realistic texts. A handful of computational proposals face the problem directly, but they are generally limited in coverage and all suffer from lack of evaluation. To fully demonstrate the usefulness of information structure, it is essential to apply a theory of information structure to the identification problem and to provide an evaluation method.

This thesis adopts a classic theory of information structure as binomial partition between theme and rheme, and captures the property of theme as a requirement of the contextual-link status. The notion of 'contextual link' is further specified in terms of discourse status, domain-specific knowledge, and linguistic marking. The relation between theme and rheme is identified as the semantic composition of the two, and linked to surface syntactic structure using Combinatory Categorial Grammar. The identification process can then be specified as analysis of contextual link status along the linguistic structure.

The implemented system identifies information structure in real texts in English. Building on the analysis of Japanese presented in the thesis, the system automatically predicts contextually appropriate use of certain particles in the corresponding texts in Japanese. The machine prediction is then compared with human translations. The evaluation results demonstrate that the prediction of the theory is an improvement over alternative hypotheses. We then conclude that information structure can in fact be used to improve the quality of computational applications in practical settings.

Comments
University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-99-07.

A COMPUTATIONAL ANALYSIS OF INFORMATION STRUCTURE USING PARALLEL EXPOSITORY TEXTS IN ENGLISH AND JAPANESE

Nobo N. Komagata

A DISSERTATION

in

Computer and Information Science

Presented to the Faculties of the University of Pennsylvania

in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

1999

Dr. Mark J. Steedman
Supervisor of Dissertation

Dr. Jean Gallier
Graduate Group Chair

Acknowledgments

I would like to thank my advisor, Mark Steedman, for his valuable advice, warm encouragement, and unfailing patience throughout my years at Penn. His extremely broad interests and deep insight have always guided my work on this thesis, especially at difficult times. Among many things I learned from Mark, I particularly appreciate his points that I needed to deliver results visibly and demonstrate the generality of an idea.

I am deeply grateful to my thesis committee members. Claire Gardent made critical comments that were difficult to respond to but gave me an opportunity to look at the thesis from a different point of view. Aravind Joshi inspired me on a wide range of issues, from formal grammars to discourse analysis. Martha Palmer read a draft of this thesis with great care and made numerous helpful points about the contents. Bonnie Webber gave me an opportunity to assist in her enjoyable AI course for three semesters. Although Ellen Prince is not on the thesis committee, I thank her for serving on the proposal committee and teaching me linguistic pragmatics.

The superb academic environment at Penn is one of the factors I cannot forget. I learned a great many basic concepts from the coursework in computer science and linguistics. I thank my professors, especially Peter Buneman, Robin Clark, Tony Kroch, Mitch Marcus, and Max Mintz. I would like to thank Mike Felker for helping me at every stage, from application to graduation. In addition, I owe much to the people who made possible the financial support I received in the form of a Dean's Fellowship, CIS departmental grants, and an IRCS Graduate Fellowship.

I am grateful for my colleagues at Penn. In particular, I had valuable comments from the members of Mark Steedman's research group, especially Beryl Hoffman, Charlie Ortiz, Jong Park, Scott Prevost, Matthew Stone, and Mike White. At various stages, I also received helpful comments from Jason Baldridge, Susan Converse, Miriam Eckert, Jason Eisner, Chung-hye Han, Gehard Jäger,

Seth Kulick, Rashmi Prasad, Anoop Sarkar, Bangalore Srinivas, and Michael Strube. Thanks also to Mimi Lipson for her help in designing a linguistic experiment. In addition, discussion with Penn graduates and visitors was a great help; many thanks to Sadao Kurohashi, Kathy McKeown, K. Vijay-Shanker, and David Weir.

My interest in computational linguistics began with my undergraduate project. I thank Tetsuya Okamoto for introducing me to the field and supervising the project. Later, I had a wonderful opportunity to be involved in a machine-translation project at Bravice with Naoya Arakawa, Akira Nagasawa, and Jun Ohta. Special thanks to Akira Kurahone, who introduced me to Categorial Grammar, the grammatical framework used in this thesis.

I would like to thank a very special couple, my mentors, Josephine and José Rabinowitz, for their encouragement, support, and the best Mexican food in town right from the next door.

Finally, I am very fortunate to have had the understanding and patience of my family from distant Tokyo, Sydney, and Québec. I thank my parents, Ichiko and Eiichi, for the way they brought me up and what they have given to me. But most importantly, my hearty thanks to my wife, Sachiko, for everything we have shared since 1983. While her expertise with medical terminology and statistics, especially the kappa statistic, was an invaluable help, the single most important thing for me has been and always will be her smile.

Abstract

A COMPUTATIONAL ANALYSIS OF INFORMATION STRUCTURE
USING PARALLEL EXPOSITORY TEXTS
IN ENGLISH AND JAPANESE

Nobo N. Komagata

Supervisor: Dr. Mark J. Steedman

This thesis concerns the notion of 'information structure': informally, organization of information in an utterance with respect to the context. Information structure has been recognized as a critical element in a number of computer applications: e.g., selection of contextually appropriate forms in machine translation and speech generation, and analysis of text readability in computer-assisted writing systems.

One of the problems involved in these applications is how to identify information structure in extended texts. This problem is often ignored, assumed to be trivial, or reduced to a sub-problem that does not correspond to the complexity of realistic texts. A handful of computational proposals face the problem directly, but they are generally limited in coverage and all suffer from lack of evaluation. To fully demonstrate the usefulness of information structure, it is essential to apply a theory of information structure to the identification problem and to provide an evaluation method.

This thesis adopts a classic theory of information structure as binomial partition between theme and rheme, and captures the property of theme as a requirement of the contextual-link status. The notion of 'contextual link' is further specified in terms of discourse status, domain-specific knowledge, and linguistic marking. The relation between theme and rheme is identified as the semantic composition of the two, and linked to surface syntactic structure using Combinatory

Categorial Grammar. The identification process can then be specified as analysis of contextual-link status along the linguistic structure.

The implemented system identifies information structure in real texts in English. Building on the analysis of Japanese presented in the thesis, the system automatically predicts contextually-appropriate use of certain particles in the corresponding texts in Japanese. The machine prediction is then compared with human translations. The evaluation results demonstrate that the prediction of the theory is an improvement over alternative hypotheses. We then conclude that information structure can in fact be used to improve the quality of computational applications in practical settings.

Contents

List of Tables

List of Figures

Notational Conventions

- *Italic*: (1) Cited word, (2) Grammatical subject (in examples) [p. 2], (3) Utterance number in Roman numerals (in discourse examples) [p. 1]
- *Math font* (appears very similar to *italic*): (1) Semantic representation [p. 81], (2) CCG category [p. 95]
- **Boldface**: (1) Technical term with definition/description, (2) Phonological prominence (in examples) [p. 32]
- Sans serif: Category variable (for type raising) [p. 103]
- SMALL CAPS: (1) Terms for referential status from Prince [1981] [p. 28], (2) Grammatical labels (Japanese)
- `Typewriter font`: Computer source code, output, or data
- Underline: (1) Attention to an element in examples (not a phonological/pragmatic feature), (2) Cancelled categories in CCG derivations [p. 96]
- Single quotes ‘ ’: (1) Technical terms in general, (2) Special character and short symbol
- Double quotes “ ”: Cited expression (more than one word)
- Parentheses ( ): (1) Argument of functional application [p. 82], (2) Presupposition (in gloss) [p. 126], (3) Utterance number in the form of $(T-U)$ corresponding to the U’th utterance in Text T [p. 182]
- Square brackets [ ]: (1) Citation, (2) Span of information-structure units (i.e., theme/rheme), (3) Functor of functional application [p. 82]

- Curly brackets { }: Span of coordination
- Angle brackets ⟨ ⟩: (1) Exclusion from information-structure analysis [p. 2], (2), Structured meaning (as in $\langle X, Y \rangle_{L-R}$ where L/R are left/right boundaries) [p. 88], (3) General meta-variable [p. 95]
- Double square brackets ⟦ ⟧: Semantic value [p. 82]
- Asterisk *: Ungrammatical sentence [p. 17]
- Number sign #: Contextually inappropriate utterance [p. 17]
- Prime $'$: Translation of linguistic expression to semantic representation [p. 82]
- Right-arrow $\longrightarrow$: CCG rule [p. 96]
- Hollow circle $\circ$: Functional composition [p. 97]
- Plus $+$: Category combination [p. 110]
- Up-arrow (superscript) $^{\uparrow}$: Type-raised category [p. 97]
- Double slash $//$: Modification structure (in semantic representations) [p. 81]
- Grammatical labels for Japanese:
  - TOP = topic marker (thematic function of *wa*)
  - CONT = contrastiveness marker (strong contrastive function of *wa*)
  - NOM = nominative case marker
  - ACC = accusative case marker
  - DAT = dative case marker
  - GEN = genitive case maker
  - COP = copula
  - COMP = complementizer
  - NML = nominalizer
  - NEG = negation
  - Q = question

Chapter 1

Introduction

This thesis concerns the notion of 'information structure': informally, organization of information in an utterance with respect to the context. In this introductory chapter, we discuss the motivation for the thesis, a brief introduction to information structure as well as a summary of the problems with previous work, and the main points and contributions of the thesis.

Motivation: Computer Applications

The necessity of incorporating information structure has been recognized but also considered a challenge in many areas of natural language processing (NLP). In this section, we begin by observing this point in three such areas: machine translation, speech generation, and writing assistance.

First, let us consider translating into Japanese the following part of a text taken from a medical case report.

(1) *i.* (Title) ⟨Osteoporosis in Active Women: Prevention, Diagnosis, and Treatment⟩

ii. Osteoporosis has been defined as "a disease characterized by low bone mass and microarchitectural deterioration of bone tissue, leading to enhanced bone fragility and a consequent increase in fracture risk."

iii. Although anyone can develop osteoporosis, postmenopausal women and young females with menstrual irregularities are most commonly affected.

iv. (cont'd)

In discourse examples like this, we label utterances with italic roman numerals. Material not considered for analysis, such as the title in the above example, is enclosed in angle brackets.

A somewhat simplified translation of the utterance (1*ii*) might look like the following:

(2) Kotusosyou_syou-wa ... byouki-to teigisaretekimasita.
osteoporosis-TOP disease-as has_been_defined
"Osteoporosis has been defined as a disease ...."

In the above, the so-called 'topic' marker *wa* is used for the grammatical subject. On the other hand, in the next utterance (1*iii*), the nominative case marker *ga* is more appropriate:

(3) ... wakai zyosei-ga mottomo ooku eikyousaremasu.
young females-NOM most commonly are_affected
"... young females are most commonly affected."

The choice of these particles *wa* and *ga* is context-dependent, as has been discussed by, e.g., Kuno [1972]. In general, it is possible to provide a context where one of these particles is more appropriate than the other. For example, where a certain symptom is described and the name of the disease is then provided as new information, the utterance (2) appears more appropriate, with *ga*-marking on the subject as follows:

(4) Kotusosyou_syou-ga ... byouki-to teigisaretekimasita.
osteoporosis-NOM disease-as has_been_defined
"It is osteoporosis that has been defined as a disease ...."

Therefore, a computer application such as machine translation must be able to identify the involved factors and select particles appropriate for the context. But there have been few reports on this issue in the machine translation literature. Nagao [1989, p. 137] points out that particle choice in relation to 'focus' (closely related to the choice of the nominative case particle *ga* above) is an issue for future study in machine translation research. No further discussion is given in the book.[1] The only project I am aware of that is specific about particle choice between *wa* and *ga* is Matthiessen and Bateman [1991, Section 7.3].

Now let us consider the entire text of (1). In the following, the grammatical subjects of the matrix clauses are italicized:

(5) *i.* (Title) ⟨Osteoporosis in Active Women: Prevention, Diagnosis, and Treatment⟩

[1]The book focuses more on Japanese-English machine translation than on the English-Japanese direction, though.

ii. *Osteoporosis* has been defined as "a disease characterized by low bone mass and microarchitectural deterioration of bone tissue, leading to enhanced bone fragility and a consequent increase in fracture risk."

iii. Although anyone can develop osteoporosis, *postmenopausal women and young females with menstrual irregularities* are most commonly affected.

iv. *An estimated 20% of women more than 50 years old* have osteoporosis.

v. Although most studies have focused on women of this age-group, *osteoporosis* is potentially more deleterious in younger women because they haven't yet attained peak bone mass, and early bone loss therefore can affect the rest of their lives.

vi. Whether patients are younger or older, *the social costs of osteoporosis* are enormous.

vii. *The yearly estimated healthcare bill for osteoporotic fractures* is between $2 billion and $6 billion.

viii. *About 200,000 osteoporosis-related hip fractures* occur each year in the United States,

ix. ⟨and⟩ *the mortality rate 1 year after fracture* is estimated to be as high as 20%.

The last compound utterance is divided into two lines for simplicity. We ignore the word *and* in (5*ix*) from analysis (considered as a discourse marker as a result of the split). The appropriate particle choice for each grammatical subject in the corresponding Japanese translation is shown in Table 1.1. The judgment is made consistently by multiple human translators (a detailed description is given in Chapter 7).

Utterance	Particle choice	Utterance	Particle choice
(*ii*)	*wa*	(*vi*)	*wa*
(*iii*)	*ga*	(*vii*)	*wa*
(*iv*)	*ga*	(*viii*)	*ga*
(*v*)	*wa*	(*ix*)	*wa*

Table 1.1: Particle Choices by Translators

Obviously, categorical choice of either *wa* or *ga* would result in an incorrect distribution. Two potential factors involved in this process are 'discourse status' [Prince, 1981] (for the current purpose, 'old'/'new') and 'definiteness' [Prince, 1992] (use of a definite determiner, etc.). For example, we might hypothesize that a discourse-old element is attached by *wa*, or a definite expression

is translated into a phrase with *wa*.[2] But neither of these factors alone can predict the appropriate particle choices as shown in Table 1.2. Our experiment, reported in Chapter 7 (for approximately 100 particle choices), shows that both of these hypotheses perform poorly.

Utterance	Particle choice	Hypothesis 1 { Disc-old →wa { Disc-new →ga		Hypothesis 2 { Definite →wa { Otherwise →ga	
(*ii*)	*wa*	Old	$\surd$	Indefinite	$*$
(*iii*)	*ga*	New	$\surd$	Indefinite	$\surd$
(*iv*)	*ga*	New	$\surd$	Indefinite	$\surd$
(*v*)	*wa*	Old	$\surd$	Indefinite	$*$
(*vi*)	*wa*	New	$*$	Definite	$\surd$
(*vii*)	*wa*	New	$*$	Definite	$\surd$
(*viii*)	*ga*	New	$\surd$	Indefinite	$\surd$
(*ix*)	*wa*	New	$*$	Definite	$\surd$
			$\surd$: correct, $*$: incorrect,		

Table 1.2: Particle Choices and Simple Hypotheses

Phenomena closely related to particle choice in Japanese have been observed in other languages as well. Word order in Turkish and Polish is not grammatically constrained (i.e., free word order) [Hoffman, 1995], but still depends on the context [Hoffman, 1996 (for Turkish); Styś and Zemke, 1995 (for Polish)].

A hypothesis put forward by a number of researchers is that the notion of 'information structure', organization of information in an utterance, is behind these phenomena despite the fact that information structure is realized differently in different languages. The importance of information structure has also been addressed in a large-scale machine translation project [Kay et al., 1994, p. 94]. But at this point, few results have been reported. Similarly, the importance of discourse processing in voice-to-voice machine translation has also been discussed [LuperFoy, 1997].

Let us now turn to the second type of application, i.e., speech generation systems. The traditional speech generation systems focus on the level within a sentence and do not usually address the issues of information structure except for deaccentuation of a 'previous mention' [Sproat, 1998, Sec. 4.1]. Steedman [1997] points out that some translation output of the Verbmobil project [Kay et al., 1994] is not contextually appropriate and that it can be improved if information structure is also considered in the system. A systematic approach to this problem has been worked out by

[2]Japanese does not have a definite marking system corresponding to that of English.

Prevost [1995], focusing on generation of intonation in English and analyzing the contrast between salient individuals.

In our example, the first sentence of the text (1*ii*) may naturally correspond to a pitch-accent pattern like (*a*) rather than (*b*) below (in the given context). Note that boldface indicates phonological prominence.

(6) *a*. Osteoporosis has been defined as "**such and such**".

b. **Osteoporosis** has been defined as "such and such".

The above contrast can be most readily seen for the case where the previous mention is deaccented and the 'new' material is pronounced prominently. But the phenomenon is not limited to such a simple pattern. There are cases where a previous mention needs to be pronounced prominently, as in the following example [Prevost, 1995, (2), p. 3]:

(7) *Q*: Does your older brother prefer baroque or impressionistic music?

A: My older brother prefers **baroque** music.

Thus, organization of information within an utterance, not just simplistic 'old' vs. 'new', is also relevant to speech generation systems.

Interestingly, the choice of phonological prominence has some relation to particle choice in Japanese. Namely, the subject in boldface is *ga*-marked and the subject not in boldface is *wa*-marked. The linguistic realization in both of these cases does not directly correspond to notions such as discourse status or definiteness, but appears to correspond to information structure.

Finally, let us consider an application of information structure in Computer-Assisted Writing systems [e.g., Komagata, 1998a]. The idea can be illustrated by the following example similar to the one found in Booth et al. [1995] (on how to write a research paper):

(8) *a*. The mitral valve could be permanently damaged if the patient has mitral valve prolapse and develops endocarditis. Medication that controls infection will not halt this damage. Only surgery which repairs the defective valve will achieve that goal.

b. If the patient has mitral valve prolapse and develops endocarditis, the mitral valve could be permanently damaged. This damage will not be halted by medication that controls infection. That goal will be achieved only by surgery which repairs the defective valve.

Booth et al. [1995] argue that (*b*) is more readable for the following reason. In each sentence in

(*b*), the information is placed in the order from 'old' to 'new', and this 'old things first' preference is at work in written English. Similar arguments have been made in the theoretical literature as well [e.g., Kuno, 1978]. But this type of advice can be overlooked even by native speakers of English, not to mention non-native speakers. For example, the readability distinction between (8*a*) and (8*b*) may not be perceived in a similar way by Mandarin speakers because the passive construction in Mandarin involves a special pragmatic function (a kind of 'negative' sense) [Cowan, 1995, p. 36]. If we assume the 'old things first' preference, and with an understanding of the mechanism underlying this phenomenon, we could develop an application such as a Computer-Assisted Writing system that could advise the user to write (*b*) instead of (*a*). Such a system could be integrated with a grammar checker, [e.g., Park et al., 1997], to provide a wider coverage in writing assistance than is currently practiced. Again, information structure is a critical element in this type of application. While previous work often made the 'old'/'new' distinction for this phenomenon, I argue that the underlying concept is also information structure in a sense discussed by Daneš [1974] as 'thematic progression'.

This rather lengthy section on motivation demonstrates that information structure is an essential element in multiple computational applications, as shown schematically in Fig. 1.1. If we can mechanically capture the effect, we can improve the quality of machine translation, assign appropriate intonation for the utterances in an extended text, and provide assistance to a writer with respect to one aspect of text readability/coherence. Thus, a solution to the first problem provides a solution to the others.

Information Structure

Let us now briefly describe the notion of information structure introduced earlier as organization of information within an utterance. Research on information structure has a long history and is couched in different names and definitions, e.g., Mathesius [1975, manuscripts from the 1920s], Halliday [1967], and Kuno [1978]; from computational viewpoints, Winograd [1972] and Kay [1975]; and more recently, Vallduví [1990].

The effects of information structure, in the sense of Vallduví [1990], are often analyzed in a question-answer context, as in the following example:

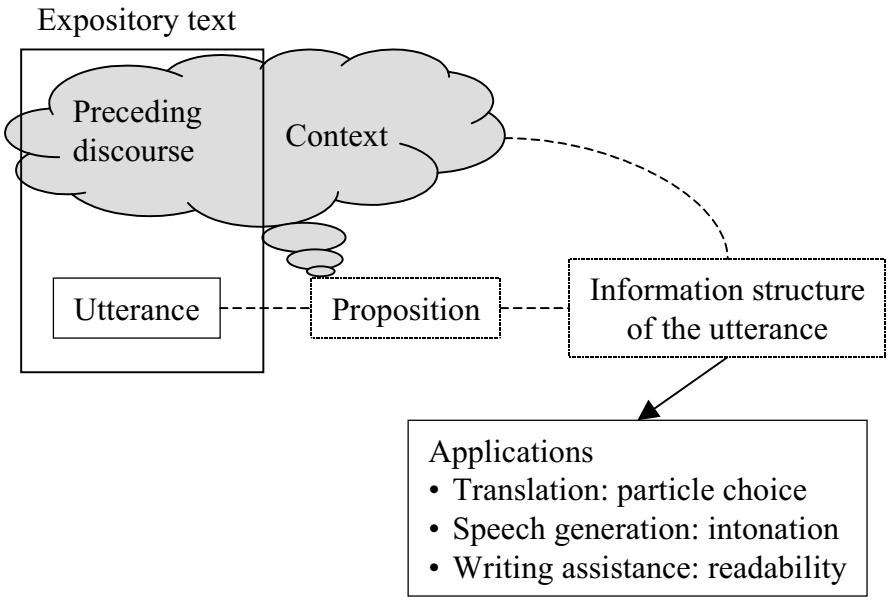


Figure 1.1: The Phenomenon under Investigation

(9) *Q*: What did the patient develop?

A: [She developed] [**endocarditis**].

The informational division in the response is clearly perceived in relation to the presupposition introduced by the question, or similarly in relation to the *wh*-phrase in the question. That is, the phrase in the response that corresponds to the *wh*-phrase in the question provides pertinent information that makes the response informative in the context. In this sense, we say that information structure manifests informational contrast between units in an utterance. This type of partition has been variously called 'theme'/'rheme', 'given'/'new', and 'topic'/'focus'. For the moment, the fine distinction between the terms is not critical.

The main concern of this thesis is mechanical identification of information structure, useful for the applications introduced in the previous section. Let us call this the **Identification Problem**, and briefly point out the problems with previous work: a group of computational approaches and another group of more theoretically-oriented work.

First, there are several algorithms proposed to identify information structure [Kurohashi and

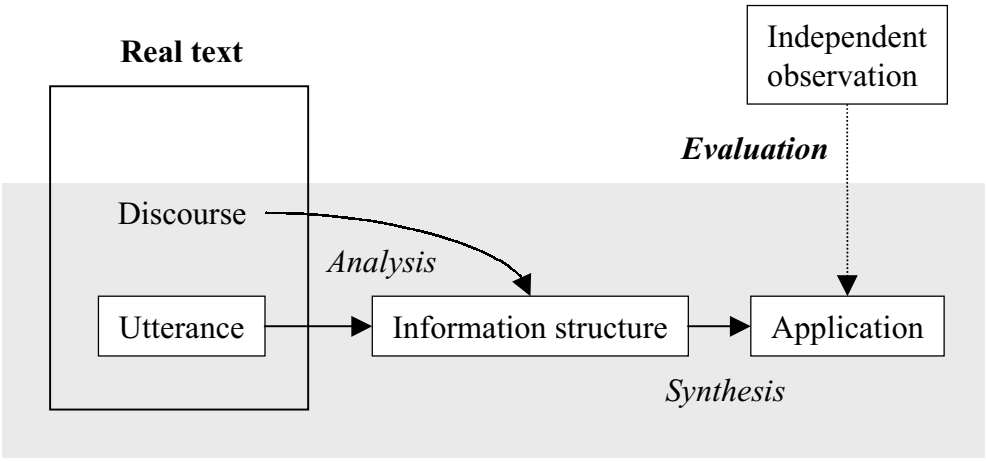


Figure 1.2: Limitations of Previous Approaches to the Identification Problem

Nagao, 1994; Hajičová et al., 1995; Hahn, 1995; Styś and Zemke, 1995; Hoffman, 1996; Komagata, 1998a]. But none of these approaches is satisfactory in terms of analyzing realistic texts and evaluating the results with respect to distinct observable phenomena. Hajičová et al. [1995], Styś and Zemke [1995], and Hoffman [1996] cannot be applied (in their proposed form) to a text of the complexity we have observed earlier, e.g., (5). Levinson [1983, p. x] questions the usefulness of information-structure study by pointing out that theories are not applicable to arbitrarily complex linguistic structures. Next, and more importantly, none of these proposals offers an evaluation procedure. Thus, the current computational approaches are limited to the shaded area in Fig. 1.2. In order to construct and make a judgment about a theory of information structure addressing the Identification Problem, we need to extend the project to the entire area of the same figure.

Next, one major problem shared by virtually all theoretical proposals on information structure is lack of explicitness. While a great many properties, e.g., referential status and linguistic marking, have been identified in relation to information structure, the results are not at the level available to computational applications (as will be demonstrated in Chapter 2). This difficulty partly arises because information structure involves the notion of inference. Since inference is an open-ended search process, attempts to involve inference in the definition of information structure face considerable difficulty [e.g., Rochemont, 1986].

Another problem with the theoretical literature is its indifference to the Identification Problem.

Some assume that the information structure is linguistically identifiable [e.g., Vallduví, 1990], which is not actually the case [e.g., Brown and Yule, 1983]. The focus of theoretical studies [e.g., von Stechow, 1981] is often on the relation between a known information structure and its referential/linguistic properties. Thus, the Identification Problem is not even discussed. Another group of researchers assume that question-answer context can be used to identify information structure in expository texts [e.g., Sgall, 1975]. Some explicitly hypothesize an implicit question for each utterance in a text [e.g., van Kuppevelt, 1995]. But the use of question-answer context is not automatically applicable to texts, and the implicit-question approach (without specifying how to obtain implicit questions) simply sidesteps the problem of identification of the right implicit question. Since information structure affects coherence and readability in both question-answer pairs and texts in a similar manner, we need a more general characterization of information structure applicable to both question-answer contexts and written discourse.

Reflecting on the above observation, it is fair to say that the Identification Problem remains open. And we have good reasons to tackle it.

Main Points

In response to the situation described above, this thesis argues for the following point.

(10) (main point of the thesis) A theory of information structure that explicates the properties of its components and their relations can be used to identify information structure in a realistic set of texts. It is also possible to provide an evaluation method that demonstrates that the proposed theory is an improvement over some alternative hypotheses underlying existing algorithms to identify information structure.

In order to be able to accept or reject the above statement, we will need to firmly grasp the concepts involved at a level we can specify and computationally implement. This thesis discusses in detail (1) how the proposed theory is developed, drawing on the existing theories of information structure, (2) what constitutes the process of identifying information structure in real texts, and (3) how the theory can be evaluated and compared with different hypotheses. Once these concepts are shared with the reader, the final question is whether the main point (10) can be accepted.

The main *theoretical* hypothesis of the thesis is that (i) information structure is informational contrast (following Vallduví [1990]) between complementary units of an utterance, i.e., 'theme' and 'rheme' [Mathesius, 1975], and (ii) only the theme is necessarily 'contextually-linked', a notion closely related to 'context set' [Stalnaker, 1978] and 'alternatives set' [Rooth, 1985]. The second theoretical point is that (i) the property 'contextual link' can be characterized in terms of 'bounds' on inference, including *zero* inference (i.e., immediately available in the context), and (ii) this bound is set by factors *external* to the logic of inference. A corollary to this second point is that contextual links can be and must be identified by logic-external properties, including discourse status [Prince, 1992], linguistic marking [Heim, 1982, among many others], and certain domain-specific knowledge. Although the Identification Problem obviously applies cross-linguistically, this thesis concentrates on a special case of English. Considering that English heavily depends on intonation for marking information structure in the spoken form, text analysis in English is not an easy task. But what we want to show in this thesis is that there is an underlying principle that applies even to written English. For other languages, language-specific modules can be replaced with appropriate ones, possibly with more encoding of information structure.

In order to delineate a theory of information structure, we need to interface the notion of information structure with components including discourse processing and surface structure. As we will see later, most traditional grammars have a crucial drawback in this regard. Their notion of surface constituency is not as flexible as the semantic units we want to consider for discourse processing. As a solution to this problem, we adopt the grammatical framework of Combinatory Categorial Grammar (CCG) [Ades and Steedman, 1982]. This enables us to explicitly state our theory of information structure as a part of the grammar itself, and provides a basis for implementation. Furthermore, in order to analyze information structure in realistic texts, we adopt the idea of 'structured meaning' [Krifka, 1992], which enriches the semantic structure with an additional degree of freedom without losing precision.

Our implementation of the information-structure analyzer demonstrates that the theory is explicit enough for the current purpose and applicable to realistic texts. But the most critical element of the entire process is evaluation of the identification process. We take advantage of the particle-choice problem in English-Japanese machine translation. Our implementation not only identifies

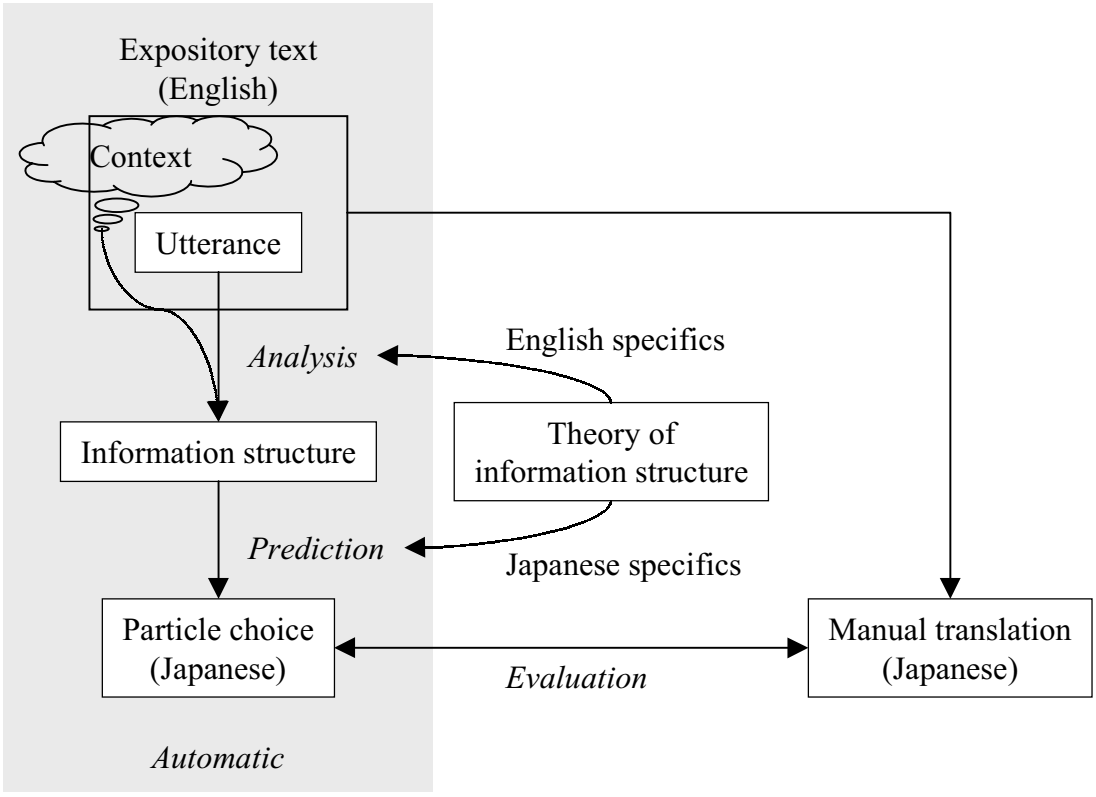


Figure 1.3: Overview of the Project

the information structure of the utterances but also predicts appropriate particles for the grammatical subjects, i.e., the choice between 'topic' particle *wa* vs. nominative case marker *ga*. The prediction is then compared with manual translations. This process is schematically shown in Fig. 1.3.

This process also requires us to understand the realization of information structure in Japanese. As will be seen in Chapter 5, the use of particles in Japanese is complex. A detailed discussion of the language provides us with a solid ground for the use of translation as an evaluation method.

At the end, we demonstrate that our theory is an improvement over the simple hypotheses 1 and 2 in Table 1.2, which underlie existing algorithms of identifying information structure. Although the experiment is limited in its scale and the scope of evaluation, its results support the claim that information structure can be used in computational applications.

Contributions of the Thesis

The main contribution of the thesis is a demonstration of identifying information structure, its evaluation, and its applicability to practical applications. This development improves the state of understanding, which has been intuitive but not objective. The demonstration consists of several key elements. First, we tackle the Identification Problem so that the results of the project are immediately available to practical applications. Second, inclusion of evaluation provides a basis for judging the main point (10). Third, by dealing with realistic texts, we challenge the skepticism about generality of information-structure analysis. Furthermore, development of an explicit theory of information structure provides a connection between theory and procedure that has been missing from existing computational approaches.

Other contributions of the thesis include the following. Use of a grammar-based parser provides a precise connection between utterance-level linguistic description and certain discourse-level concepts. We adopt a system of structured meaning that is more comprehensive than existing theories. Finally, the analysis of information-structure marking in Japanese provides information useful for research and education involving this language.

Overview

This thesis is organized in the following way. In Chapter 2, we start our study of information structure by defining the Identification Problem for information structure. This leads us to questions to be investigated in the literature review. The chapter first looks at a number of theoretical proposals about information structure. Information structure is analyzed in connection to referential status, contrastiveness, and linguistic form. This chapter also discusses the internal structure of information structure, including the question whether it is recursive or not. After this, we review several computational approaches to the Identification Problem.

Chapter 3 proposes a theory of information structure as a basis for the solution to the Identification Problem for expository texts. The theory is based on the idea of ‘information packaging’ [Vallduví, 1990], and explicates this as a binomial partition between ‘theme’ and ‘rheme’. We hypothesize that a crucial property in distinguishing these components is ‘contextual linking’ and

present a way to characterize it in terms of discourse status, domain-specific knowledge, and linguistic marking. The chapter also addresses a potential problem associated with constituency and discontiguous cases of information structure and provides a solution based on the idea of 'structured meaning' as a structure of semantic representation [Krifka, 1992].

Chapter 4 bridges the theory and an implementation. In order to provide a computational framework that can recognize constituents in accordance with information-structure partitions, we adopt Combinatory Categorial Grammar (CCG) [Ades and Steedman, 1982]. We show that specification of 'contextual link' can be formalized within the framework, and analysis of discontiguous information structure can also be spelled out.

In Chapter 5, we carefully sort out the conditions under which Japanese particles can be considered markers for information structure. The task is rather complicated because of the contrastive semantics also involved in these particles. Once this is done, we apply this analysis in the prediction of particle choice from information structure. This provides the basis for the evaluation of the analysis of English through comparisons between mechanical prediction and the corresponding human translation.

The next step in Chapter 6 is to implement an information-structure analyzer built on a CCG parser. We first address the practicality of our CCG parser, considering the issue of so-called 'spurious ambiguity', a problem for CCG and related Categorial Grammar formalisms. The chapter shows that existing technologies provide practical solutions to this problem. Second, we describe the module responsible for analyzing information structure based on the formalization of the proposed theory.

In Chapter 7, we evaluate the theory through comparison of the particle prediction made by the system and that made by human translators. We describe the experiment data and the evaluation procedure in detail. The results are compared with two simple hypotheses and a chance result. An extensive discussion of the results is also provided.

In the concluding chapter, we summarize the results of the thesis and discuss its contributions, and then address some directions for future work.

Chapter 2

Information Structure: The State of the Art and Open Questions

In this chapter, we review existing theories of information structure and computational approaches to identifying information structure. We first point out that some existing definitions of information structure fail to explicate the properties of its components and the relation between the components. The next point is that most theoretical proposals about information structure are indifferent to the Identification Problem and lack the explicitness required for formalization and implementation. Finally, we observe that existing computational approaches do not yet provide a solution to the Identification Problem due to their limited coverage, lack of evaluation, and missing connection to theories.

To clarify our goal, we begin this chapter with a discussion of the Identification Problem for information structure. After presenting an informal view of information structure, we move to the review of theoretical and computational proposals in that order.

2.1 The Identification Problem

In the Introduction, we noted that the Identification Problem for information structure is necessary for applications such as machine translation, speech generation, and computer-assisted writing. This section explores this problem more in detail and identifies the associated subgoals.

The **Identification Problem** takes the following form. Given a text such as the one shown

below, the information structure consisting of two components, say, 'theme' and 'rheme', for each utterance except for the title must be identified (the text is taken from our experiment data, which will be discussed in Chapter 7).

(11) Title: Osteoporosis in Active Women: Prevention, Diagnosis, and Treatment

Osteoporosis has been defined as "a disease characterized by low bone mass and microarchitectural deterioration of bone tissue, leading to enhanced bone fragility and a consequent increase in fracture risk." Although anyone can develop osteoporosis, postmenopausal women and young females with menstrual irregularities are most commonly affected. An estimated 20% of women more than 50 years old have osteoporosis. Although most studies have focused on women of this age-group, osteoporosis is potentially more deleterious in younger women because they haven't yet attained peak bone mass, and early bone loss therefore can affect the rest of their lives.

Now, suppose that a hypothetical procedure identifies the information structures as follows:

(12) Title: Osteoporosis in Active Women: Prevention, Diagnosis, and Treatment

[Osteoporosis]$_{Theme}$ [has been defined as "a disease characterized by low bone mass and microarchitectural deterioration of bone tissue, leading to enhanced bone fragility and a consequent increase in fracture risk."]$_{Rheme}$ [Although anyone can develop osteoporosis]$_{Theme}$, [postmenopausal women and young females with menstrual irregularities are most commonly affected]$_{Rheme}$. [An estimated 20% of women more than 50 years old]$_{Rheme}$ [have osteoporosis]$_{Theme}$. [Although most studies have focused on women of this age-group]$_{Theme1}$, [osteoporosis]$_{Theme2}$ [is potentially more deleterious in younger women because they haven't yet attained peak bone mass, and early bone loss therefore can affect the rest of their lives]$_{Rheme}$.

At this point, one may naturally ask questions such as the following:

1. What is 'information structure'? In other words, what do we want to identify? How to separate information structure from various related properties?

2. How can these information structures be identified? Is the procedure related to *any* theory of information structure?

3. How can we say whether the identified information structures are correct with respect to our goal?

The extent of discussion responding to the first question is enormous. But the foci of attention and points of view are quite diverse. Also reflecting the complexity involved in the question, it is fair to say that there are no uniformly agreed answers to this question. In addition, looking at this question from the entire span of the Identification Problem, many proposals are not sufficiently explicit for the next two steps.

The second question has received much less attention. Although several proposals have been made, each one of them has weaknesses in the coverage and/or theoretical foundation. Finally, the third question has rarely been addressed. In order to complete the entire process of the Identification Problem, this question must be answered. In the rest of this chapter, we explore these three questions in relation to previous work.

Before proceeding, it is illuminating to briefly mention closely related work by Heine [1998] and Murata and Nagao [1998]. Their focus is identification/generation of definiteness (in English) in Japanese-English machine translation. This problem is in a sense the opposite direction of the Identification Problem. But it is a problem distinct from the Identification Problem for information structure because generation of definite marking in English requires a different set of criteria. For example, we will see that definiteness marking within an embedded clause cannot be predicted from information structure (see Subsection 2.3.3).

2.2 What is Information Structure?

This section reviews the phenomenon under discussion, observes difficulties with previous definitions of information structure, and introduces a characterization of information structure that serves as the basis for subsequent discussion. At the end, the assumptions and qualifications for the present work are described.

Phenomenon under Discussion

Let us start from some observations involving a question-answer pair. Throughout this work, the **boldface** in examples indicates a pitch accent.[1]

(13) *Q*: Who did Felix praise?

A_1: Felix praised **Donald**.

A_2: # **Felix** praised Donald.

A_3: # Felix **praised** Donald.

While the choice (A_1) is appropriate as a direct response to the question, the other two preceded by '#' are not. The symbol '#' is used as contextual inappropriateness throughout the present work, cf. the use of '*' for ungrammaticality. In this case, placement of pitch accent is relevant to the delivery of information. Similarly, the following distinction can also be observed.

(14) *Q*: Who did Felix praise?

A_1: It was **Donald** whom Felix praised.

A_2: # It was **Felix** who praised Donald.

In the above case, syntax (in conjunction with intonation) has an effect similar to that of intonation in the previous example. All of the above responses in (13, 14) are grammatical, and presumably share the same propositional (truth-conditional) meaning. But they have distinct felicity conditions depending on the phonological or syntactic realization. This observation about a direct response to a question lets us believe that there is a pragmatic aspect in addition to truth-conditional semantics, which may be realized in distinct linguistic forms. This way of checking information structure is commonly called the **question test** [e.g., Sgall, 1975]. While the question test is useful for informal analysis of information structure, we do not adopt the position that the question test can always be used to identify information structure. There are complicated cases. For example, a response to a question may be embedded in a complex utterance, or responses to multiple questions may be combined into an utterance. We will explore a theory of information structure that captures the intuition behind the question test but also applies to arbitrarily complex structures in expository texts.

[1] In many papers, a pitch accent is indicated by UPPERCASE or SMALL CAPS. When we cite examples from them, these conventions are translated into **boldface**. In this and the following examples, all occurrences of pitch accent correspond to H* tone in the notational system of Pierrehumbert and Hirschberg [1990].

The phenomena related to information structure are observed in various languages in a number of ways. In English, the function of intonation related to the above point is reported in Pierrehumbert and Hirschberg [1990, Sections 5.1 and 5.3]. Certain types of pitch accents, e.g., represented as L+H* and H*, are argued to have distinct functions related to the contrast seen in (13) [Steedman, 1991a]. In addition, various syntactic forms such as topicalization, left dislocation, cleft, VP preposing, inversion, heavy NP shift, *since*/*because*, etc. have been studied in this connection [Prince, 1984; Ward, 1990; Birner, 1994; Lambrecht, 1994; among others]. These and other types of syntactic realization are extensively discussed in, e.g., Lambrecht [1994]. More visible relations to syntactic structure are observed as word order in Catalan [Vallduví, 1990], Czech [Sgall et al., 1986], Hungarian [Kiss, 1987], Russian [King, 1995; Paducheva, 1996], Turkish [Hoffman, 1995, citing earlier work], Polish [Styś and Zemke, 1995], and Finnish [Vallduví and Vilkuna, 1998]. Another form of realization is through morphology in Japanese [Kuno, 1972], and Korean [Wee, 1995]. Vallduví and Engdahl [1996] present an extensive cross-linguistic review also including Dutch and German.

The above observation urges us to derive a general description of the phenomenon across languages. Since linguistic realization is quite diverse, it is reasonable to consider that such linguistic marking is arbitrary [Prince, 1998, p. 282].

Returning to an earlier example repeated below, we assume that the informational statuses of "*Felix praised*" and "*Donald*" are distinct.[2]

(15) *Q*: Who did Felix praise?

A: [Felix praised] [**Donald**].

And this informational contrast affects the felicity of the utterance. Although the above illustration uses a question-answer context for presentation purposes, the same phenomenon is observed in written texts, as in (12) in the previous section. Due to a lack of prosodic information in texts, languages like English lose certain properties that may be marking information structure. In some cases, punctuation may be used to supplement prosody. But other languages that mark information structure non-prosodically may retain more linguistic properties relevant to information structure. Considering that reading in English does not seem to suffer from lack of direct information-structure marking, we assume that there is an underlying mechanism of identifying information

[2] A related but distinct notion of information structure is developed in Roberts [1996, 1998].

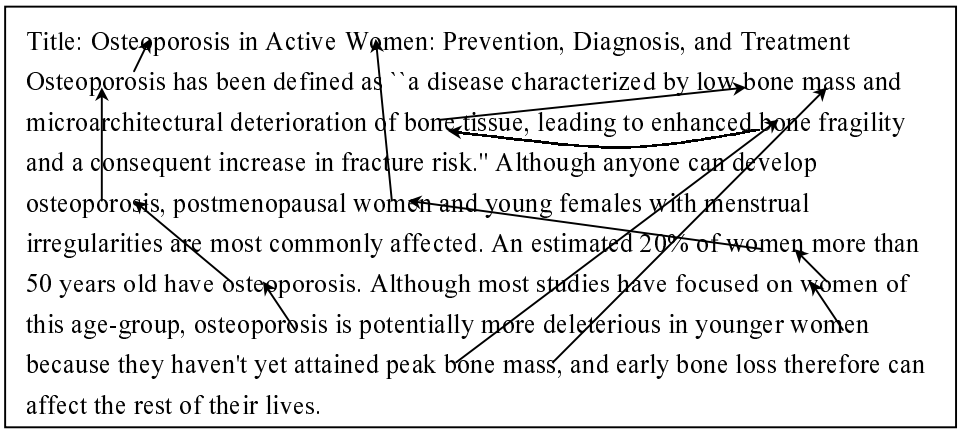


Figure 2.1: Text Link

structure that works for all the cases including written English.

At this point, we should note that our notion of information structure is orthogonal to the notion of 'discourse topic' [Brown and Yule, 1983, Section 3.3 (for review)]. An illuminating (informal) definition of **discourse topic** is that it is the title of a text [Brown and Yule, 1983, p. 71]. In general, discourse topic is a phrase (or a proposition, depending on the definition) associated with a text, and is *not* about the informational contrast within an utterance. As a consequence, a discourse topic may or may not be the theme of an utterance.

There is another group of work also orthogonal to the present approach. This group applies statistical methods to analyze text link (their 'topic') in a large corpus for speech recognition [Sekine, 1996; Jokinen and Morimoto, 1997] and discourse segmentation [Reynar, 1998]. The idea of text link is shown in Fig. 2.1. The focus of this group is a *macro* view of the discourse, and not the utterance-internal information structure we are looking at.

Difficulty with Previous Definitions

To capture the phenomenon discussed above, let us take a look at two definitions of information structure. First, Vallduví [1990, p. 18] provides the following, as a concept underlying information structure.

(16) INFORMATION PACKAGING: A small set of instructions with which the hearer is instructed by the speaker to retrieve the information carried by the sentence and enter it into her/his knowledge-store.

This definition is too broad as a starting point to work on the phenomenon of information structure. In fact, it equally applies to 'instructions' for speech acts. For example, it *could* be used to describe the distinction between locutionary act (reference) and illocutionary act (conventional force associated with it) [Austin, 1962].

Here is another definition from Lambrecht [1994, p. 5].

(17) INFORMATION STRUCTURE: That component of sentence grammar in which propositions as conceptual representations of states of affairs are paired with lexicogrammatical structures in accordance with the mental states of interlocutors who use and interpret these structures as units of information in given discourse contexts.

This appears to contain critical elements of information structure. But it could apply to, say, presupposition projection [Gazdar, 1979]. For the investigation of the Identification Problem for information structure discussed in the previous section, both of these definitions seem to allow arbitrary instance of a theory and implementation.

Although both of the above definitions are an attempt to clarify the long-standing vagueness associated with the notion of information structure, they are not successful as a definition of information structure. To avoid problems like this, even the top-level characterization of information structure should mention the involved components and properties associated with them.

Information Structure as Semantic Partition between Theme and Rheme

Let us first observe Vallduví's [1990, p. 3] intuition behind information packaging: "speakers seem to structure or package the information conveyed by a sentence at a given time-point" (following earlier work of Chafe [1976] and Prince [1986]). As a simplest model, we consider a structure of two components that differ in terms of delivery of information. For example, as seen earlier, the response in a question-answer pair exhibits this point.

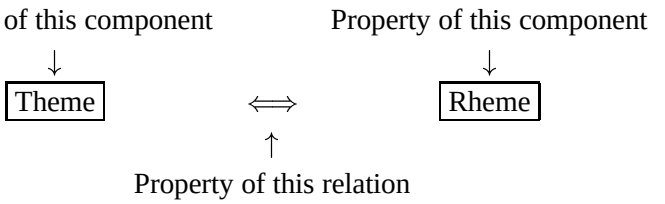
(18) *Q*: Who did Felix praise?

A: [Felix praised]$_{Theme}$ [**Donald**]$_{Rheme}$.

As in the above, we call the two components **theme** and **rheme**, following Mathesius [1975, p. 81] and Halliday [1967, p. 211]. The choice of the terminology is mainly to avoid related, but heavily overloaded terms such as 'topic' and 'focus', or 'old' and 'new' (for an extensive review of terminologies, see Vallduví [1990, Chapter 3]). But we do not follow Halliday [1967, p. 212] who states that a theme is the utterance-initial constituent. Now, our starting point is to characterize **information structure** as the abstract representation of such an organization of informational components.

To be able to provide a solution to the Identification Problem, we need to clarify the properties associated with theme, rheme, and the relation between them, schematically shown below.

(19)



This corresponds to Vallduví's [1990, p. 23] intuition about information structure as a *relational* notion. While Vallduví departs from the binomial partition (see in Subsection 2.3.4), we pursue this binomial model in order to maintain a clear and simple notion for the relation between theme and rheme.

We now describe a preliminary version of the main hypothesis about information structure as follows:

(20) **Main Hypothesis** (preliminary version)

a. The theme is always linked to the context (but rheme is not necessarily linked to the context).

b. The rheme is always contrastive, in a broad sense (but theme is not necessarily contrastive).

c. The information structure of an utterance is a complementary, semantic partition between theme and rheme.

The first point (20*a*) basically follows many previous proposals [Chomsky, 1971; Jackendoff, 1972; Sgall et al., 1986; Rochemont, 1986; Prince, 1992]. These proposals are discussed in detail in Section 2.3.2. A more precise characterization of the idea awaits Chapter 3.

A traditional characterization of rheme is to associate it with some kind of 'newness' [e.g., Jackendoff, 1972]. We will see that this position cannot be maintained (Subsection 2.3.1). Instead, the second point (20*b*) associates rheme with a general notion of 'contrast' such as proposed by Rooth [1985, (Alternative Semantics)]. This point is discussed in Subsection 2.3.2.

The third point (20*c*) says that theme and rheme are the only components. This also requires that a theme and a rheme combine into a proposition corresponding to the utterance in question.

Unlike the previous definitions (17) and (16), the characterization (20) at least clarifies the involved components and the properties to investigate.

In the rest of this chapter, we review previous work in relation to this informal idea. Not surprisingly, the idea is partially shared by many previous proposals. Nevertheless, we will see that every previous proposal differs from the idea in one way or another. By the end of this chapter, we will have observed that we cannot just adopt a single previous proposal as a basis for formalization and implementation along the line of (20). The main hypothesis (20) is then made more precise in the next chapters. Before moving on to the literature review, let us discuss some assumptions and qualifications.

Assumptions and Qualifications

As Vallduví [1990, Section 2.3] reviews, study of information structure is connected to various areas of linguistic studies. The course of the present work, therefore, must focus on the issues most strongly connected to the Identification Problem. We state some qualifications for the following four areas: contrastiveness, inference, reference resolution, and discourse structure.

In the main hypothesis, contrastiveness is an essential property associated with rheme. Although we review the literature in this respect, we exclude formalization and implementation of contrast. For one thing, contrastiveness is a topic on its own, which deserves a separate study [e.g., Rooth, 1985]. For another thing, its implementation is extremely difficult [Prevost, 1995 (for a small-scale implementation)]. In practice, we can achieve results useful for practical applications, as demonstrated in later chapters.

As we will see shortly in Subsection 2.3.1 (and in other sections as well), the notion of 'contextual link', the required property of theme, involves inference. While we discuss the way inference is involved in the Identification Problem, we exclude from discussion the *mechanism* of inference.

Although inference has been well recognized as a source of linguistic activity [e.g., Grice, 1975] and an active area in Artificial Intelligence (AI) [e.g., Russell and Norvig, 1995 (a standard text)], the state of the art is not yet at the level that we can incorporate it into our theory of information structure. Our position is that study of information structure can be done sufficiently well for practical merits without depending on the understanding of general mechanism of inference, and that the places where we fail are due to the cases where even the state of the art in the inference study does not offer a general solution.

Next, we assume that the result of reference resolution is available prior to analysis of information structure, and exclude the discussion of reference resolution itself. Reference resolution is another difficult problem on its own, theoretically and practically [e.g., Grosz et al., 1995; Hobbs, 1979]. For the purpose of identifying information structure, not knowing the correct referent does not necessarily pose a problem. For example, reference resolution of a definite expression is in general a challenging problem, but a definite expression generally provides sufficient information for the purpose of identifying information structure. That is, it in general refers to *some* entity in the context.

Finally, it is often argued that the discourse structure prior to an utterance affects reference resolution in the utterance [Grosz and Sidner, 1986; Mann and Thompson, 1988]. Now, suppose a case where multiple information structures are ambiguously available (i.e., consistent with the theory). In a way similar to reference resolution, it is quite possible that the discourse structure prior to an utterance may affect disambiguation of the available information structures. We limit our discussion to a theory of information structure that admits possible information structures, much like the way a competence grammar licenses all (and only) grammatical sentences. Although we exclude disambiguation by discourse structure, our implementation includes some heuristics for disambiguation for practical reasons.

2.3 Previous Theories of Information Structure

In his influential textbook, Levinson [1983, p. x] casts a doubt on information structure in the following manner: “the whole area may be reducible to a number of different factors: to matters of presupposition and implicature on the one hand, and to the discourse functions of utterance-initial

(and other) positions on the other." This is a question crucial for the study of information structure, and the discussion continued until Vallduví's [1990] demonstration against the proposition. Since this point illuminates the characteristics of information structure, this section reviews previous work mainly in relation to related properties, to which information structure was considered reducible.

The main goal of the review is to examine theories of information structure for application to the Identification Problem. Accordingly, we will pay close attention to the following check points: (1) Is the Identification Problem acknowledged? (2) Is the coverage of a theory good for realistic texts? and (3) Is the proposal under consideration sufficiently explicit for formalization and computational implementation? At the same time, this review shows that no theory singly delineates the properties addressed in the characterization (20).

In the rest of this section, we discuss information structure in relation to referential status, contrastiveness, and linguistic form. The last subsection discuss several proposals on how to partition information structure.

2.3.1 Referential Status of Theme and Rheme

In this subsection, we review the literature in the following way. Theme and rheme must be seen in relation to some referential property. But we reject the idea that information structure is reducible to referential status. After a closer look at referential status, we revisit the property of theme in connection to inference. The conclusion of the subsection is that we can capture the property of theme in relation to inference but, without depending on the problem of inference itself. At the end, we also discuss the semantic types of referents.

'Functional' Approaches: Recognition of Contextual Effect

The use of the terms 'theme' and 'rheme' dates back to Mathesius's [1975, p. 81] manuscript from 1920s (Mathesius cites even earlier work), replacing more obscure terms 'psychological subject/predicate'. The properties of theme and rheme are characterized informally as 'given' and 'new', respectively [p. 82]. Thus, by this time, properties of theme and rheme in relation to referential status had already been observed. The major contribution of the work is a clear separation of information structure from propositional (truth-conditional) meaning, and its analyses in relation

to word order (linguistic form), esp. in Czech. Mathesius calls the approach Functional Sentence Perspective (FSP) and stimulates the Prague School linguists and others to date. Halliday [1967] develops an extended system of functional (systemic) grammar. The general approach of Halliday has been applied to natural language understanding [Winograd, 1972] and generation [Matthiessen and Bateman, 1991]. Another proposal directly following FSP is due to Kay [1975], but this line has not been followed up very much. Kuno [1978] also extends this tradition and discusses pragmatic effects on English and Japanese grammar. We will come back to Kuno's work in Chapter 5.

One problem with FSP is that the properties of theme and rheme are not clearly characterized in Mathesius [1975] and also in many of the Prague school research, as mentioned in Contreras [1976, p. 16]. This tendency is still observed in more recent work including Sgall et al. [1986]. Sgall et al. [1986, Section 3.4] define 'topic' and 'focus' (corresponding to 'theme' and 'rheme') partly in terms of the notions 'Contextual Bound' (CB) and 'Non-bound' (NB) [Sgall et al., 1986, p. 178]. But the notions of CB and NB escape further clarification. They provide an operational criteria to distinguish the two that "may be found in the question test and in similar procedures" [p. 86]. Recently, an attempt of formalization has been made. For example, Peregrin [1996] describes information structure (their 'topic-focus articulation' or TFA) concisely and clearly [4. and 5. on p. 237]. This is a welcome direction, as we can evaluate the theory. But Peregrin's [1996, p. 239] formalization is too limited, as it states that the subject of an utterance (in English) is connected with a presupposition. But, as we have seen in (6) on page 5, the subject in English can be a rheme. In this regard, Halliday's [1967, p. 212] characterization of 'theme' as the utterance-initial constituent is not realistic either. We have already seen that information structure is more flexible. Another characterization of theme in Halliday [1967, p. 212] as 'point of departure' hardly delineates the involved idea.

Although theme/rheme properties in this tradition are not as clear as they should be, FSP researchers are well aware of the Identification Problem. Daneš's [1974] analysis of thematic progression, a kind of discourse structure that connects a theme to an element in the discourse, is applied to real texts. Thus, at least the theme of each utterance must be identified. The idea of thematic progression has been applied to machine translation [Papegaaij and Schubert, 1988]. But the exposition of this material is not explicit enough for me to evaluate the effectiveness and

correctness of the procedure. Hajičová et al. [1995] along with the associated earlier work provides a computational procedure to identify information structure, to which we will return in Section 2.4.

Outside the above-mentioned work, the Identification Problem is rarely acknowledged in the theoretical studies. A common method of fixing the information structure of an utterance is to apply the 'question test' [Sgall, 1975]. Several proposals extend this idea and assume 'implicit questions' to analyze information structure in texts [e.g., van Kuppevelt, 1995, p. 110; Roberts, 1996, p. 93; Büring, 1997a, p. 178]. They hypothesize that every utterance in a text has a corresponding implicit question. The most fundamental problem with this approach is that it simply sidesteps the issue to another area. None of these analyses offers a precise way to identify the right implicit question. In addition, if we need to consider a set of ambiguous implicit questions, the set could be unbounded due to all sorts of, say, adverbial questions, unless it is constrained in a certain way. I am not aware of any practical use of this approach, e.g., text analysis or implementation.

As for coverage of realistic data, FSP researchers vary greatly. While the study of thematic progression [e.g., Daneš, 1974] commonly analyzes real data, more theoretical analysis such as Sgall et al. [1986] deal with mostly short prepared examples. In the former case, it is not clear how to identify thematic progression, and in the latter case, it is not clear whether their analysis can generally cover realistic data.

Information Structure cannot be Reduced to Referential Status

As the connection between theme and context is observed by FSP researchers, a thought was developing that information structure might be reduced to other properties [Levinson, 1983, p. x]. Chafe [1976] compared notions such as 'givenness', 'contrastiveness', 'definiteness', 'subjects', and 'topics'. But Reinhart [1982] and von Stechow [1981] seem to give the clearest argument against information structure being reduced to referential status. Subsequently, this point is adopted by Vallduví [1990, Subsection 2.3.2] in favor of his analysis of information packaging as an autonomous level of representation.

The following example taken from Reinhart [1982, p. 18] demonstrates the point that information structure is not just referential status:[3]

[3]A similar example is found in von Stechow [1981, p. 96], which is actually a response to an earlier version of Reinhart [1982].

(21) *Q*: Who did Felix praise?

A: [Felix praised]$_{Theme}$ [**himself**]$_{Rheme}$.

Reinhart [1982] points out that the referent of *Felix* and *himself* are identical. But the information structure indicated in the example is fairly clear from in this type of question-answer context. This results in a situation where both the theme and the rheme have the same referential status. Rochemont [1986, p. 52], building on Culicover and Rochemont [1983], suggests a related idea in a different way. He distinguishes two types of rheme (his 'focus'): 'presentational' and 'contrastive'. *Presentational* rheme is roughly a 'new' element and *contrastive* rheme is not 'new' (or 'c-construable' in his terminology) and stands in contrast to some other element.[4] This implies that the referential status of a rheme cannot be fixed. The same point that information structure is not just reference is also made by Hoffman [1998] as she compares the roles of information structure and reference resolution applying a Centering-based theory [Grosz et al., 1995].

After separating information structure from referential status, Reinhart [1982] attempts to characterize theme in terms of the notion of 'aboutness' within formal semantics, adopting Stalnaker's [1978] idea of 'contextual set'. But such a notion is inherently knowledge-level, and requires powerful mechanism of inference, as studied in the area of Artificial Intelligence (AI). Formalization of this kind does not necessarily make the situation more explicit.

There is another attempt to provide a means of integrating information structure within semantic representation [von Stechow, 1981]. This is an important step, and we follow some of his ideas. But the discussion is limited to question-answer context and ignores the critical elements of information structure in real texts.

More on Referential Status

We have started from an intuition developed by FSP that information structure is related to referential status, but rejected the possibility that information structure *is* referential status. One important development about referential status in this connection is that there are more than just 'old' and 'new'.

Prince [1981] analyzed three distinct notions of 'givenness' floating around at that time: (i) givenness in terms of predictability/recoverability [Halliday, 1967; Kuno, 1972], (ii) givenness in

[4]Choi [1996, p. 97] cites Dik for a similar distinction between 'completive' and 'contrastive' foci.

terms of saliency [Chafe, 1976], and (iii) givenness in terms of 'shared knowledge' [Clark and Haviland, 1977]. After noting the subsumption relation between these three rather heterogeneous notions, she proposed a taxonomy in terms of 'assumed familiarity', distinguishing EVOKED, INFERRABLE, and NEW referents. Note that we use SMALL CAPS for these terms throughout this thesis to identify the usage as we are discussing here. EVOKED referents are those textually or situationally evoked in the discourse. INFERRABLE referents are those not evoked in the discourse but the speaker believes that the hearer can infer through non-linguistic means, such as world knowledge. Finally, NEW referents are those new to the hearer (BRAND-NEW) or those known by the hearer but neither evoked in the discourse nor inferred (UNUSED). Among these three types, it is inferrable that complicates the situation most, due to involvement of inference.

Prince [1992] also introduces the notion of **discourse status**: **discourse-old** vs. **discourse-new** depending on whether the referent is introduced in the discourse. Yet another notion is **hearer status**: **hearer-old** vs. **hearer-new** with respect to the speaker's belief about hearer's knowledge. The terminology introduced above is summarized in Table 2.1.

Class	Subclass	Discourse status	Hearer status
EVOKED	Textually EVOKED	Old	Old
	Situationally EVOKED	New	Old
INFERRABLE		New	Old/New
NEW	UNUSED	New	Old
	BRAND-NEW	New	New

Table 2.1: Taxonomy of Assumed Familiarity (adapted from Prince [1981, 1992])

The notion of inferrable is closely related to 'bridging' [e.g., Clark and Haviland, 1977], and is also captured by more general notions of 'accommodation' [Lewis, 1979] and 'presupposition' [Beaver, 1997, for an extensive review].

Revisiting the Referential Status of Theme

The earlier discussion shows that the referential status of rheme cannot be fixed. But, now that we know more about referential status as seen above, we should be able to say more about theme.

Reinhart [1982, p. 21] separates theme from 'oldness' by excluding INFERRABLE (her 'semantic link') from her 'old'. But INFERRABLE and EVOKED referents typically share linguistic

marking such as definite expression for NPs [Heim, 1982]. It is also argued that for a NP, inference is invoked by definite expression when the referent is not readily available [Bos et al., 1995; Poesio and Vieira, 1998]. Birner [1997] argues that VPs and adverb phrases too share linguistic marking between EVOKED and INFERRABLE. Considering these cases, it seems more problematic to completely separate INFERRABLES from EVOKED.

Following Reinhart, Vallduví [1990, p. 25] also separates themehood from discourse-oldness. He states that information packaging is *orthogonal* to referential status [Vallduví, 1990, p. 26]. But we need to take a closer look at this point. Vallduví [1990, p. 26] himself discusses that hearer-oldness as a *necessary* (but not sufficient) condition for topichood. Then, neither of them are in fact against the idea that theme is *not* BRAND-NEW, i.e., some combination of EVOKED and INFERRABLE.

Let us consider EVOKED and INFERRABLE themes in the following two examples:

(22) *i.* John has a house.

ii. [The house]$_{Theme}$ [looks exotic]$_{Rheme}$. (EVOKED)

(23) *i.* John has a house.

ii. [The door]$_{Theme}$ [looks exotic]$_{Rheme}$. (INFERRABLE)

For both of the above responses, it is natural to identify analogous information structures.

This observation is consistent with many other characterizations of theme/rheme (and related) distinctions. For example, Chomsky [1971, p. 199], Jackendoff [1972, p. 230], and Zubizarreta [1998, p. 1] discuss 'presupposition' roughly corresponding to our theme, but is distinct from the usual notion discussed by Beaver [1997]. Their 'focus' corresponds to our rheme in that it is informationally in contrast with theme (their 'presupposition'). But they explicitly state that 'focus' is "the information in the sentence that is assumed by the speaker not to be shared by him and the hearer [Jackendoff, 1972, p. 230] and "nonpresupposed part of the sentence" [Zubizarreta, 1998, p. 1]. This distinction is basically the one between EVOKED/INFERRABLE vs. BRAND-NEW. Note that we have already rejected the simplistic characterization of rheme as BRAND-NEW [cf., Jackendoff, 1972]. Sgall et al. [1986, p. 178] distinguish 'Contextual Bound' and 'Non-Bound' (page 25). Although they do not give a precise definition, Contextual Bound seems to share the property of EVOKED/INFERRABLE. Rochemont [1986, p. 47] introduces the notion of 'c-construable', which again appears to be very close to EVOKED/INFERRABLE. To some extent,

this also corresponds to hearer-old [Prince, 1992, Section 2.2.2] and the idea of 'shared topicality' Gundel [1985].[5]

Then, we should not completely abandon the relation between information structure and referential status as Vallduví [1990, p. 25] states, but should take advantage of the relation between theme and EVOKED/INFERRABLE observed by many researchers. The tentative conclusion here is that the property of theme we mentioned in (20) is related to the referential status EVOKED/INFERRABLE.

Difficulty with Inference

If INFERRABLE is involved in the property of theme, we need to address the issues involving inference. Naturally, this is a difficult task, as can be seen in a few proposals discussed below.

Reinhart [1982, Section 6.4] observes the role of INFERRABLE (her 'semantic link'), but does not explicate how to deal with it. Rochemont [1986, (30), p. 47] starts his definition of 'c-construability' in a fairly formal manner: "A string *P* is *c-construable* in a discourse δ if *P* has a semantic antecedent in δ." Then, another definition for 'semantic antecedent' (31): "A string *P* has a semantic antecedent in a discourse δ, $\delta = \{\phi_1, ..., \phi_n\}$, if, and only if, there is a prior and readily available string P' in δ, such that the uttering of P' either formally or informally entails the mention of *P*." But, then, formal/informal entailment does not get the same level of explicitness.

Bos et al. [1995] analyze the problem of reference within the framework of Discourse Representation Theory (DRT) [Kamp, 1981]. Bos et al. [1995, Section 3.3] classify three kinds of anaphoric relations:

(24) *a*. An antecedent is available in the discourse

b. An 'implicit' antecedent is available in the discourse (after failing the previous step): bridging[6]

c. No antecedent is available in the discourse (after failing the previous step): accommodation

Integrating a constrained form of inference this way has limitations. According to Bos et al. [1995], the shift from (*b*) to (*c*) depends on the availability of a *suitable* anchoring referent. But the

[5] Additional references related to this point include: Dryer [1996], van Kuppevelt [1996]. But we do not consider hierarchy of activation levels, cf. Chafe [1994].

[6] Jäger [1996] has a formal account of bridging based on dynamic semantics.

inference process involved in bridging is presumably a *general* logical process. Then, how can a system know when to fail? On the other hand, while their accommodation always saves the reference in question, we know that accommodation *can* fail. It seems more reasonable to assume that bridging and accommodation are not that different as proposed by Bos et al.

The conclusion of this subsection is as follows. Although information structure cannot be reduced to referential status, theme still has a property that is based on referential status involving inference. The previous work reviewed here fails to explicate this observation and thus not applicable to the Identification Problem. What we need is a theoretically sound, yet formalizable/implementable idea for this condition.

2.3.2 Information Structure vs. Contrast

In this subsection, we separate the notion of contrast from rheme and characterize rheme in terms of a general notion of Alternative Semantics [Rooth, 1985] that can be applied to both contrast and rhemehood.

Distinct Notions Associated with 'Focus'

The term 'focus' is heavily overloaded. Thus, it is important to delineate various notions associated with it. 'Focus' as used by Sgall et al. [1986] and Vallduví [1990] basically corresponds to our 'rheme'. Another group of researchers [e.g., Ladd, 1996, p. 160] use 'focus' as a notion closely linked to phonological properties readily observed at the word level, independent of information structure. While we distinguish these two notions, a more important point is actually to relate these two notions in a systematic way. Note that so-called 'AI-focus' [Grosz and Sidner, 1986, p. 179] is a way to organize referents based on their salience and should be considered distinct from other uses of 'focus' [Vallduví, 1990, p. 46].

The intuition we start from is that information structure is about the informational relation between units within an utterance and contrastiveness is a relation about referents not limited to those within an utterance. Thus, a rheme must always be seen in relation to a theme (possibly null) and a contrast must always be seen in relation to another referent in the context (see Fig. 2.2).

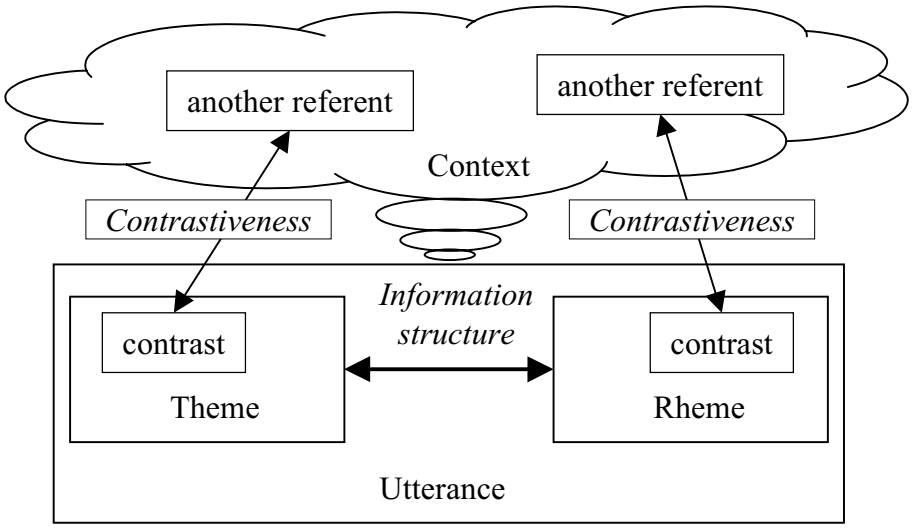


Figure 2.2: Information Structure vs. Contrast

Contrast in Relation to Phonological Prominence

First, let us explore the notion of 'contrast' in relation to **phonological prominence**. We consider a phonological notion of prominence at the perceptual level involving pitch, loudness, duration, and quality [Laver, 1994, p. 450], e.g., in relation to pitch accent (in English) [Ladd, 1996, p. 46, citing Bolinger (1958) and Pierrehumbert (1980)]. We continue to use **boldface** to indicate a word (in examples) where a prominence falls, as in the following example [Ladd, 1996, (5.1), p. 162].

(25) I didn't give him three francs, I gave him **five** francs.

We use the term prominence, rather than pitch accent, to cover cross-linguistic variation in realizing the similar notion in potentially-distinct acoustic properties. Then, as in Ladd [1996, p. 160], "[i]t is now generally accepted that sentence accentuation reflects – in some way – the intended *focus* of an utterance". This position is also taken by Jackendoff [1972, p. 229] and Gardent and Kohlhase [1996]. In the present work, we use **contrast** (instead of focus) for the semantic effect associated with prominence *on a word*. In the above example, the prominent word *five* is in contrast to *three*. On the other hand, we may call the complement of a contrast **background**.

Projection of Contrast

The notion of contrast is complicated because it can 'project' to a more complex linguistic structure.[7] For example, Ladd [1996, p. 162] distinguishes between 'narrow focus' corresponding to our contrast on a phonological word and 'broad focus' spanning a more complex structure such as "*five francs*" in the following example [Ladd, 1996, (5.4)]:

(26) I didn't give him a sandwich, I gave him five **francs**.

Even for the case where only *francs* is prominent, the phrase "*five francs*" (its interpretation) is in contrast with "*a sandwich*" in this example. As stated in Ladd [1996, p. 161], this is a phenomenon distinct from word-level 'contrast'.[8] Accordingly we may specifically distinguish **projected contrast** from (word-level) contrast.

As Krifka [1992] points out, and Halliday [1967] and Steedman [1991a] state more explicitly, there appears to be a connection between information structure and contrast. One complication arising from contrast projection is that a rheme may coincide with a broad focus or a single-word contrast. This is the intersection of (possibly projected) contrast and information structure.

Contrast within Theme and Rheme: Two-level Analysis

We now demonstrate that the notion of contrast (at the word level) needs to be considered independent of information structure.

Steedman [1999, (31)] provides the following example involving separation of the two notions.

(27) *Q*: I know that Mary envies the man who wrote the musical.

But who does she **admire**?

A: [Mary **admires**]$_{Theme}$ [the woman who **directed** the musical]$_{Rheme}$.
L+H* contrast — H* contrast

Note: 'L+H*' and 'H*' are argued to mark theme and rheme, respectively [Steedman, 1991a].

[7] Winkler [1997] is a good review on focus projection especially in connection to syntactic structure. Another recent work is Gussenhoven [1999].

[8] Hockey [1998, p. 226] discusses the role of amplitude and duration in marking the entire span of rheme (her 'focus') in English and Hungarian.

In addition to the projection of contrast from *directed* in the rheme, there is another contrast *admires* in the theme.[9] The two instances of contrast above receive distinct pitch accents corresponding to theme and rheme [Steedman, 1991a]. Halliday [1967] too discusses the two levels: 'information structure' [p. 199] roughly corresponding to our information structure and the distinction between 'new' and 'given' [p. 204] (corresponding to our contrast/background).[10]

Vallduví and Vilkuna [1998, p. 85] also distinguish 'rheme' ('focus' in Vallduví [1990]) from 'kontrast' (their new terminology). But their kontrast is a notion associated with a constituent, and thus is intermediate between our rheme and our (word-level) contrast. Their analysis would make the projection problem of contrast unnecessarily complicated.

The distinction between rheme and contrast is not always understood as the above. Since the term 'focus' is overloaded, analyses often mix the two notions. For example, Pulman [1997, p. 74] uses the term 'focus' citing the Prague School (some work preceding Sgall et al. [1986]), Selkirk [1984], and Rooth [1985]. But this introduces a complication because 'focus' of the Prague School basically corresponds to our rheme, and that of Selkirk [1984] and Rooth [1985] corresponds to our contrast (and its projection). The subsequent description of broad and narrow foci does not illuminate the discussion. He distinguishes narrow and broad foci based on constituent size, which is misleading. It is not clear why his approach works for the case of broad focus without discussing focus projection.

Property of Rheme: Projection from a Contrast

We have shown above that a theme *can* contain a contrast. But, as seen in many earlier examples, a theme does not always contain a contrast. But a rheme is always projected from a contrast [e.g., Jackendoff, 1972, p. 229; Rochemont, 1998, p. 337]. The main point in this subsection is to examine (20*c*) of the main hypothesis: "a rheme is always contrastive". At this point, let us recall Rochemont's [1986, p. 52] distinction between 'presentational' and 'contrastive' rhemes. This suggests that there is a rheme that is *not* contrastive. For example, in the example (18) repeated below, the response may be considered to include a presentational rheme (without further

[9] Prevost [1995, p. 67] calls the contrasts in a theme and a rheme 'theme-focus' and 'rheme-focus', respectively.

[10] Fries [1994, p. 234] calls Halliday's information structure and given/new distinction thematic structure and 'information structure', respectively. Brown and Yule [1983, Chapters 4 and 5] use theme/rheme in the sense of Halliday [1967], and information structure for Halliday's [1967] given/new.

contrasting information).

(28) *Q*: Who did Felix praise?

A: [Felix praised]$_{Theme}$ [**Donald**]$_{Rheme}$.

But Jackendoff [1972, p. 246] observes that a negative response such as the following is possible in the same context.

(29) *Q*: Who did Felix praise?

A: [Felix praised]$_{Theme}$ [**nobody**]$_{Rheme}$.

This suggests that there is no presupposition for the existence of an individual who was praised by Felix. The rheme in (28), *Donald*, is in contrast at least with *nobody*. Thus, as argued by Büring [1997b, p. 40], it is possible to abstract away from Rochemont's distinction.

Alternative Semantics

There is a general way to capture the semantics of contrast, i.e., Alternative Semantics [Rooth, 1985]. The idea is that the notion of contrast can be defined by considering an 'alternatives' set where the elements in contrast are marked. For example, for an alternatives set {"[John]$_c$ is tall", "[Mary]$_c$ is tall"}, *John* is in contrast with *Mary*. In other words, the alternatives set is obtained by making an appropriate substitution in the contrastive element. The selection of these contrastive elements can span an arbitrarily long distance across utterances in a discourse (or in the context in general). Therefore, the exact nature of how such contrast is analyzed is obviously beyond the grammar for the utterance level [discussion in Rooth, 1992]. Partee [1999, p. 214] comments on this point that Rooth [1992] is an extreme of degrammaticalized analysis of contrast (her 'focus'). But, since contrast spans across discourse (and situational context as well), it is natural and necessary for a theory of contrast to have a degrammaticalized component.

An advantage of Alternative Semantics is that it can be applied to a projection of contrast in a general way. Now that we consider a rheme as a projection of contrast, the rheme can be seen in terms of the alternatives set associated with the theme [Steedman, 1999, Section 5.3]. It is this view that a property of rheme is contrastiveness in a general sense (20*c*).

But Alternative Semantics does not automatically solve the problem of identifying rheme through contrastiveness. It is a general framework that can be used for accounting for the semantics

of (narrow and projected) contrast *and* that of rheme. To complete the analysis of contrastiveness, we must have a mechanism of identifying the alternatives set, which is extremely difficult to formalize and implement. On the other hand, in order to apply it to the Identification Problem of information structure, we also need to know the relation between rheme and theme. This latter point is not clearly stated in Rooth [1992, p. 84] when he argues that Alternative Semantics can be applied to the analysis of question-answer context. We will address the relation between theme and rheme in the next chapter.

Dynamic Semantics: Connection to Procedural Accounts

Alternative Semantics can also be connected to procedural ideas through 'dynamic semantics' [Stalnaker, 1978, (an earlier work)]. In this tradition, the meaning of an utterance is considered as a potential to change context. The representation of context differs among proposals. For example, Stalnaker [1978, p. 321] has it as a set of propositions. Heim [1982] has it in terms of files in File Change Semantics (FCS). Kamp [1981] has it as Discourse Representation Theory (DRT). More recent work relevant to our case are: Asher [1993] and Hendriks and Dekker [1996].[11] The idea of dynamic semantics is adopted in recent analyses of information structure McNally [1998, Section 3.2] and Steedman [1999, Subsection 5.3.1].

In this subsection, we have separated the notion of contrast from information structure, and observed a requirement that a rheme (semantic unit roughly corresponding to a constituent) be projected from a contrast (word-level property). While this identifies a property of rheme useful for theoretical analysis, its formalization and implementation for the Identification Problem remains open.

2.3.3 Information Structure and Linguistic Form

This subsection explores direct linguistic marking of information structure and argues that information structure cannot be obtained from linguistic form alone.

[11] Atlas [1991] may also be included in this group.

Linguistic Marking of Information Structure as a Matrix-level Phenomenon

It is generally accepted that linguistic marking of information structure exists [e.g., Vallduví and Engdahl, 1996]. But very little has been said about properties generalizing various distinct forms of information-structure marking. One reason may be that linguistic marking of information structure is arbitrary [Prince, 1998, p. 282]. As a tool to analyze linguistic marking of information structure, I would like to examine the following hypothesis:

(30) (hypothesis) Linguistic marking of information structure is a matrix-level ('root') phenomenon, i.e., *non-recursive*.

Naturally, this view is consistent with most proposals of information structure including our main hypothesis (20), which is non-recursive (the idea of recursive information structure is reviewed in Subsection 2.3.4). This is in contrast with the use of, say, definite determiner, which is recursive along linguistic structure. A consequence of the above hypothesis is that Levinson's [1983] complaint about lack of projection analysis for information structure is not actually applicable to information structure itself. But it may apply indirectly to information structure through other types of linguistic marking, e.g., definiteness. Let us now examine some examples of information-structure marking discussed in the literature.

First, it is generally held that prosodic structure is non-recursive [Selkirk, 1984; Pierrehumbert and Beckman, 1988; as reviewed by Ladd, 1996, p. 238]. If certain pitch accents, e.g., L+H* and H* as shown in Pierrehumbert and Hirschberg [1990] are associated with theme and rheme [Steedman, 1991a], respectively, such a pitch accent, associated with a word, may recursively project through linguistic structure. But the prosodic units projected from pitch accents do not embed another unit, as formally shown in Steedman [1999, Section 5.6]. Thus, there is no conflict between prosodic structure that marks information structure and the hypothesis (30). Note that Ladd [1996, p. 245] himself argues for recursive prosodic structure, but this means that prosodic structure can recursively associate with linguistic structure and is not a position contrary to what has been said above.

Although English does not have an extensive set of direct information-structure markers (compared to languages like Catalan), there are many special constructions whose functions have been discussed in relation to information structure. Among these, inversion cannot be embedded while subordinators *since*/*because* can (examples and more details in Subsection 3.3.2). The hypothesis

predicts that the former can be but the latter is not a direct information-structure marker.

A strong support for non-recursiveness of information-structure marking comes from particle use in Japanese. While the detailed discussion awaits Chapter 5, it is illustrative to point out that *thematic* function of particle *wa* is only available at the matrix level. In addition, a constituent extracted from an embedded level can also be marked in this way. Direct information-structure marking is a basis for our evaluation method. Later, we use particle choice in Japanese in the evaluation process.

Discussion of languages other than English and Japanese is beyond the scope of the present work, but I am very much interested in analyses for or against the hypothesis (30).[12] The prediction of the hypothesis is that recursive linguistic marking is not a direct information-structure marking. For example, is it really the case that a clause (IP), regardless of matrix or embedded level, is 'configured' according to information structure, e.g., in Russian [King, 1995]?

The theme-first principle is certainly a controversial one as a linguistic marking of information structure [Lambrecht, 1994, Section 4.5, for a detailed discussion]. There are some experimental results showing that passivization is associated with information-structure effect [Most and Saltz, 1979]. But the current work is not committed to accept that theme-first principle applies universally, or even language-specifically, as information-structure marking. But we do consider a certain cases of preposing, e.g., utterance-initial modifier, as a contextual-link marker based on de Swart's [1999] analysis. More detail is described in Subsection 3.3.1.

Information Structure cannot be Recovered Solely from Linguistic Form

We have seen above that information structure may be marked linguistically. In this connection, Vallduví [1990, p. 6] states that "[i]nformational understanding and the packaging instructions that encode it must obviously be recoverable from the overt structure of any language". This is a very strong statement suggesting that the linguistic structure completely identifies the information structure. We have to disagree with this position following Brown and Yule [1983, p. 188] who state that linguistic form alone is not enough to identify information structure.

The following example from Steedman [1991a, p. 285] demonstrates that exactly the same

[12] Kiss [1995] discusses a number of languages in relation to the idea of 'discourse configurationality'.

linguistic forms including prosody may have distinct information structures depending on the context.

(31) *a*. [They are]$_{Theme}$ [a good source of **vitamins**]$_{Rheme}$. (in response to "What are legumes?")

b. [They are a good source of]$_{Theme}$ [**vitamins**]$_{Rheme}$.

(in response to "What are legumes a good source of?")

Similarly, in Japanese, exactly the same utterance including phonological marking can be ambiguous with respect to information structure (assuming that there is no phonologically marked distinction between theme and rheme, and particle *wa* can be used for a theme and contrastiveness, as will be discussed in Chapter 5). Here, the following grammatical labels are used TOPic, CONTrastive, ACCusative.

(32) *Q*: "What did Ken and Naomi do?"

A: [**Ken**-wa]$_{Theme}$ [**banana**-o tabeta]$_{Rheme}$.
Ken-TOP/CONT banana-ACC ate
"Ken (but not Naomi) ate a/the banana."

(33) *Q*: "Between Ken and Naomi, who ate the banana and the mango?"

A: [**Ken**-wa]$_{Rheme}$ [**banana**-o tabeta]$_{Theme}$.
Ken-CONT banana-ACC ate
"Ken (but not Naomi) ate the banana."

Vallduví's [1990] position indeed suggests that information structure *cannot* be affected by the context. This reduces identification of information structure to parsing. Possibly for this reason, Vallduví [1990] does not address the problem of identifying information structure in texts, and only works on examples that do not show the problem of ambiguous information structure. Nevertheless, Vallduví and Engdahl [1994, p. 531] state that "no syntactic constituency is required for any informational unit as long as inheritance of INFO-STRUCT values proceeds in the permitted fashion". This seems to discount Vallduví's [1990] position that information structure can be completely derived from surface structure.

We have seen that information structure cannot be identified from linguistic form alone. We have also noted that linguistic marking of information structure is relatively impoverished in written English. But it seems that linguistic communication in written English does not suffer from

potentially 'defective' information structure. In the next chapter, we develop the main hypothesis (20) in terms of properties including definiteness, which is systematically employed in English.

2.3.4 Internal Organization of Information Structure

This subsection examines different ways of organizing components of information structure: i.e., recursive structure, binomial and trinomial partition, and graded multiple partitions.

Recursive Information Structure

Our main idea about information structure (20) assumes that it is non-recursive. We have also stated a hypothesis, (30), that linguistic marking for information structure is matrix-level. But some argue that information structure is recursive [i.e., Kiss, 1987; Hoffman, 1995, p. 145; Partee, 1996, p. 77].

Let us examine the following example from Partee [1996, (31), p. 82]:

(34) What convinced Susan that our arrest was caused by **Harry** was [$_{FOC1}$a rumor that [$_{S3}$ someone had [$_{FOC3}$ witnessed Harry's confession.]]]

Partee analyzes the structure for this utterance in the following way:

(35)
$$\frac{\dfrac{\underline{TOP2 \quad FOC2}}{S2}}{TOP1} \qquad \frac{\dfrac{\underline{TOP3 \quad FOC3}}{S3}}{FOC1}$$
$$\overline{\qquad\qquad S1 \qquad\qquad}$$

Partee [1996, p. 67] is specific about her 'topic' and 'focus' are Praguian [Mathesius, 1975; Sgall et al., 1986, etc.]. But there are two points we may argue against recursive information structure. First, there is no standard way to identify information structure recursively, cf. 'question test' [Sgall et al., 1986], which is non-recursive. Second, commonly observed direct information-structure marking is non-recursive, as we have seen for the hypothesis (30). With a focus on the contextual status of a clause, Partee's [1996] analysis is more in line with formal analyses of 'presupposition' [e.g., Beaver, 1997]. The problem of presupposition projection is widely discussed in relation to linguistic structure [e.g., Gazdar, 1979; Karttunen and Peters, 1979]. Once contrastive elements [Rooth, 1985] involved in the utterance are identified, two-level analysis (page 2.3.2) of

Steedman [1991a] seems sufficient for the above example. A convincing demonstration of recursive information structure would identify a test comparable to question-test for arbitrary linguistic structure or find recursive linguistic marking that directly marks information structure.

One motivation often found behind recursive information structure is to identify information structure with tripartite quantification structure [Partee, 1996] (also to some extent in Partee [1999]). A quantification structure has the form $Quantifier(Restrictor, Scope)$ commonly used in formal semantics. Applying this connection, Szabolcsi [1983b], Rooth [1985], and Sgall et al. [1986] argue that information structure is truth-conditional.[13] Szabolcsi [1983b, Section 3.1] explicitly states exhaustivity as the cause of this point, and the same situation is implicit in Sgall et al. [1986, p. 62] as well. For this matter, I follow Horn [1981, p. 132] and Vallduví [1990, Section 7.1] in that exhaustivity is conversational implicature [Grice, 1975] (for English, not a direct counterexample to Hungarian examples in Szabolcsi [1983b]). Kuno [1972] also states the exhaustivity effect for a Japanese particle *ga* , but rejected by Shibatani [1990, p. 271] as epiphenomenal (more discussion in Chapter 5).

Binomial Partition

The rest of this subsection reviews some proposals on non-recursive information structure. The classic partition of information structure is the binomial one, e.g., early Prague School [Mathesius, 1975], and [Chomsky, 1971; Jackendoff, 1972; Halliday, 1967; Steedman, 1991a]. But its simplicity is also associated with some problems. For example, Vallduví [1990] argues that neither topic-comment nor focus-background can properly represent the partition commonly observed in natural data. In general, the complexity of realistic texts poses a challenge to binomial partition.

First, let us consider the following example from Vallduví [1990, (42), p. 55]:

(36) *Q*: What does John drink?

A: [John]$_{Link}$ [drinks]$_{Tail}$ [beer]$_{Focus}$.

Vallduví [1990] proposes a trinomial partition of information structure such that our theme is further divided into two subcomponents. His 'link' and 'tail' jointly correspond to our 'theme', and 'focus' to our 'rheme'. His argument, then, is that focus-background partition would result in

[13]Relevant other papers include: Szabolcsi [1981], Szabolcsi [1983a], Erteschik-Shir [1997], and Erteschik-Shir [1998], Lee [1993], and Jäger [1999].

"focus = *beer*" and "background = *John drinks*" and topic-comment partition would result in "topic = *John*" and "comment = *drinks beer*", and that neither of them capture the information structure properly. While the focus-background partition directly correspond to the informational division of the question, topic-comment structure (as presented by Vallduví) does not. There are two points to make here. One is about semantic types for referent, and the other is about accommodation of a theme. In the following, we discuss these points in turn.

Most studies of reference in relation to information structure deal only with (discourse) referents [Karttunen, 1976] of the individual type, corresponding to referential NPs. For example, Reinhart [1982, p. 5] limits the discussion of theme to NPs. This also applies to Vallduví [1990, Chapter 4] adopting an analogy of File Change Semantics (FCS) [Heim, 1982], and Hoffman [1996] adopting a version of Centering theory [Grosz et al., 1995]. But a question like (36*Q*) partitions information where subject-verb sequence is a unit of information, as observed by Steedman [1991a, p. 260]. In general, any linguistic units that are extractable or can be coordinated may well be an information-structure unit [Steedman, 1996]. In accordance to this observation, Vallduví and Vilkuna [1998, p. 82] seem to have dropped File Change Semantics in favor of a more general extension of Discourse Representation Theory [Kamp, 1981; Heim, 1982], an extension due to Asher [1993] to deal with 'abstract objects'. A consequence of this more general view of referent allows us to analyze "*John drinks*" in (36*A*) as a unit of information structure even though it is not traditionally considered a constituent. Thus, as long as we have a means to account for such constituents, e.g., Combinatory Categorial Grammar [Ades and Steedman, 1982], this type of division is not a problem for binomial partition. Then, we need some other explanation for separating *John* in (36*A*) as Vallduví's [1990] 'link'.

The other point is the possibility of accommodating a theme. Although a direct response to a question such as (36*A*) is what we usually expect, we may also encounter unexpected responses, including completely irrelevant ones. Note that question test as a tool to identify the information structure of a response is only good for a direct response. But even for non-direct response, we will find a certain information structure depending on the context. Let us consider the following example with ambiguous information structure.

(37) *Q*: Who did Felix praise?

A_1: [Felix praised]$_{Theme}$ [Donald]$_{Rheme}$. (direct response)

A_2: [Felix]$_{Theme}$ [praised Donald]$_{Rheme}$.

A_3: [Felix praised Donald]$_{Rheme}$.

As long as the theme is linked to the context (including the null case) and the complementary rheme is a projected contrast, any of the above information structures are possible, which is consistent with our main hypothesis (20). Although the contextual force of a question is very strong, it cannot completely specify the response. There is a room for the respondent to *accommodate* a distinct theme (see Subsection 2.3.1 for accommodation). Thus, theoretically, the following ambiguity for (36) is possible.

(38) *Q*: What does John drink?

A_1: [John drinks]$_{Theme}$ [beer]$_{Rheme}$.

A_2: [John]$_{Theme}$ [drinks beer]$_{Rheme}$.

Note that the above analysis observes an ambiguity, but *not* a coexisting parallel structures, as in Vallduví [1990]. Without additional contextual information, (A_1) is the most likely response. (A_2) may still be available if, e.g., the context is specifically about John and elaborating various properties of John. In summary, we accept the possibility of information structure like (38A_2), but it can be analyzed within the binomial partition approach.

There is another type of problem for binomial partition. Let us take a look at another example from Vallduví [1990, (56a)], assuming a question "What did the farmer do with the broccoli to the boss?".

(39) [The farmer]$_{Link}$ [already **sent**]$_{Focus}$ [the broccoli to the boss]$_{Tail}$.

In this case, the theme ($Link + Tail$) is *discontiguous*. Related examples are found in Büring [1997b, (4,5), p. 3].

(40) *i.* Guess who went to the central station after Smith left the pub.

ii. After Smith left the pub, [**Jones**]$_F$ went to the central station.

Again, the theme (i.e., the complement of *John*) is discontiguous. This case is a problem for binomial partition that assumes complete syntax-semantic parallelism. But it is still possible to construct a semantic unit covering the discontiguous themes. We will explore a principled method to link such a semantic structure with syntax in the next chapter.

Trinomial Partition

In an attempt to avoid the problem with binomial partition, Vallduví [1990], Büring [1997b], and Hoffman [1995] adopt a trinomial partition.[14] For Vallduví [1990] and Büring [1997b, p. 54], it is a way to mediate both topic-comment and background-focus partitions, also suggested by Jacobs [1986, p. 104].

Vallduví [1990, p. 57] proposes a trinomial partition of information structure "*Link – Focus – Tail*". This corresponds to our "*Theme – Rheme – Theme*" case as Vallduví's [1990] 'link' and 'tail' are in contrast with his 'focus', e.g. (36). But this partition does not generalize to cases such as the following [p.c., Steedman 1998]:

(41) *Q*: I know what team Fred wants to win the Cup, but which team does Alice want to lose which contest?

A: [Alice wants]$_{Theme}$ [**Australia**]$_{Rheme}$ [to lose]$_{Theme}$ [the **Ashes**]$_{Rheme}$.[15]

Hoffman [1995, Chapter 5] proposes a slightly different trinomial partition "*Topic – Focus – Ground*". But she combines 'focus' and 'ground' as 'comment' in contrast to her 'topic', and only considers contiguous partitions between 'topic' and 'comment'. Thus, it is not a solution to the problem of discontiguous information structure. Similarly, Fries [1994, p. 234] divides rheme into N-Rheme (last constituent) and the rest (assuming Halliday's theme).

We have separated the notion of 'contrast' from information structure, and have accepted that contrast can appear freely within a theme or a rheme [Halliday, 1967; Steedman, 1991a]. Thus, partitions between 'contrast' and 'background' within a theme or a rheme can be accounted for without problem. This approach can cover Hoffman's [1995] and Fries's [1994] analyses more generally. In summary, trinomial approaches do not seem to be a solution to the discontiguity problem.

Another question about these trinomial partitions is how can we *define* such further divisions of theme and rheme. It is not entirely clear how the two theme components in the above examples are distinct in a systematic manner. Although Catalan seems to split a theme across the rheme, such a distinction between the two theme components seems language-specific and does not show up in other languages in a systematic way. Hendriks and Dekker [1996, p. 350] also argue against the

[14] A similar observation is made in Foley [1994, p. 1680], which is a fairly extensive encyclopedia entry.

[15] With or without L+H* on the themes.

status of ‘tail’ [Vallduví, 1990] that it complicates analysis and processing of information structure (they show an example to demonstrate such a complication).

Communicative Dynamism

Another, more complicated approach is Communicative Dynamism (CD) [Firbas, 1964], developed within the Functional Sentence Perspective (FSP) approach. CD is a degree of contribution to the development of the communication by sentence elements. Firbas [1964, p. 272] states that Communicative Dynamism is not dependent on ‘unknown’ vs. ‘known’. Communicative Dynamism is by definition a *gradable* concept. While it may well be the case that information ordering is graded, it is hard to grasp the idea cross-linguistically in terms of observable phenomena. While information ordering may be faithfully realized in a language like Czech, it is not readily observable in other languages to the level we can generally see for the contrast between theme and rheme. Second, there is no generally accepted ‘semantics’ for such grading. Finally, in relation to the first two points, it is extremely hard to evaluate. Thus, CD is not appropriate for the current purpose. Note that we do not deny the possibility of multiple divisions. There may be factors that are beyond the current scope and have not been clarified in the previous work.

Summary

In any of the reviewed cases, there are some kinds of problems. Since additional complexities associated with multiple partitions do not solve the problem as a whole, we assume the classic and simplest case, binomial partition (Ockham’s razor). The problem with binomial partition, namely discontiguous information structure is addressed in detail in the next chapter.

2.4 Previous Proposals for Identifying Information Structure

There are several proposals directly addressing the Identification Problem [Kurohashi and Nagao, 1994; Hahn, 1995; Hajičová et al., 1995; Styś and Zemke, 1995; Hoffman, 1996; Komagata, 1998a]. This subsection reviews these proposals. We also discuss application of information structure to natural language generation at the end because this computational application too involves the Identification Problem

While each one of these approaches has particular problems of its own, there are more fundamental problems shared by these approaches: namely, limited coverage, lack of evaluation, and unclear theory-procedure relation. The following review pays close attention to these points.

Kurohashi and Nagao, 1994

The main point of Kurohashi and Nagao [1994] is that 'discourse structure' in Japanese in the sense of Grosz and Sidner [1986] and Mann and Thompson [1988] can be identified through surface information. Discussion of their main goal is naturally beyond our scope, but we must investigate the component involving the notion of information structure, namely the problem of identifying information structure (their 'topic'/'non-topic') in Japanese. Their method basically consists of observing the distribution of particles *wa* (so-called 'topic marker') and *ga* (nominative marker) without using contextual information.[16] Analysis and the use of these particles are important aspects of text analysis in Japanese, and we follow this direction. But the functions of these particles are complex and we cannot simply say that *wa* and *ga* mark theme and rheme, respectively (see Chapter 5 for more detail). Moreover, there are utterances lacking these particles (as arguments can be dropped in Japanese), still with clear information structure depending on the context.[17] Kurohashi and Nagao [1994] are also limited in explicating the theory-procedure relation with respect to the description of (partial) relation between Japanese particles and information structure. Finally, their analysis only contains a language-specific element of information structure. Since

[16] Kurohashi and Nagao [1994] also apply a few additional structural cues, which are not clear from the paper.

[17] The following example demonstrates that information structure is not necessarily marked by *wa* or *ga* (grammatical labels: TOPic, ACCusative, and Question):

(1) *Q*: Ken-wa Montana-to Oregon-de nani-o sita-no?
Ken-TOP Montana-and Oregon-at what-ACC did-Q
"What did Ken do in Montana and Oregon?"

A: [**Montana**-de]$_{Theme(contrastive)}$ [**sukii**-o site,...]$_{Rheme}$
Montana-at ski-ACC did
"He **skied**$_{H^*}$ in **Montana**$_{L+H^*}$,..."

(2) *Q*: Ken-wa doko-de sukii-to sukeeto-o sita-no?
Ken-TOP where-at ski-and skate-ACC did-Q
"Where did Ken ski?"

A: [**Montana**-de]$_{Rheme}$ [**sukii**-o sita.]$_{Theme}$
Montana-at ski-ACC did
"He skied in **Montana**$_{H^*}$."

our position is that the notion of information structure applies cross-linguistically and that it contains universal elements, the approach of Kurohashi and Nagao [1994] does not apply to analysis of other languages. Since their goal is identification of discourse structure, no direct assessment of the information structure is provided.

Hajičová and others, 1995

Following the tradition of the Prague school, e.g., Sgall et al. [1986], Hajičová et al. [1995] proposed an algorithm to identify information structure (their 'topic' and 'focus').[18] Their algorithm is an implementation of a series of theoretical works, it addresses the theory-processor relation more strongly than others.

But there still remains a question about theory-processor relation. Although they discuss a contextual factor in terms of their 'Contextual Bound' (CB) and 'Non-Bound' (NB) (p. 25 in Subsection 2.3.1), their algorithm actually assigns a CB/NB status through structural analysis [p. 89-90], as seen below.

(42) (a) After the dependency structure of the sentence has been identified by the parser, so that also the underlying dependency relations (valency positions) of the complementations (to the governing verb) are known, the verb and all the complementations are first assumed to be NB, i.e., to belong to the focus, which we denote by f.

(b) (omitted: three conditions for the case where the verb is rightmost)

(c) If the verb does not occupy the rightmost position, then:

(ca) the verb itself is understood as t [topic], if it has a very general lexical meaning (see above), or as f if its meaning is very specific, or else as ambiguous (t/f);

(cb) the complementations preceding the verb are denoted as t, with the exception of an indefinite subject and of a specific (i.e., neither general nor indexical; see above) Temporal complementation; either of the latter two is characterized as t/f;

(cc) (omitted: ten more conditions for various cases)

The condition (cb) thus predicts that a definite subject is a topic as they do in their example (3) "The neighbor met him yesterday" [p. 91]. But, as the following example shows, a definite subject

[18] Two closely related papers are Hajičová [1991] and Hajičová et al. [1993].

with a verb not at the rightmost position can be a rheme.

(43) *Q*: Who met him yesterday, the neighbor or the gardener?

A: The neighbor met him yesterday. Hajičová et al. [1995, (3), p. 91]

"*The neighbor*" in (43*A*) must be analyzed as the theme (or its part) of the utterance. As we have discussed earlier, linguistic form alone cannot fix the information structure.

We agree that certain linguistic marking such as definiteness plays an important role in identifying information structure, and we will use that property. But we cannot underestimate the contextual effect. The algorithm depends too much on structural and lexical information and has very little contextual information in it. The coverage of the algorithm is limited to simple sentences in English. They comment on the extension of the proposal to more complex constructions [p. 93]. But their algorithm [p. 89-90] is already a sequence of *seventeen* conditional statements. Even if it can be extended to more complex cases, it would be hard to see the underlying generality. Finally, no evaluation is discussed.

Hahn, 1995

Hahn [1995] argues that thematic progression [Daneš, 1974] can be formalized, be applied to real-world texts, and provide a means to view text coherence. The implementation consists of partial parsing, processing of 'frame' representation including relations between entities, and processing of theme/rheme according to how the theme of an utterance is connected to an antecedent in the context. The system works on realistic data taken from computer-related journals. This approach has a strength in dealing with real data, unlike many other approaches discussed here. The contextual information is well handled as well.

The problems with this approach include the following. Although Hahn [1995, p. 215] argues that full parsing is infeasible for such a task, there is a cost associated with adopting partial parsing. For example, the information obtainable from complex NPs can be misused. In addition, special constructions such as 'cleft' and 'topicalization' cannot be identified without ad-hoc treatment. A systematic analysis of sentence construction requires full parsing. Furthermore, the system appears to be limited to individual-type themes. It could not identify a theme such as "*Felix praised*" (18) seen earlier. There is little discussion about how his implementation is related to a theory of information structure. Again no evaluation method is provided.

Styś and Zemke, 1995

Styś and Zemke [1995] proposes a method to improve the quality of English-Polish machine translation. Their point is that word order in Polish depends on salience and this information can be obtained in English through linguistic analysis including Centering theory [Grosz et al., 1995, as well as much earlier work cited there]. Their approach is actually more in line with Communicative Dynamism (CD) [Firbas, 1964] because their theory adopts 'gradation' of salience, not binomial contrast between theme and rheme. They obtain such results by applying gradation to Centering analysis, utterance construction type, definiteness, constituent length, etc. There is no doubt that information structure is related to most, if not all, of these properties. But the use of graded salience makes evaluation of this approach extremely difficult. Accordingly, no evaluation is discussed. Furthermore, an ad-hoc weighting of these properties does not seem to be well-founded in terms of available theories of information structure. Styś and Zemke [1995] mainly deal with the transitive construction including clefted cases [Section 5 (Conclusion)], and need to extend their limited coverage for a more realistic set of data.

Hoffman, 1996

Hoffman [1996] proposes a method to improve the quality of English-Turkish machine translation through the use of information structure. The key element of the proposal is identification of information structure in English through a combination of contextual information and linguistic form, including Centering analysis [Grosz et al., 1995]. This in principle combines the strengths of Hajičová et al. [1995] and Hahn [1995]. Hoffman [1996] characterizes theme (her 'topic') in terms of referential preference based on a version of Centering theory [Grosz et al., 1995], and rheme (her 'focus'), in terms of 'discourse-newness' and 'contrastiveness', corresponding to the distinction of Rochemont [1986].

The main contribution of the proposal is the following two algorithms:

(44) Topic algorithm:

a. Choose Cb (if available) as the topic.[19]

[19]Cb is the highest-ranked referent in the reference list (Cf) of the previous utterance also present by the current utterance.

b. Choose the first entity in the Cf list (if available).[20]

c. Choose a situation-setting adverb (if available).

d. Choose the subject.

(45) Focus algorithm:

a. Choose a discourse-new

b. Choose a contrastive element

Use of these algorithms is demonstrated in Hoffman's (5), which can be shown as follows ('topic' and 'focus' are indicated with the rule that is used to identify it):

(46) *i.* Pat (*Focus* (45*a*)) will meet Chris (*Focus* (45*a*)) today (*Topic* (44*c*)).

ii. There is a talk (*Focus* (45*a*)) at four (*Topic* (44*c*)).

iii. Chris (*Focus* (45*b*)) is giving the talk (*Topic* (44*a*)).

iv. Pat (*Topic* (44*b*)) cannot come (*Focus* (45*a*)).

One of the weaknesses of Hoffman's algorithms is its lack of connection to a theory of information structure. For example, it is not at all clear why *today* in (*i*) *must* be the topic. Information structure is characterized in terms of combination of referential status and other properties on the involved components. It does not capture the *relation* between theme and rheme in a way we are interested in.

Another problem is its limitation in recognizing 'referents' corresponding to complex linguistic structures. In the following example similar to the one given in her paper, the theme algorithm will pick up "Chris" as the theme of (*ii*), among the possible candidates underlined below.

(47) *i.* Chris will give the talk. [Chris, talk]

ii. But, Pat doesn't think that Chris will give the talk. [Pat, Chris, talk]

But the clause "that Chris will give the talk" is most likely the theme of (*ii*).

Hoffman [1996] tackles cases involving adverbs and complement clauses, but demonstrates her algorithm only for a few prepared texts, not realistic data. She also mentions the role of INFERRABLE, which is a critical element in identifying information structure, but does not specify how to identify them. Finally, there is no evaluation is presented.

[20] Cf is the list of discourse referents in the utterance.

Komagata, 1998

In the precursor to the current work [Komagata, 1998a], I proposed a theory of information structure and an algorithm to identify information structure to be used for a Computer-Assisted Writing system. The goal of the system is to detect text readability with respect to information structure. The mechanism of the identification process is that theme has a property 'contextual link' , which is realized as either discourse-old or linguistically-marked inferrables like Hoffman [1996]. Then, the theme-rheme structure is observed as the last semantic composition.

Some problems with this work are that the theory is overly simplistic. For example, the only considered linguistic marking for inferrable was definiteness. The theory assumed binomial partition of information structure where theme and rheme are contiguous, which is not necessarily the case (see Subsection 2.3.4).

In an attempt to address lack of evaluation in previous work, I proposed a method based on text readability. Assuming that 'theme first' preference is at work in written English (following Mathesius [1975, p. 81], [Halliday, 1967], and [Kuno, 1978] in a slightly weaker form), I adopted the FSP-type approach that a pattern of "*Theme – Rheme*" is more readable than one of "*Rheme – Theme*". Although certain effects have been observed, the paper did not provide an objective way of measuring the effects. As mentioned in Subsection 2.3.3, 'theme first' preference is controversial. The present work does not assume this position in any strong form.

Identification Problem in Natural Language Generation

Natural language (NL) generation is one area where theories of information structure are successfully applied. This involves contextually appropriate generation of intonation in English [Prevost and Steedman, 1993; Prevost, 1995; Prevost, 1996] and that of word order in Turkish [Hoffman, 1994; Hoffman, 1995; Hoffman, 1996].[21] Such approaches are possible due to direct linguistic marking of information structure. Although the Identification Problem in its original form is not a part of NL generation, there are some connections between them.

First, an assumption common to the above-mentioned NL-generation approaches (except for Hoffman [1996], which also presents an information-structure identification algorithm) is that the information structure is available for each utterance in the given contexts. Prevost [1995] also

[21] Günther et al. [1999] is another example.

works on short discourse, but his examples are limited to the cases where the subsequent utterances share the same theme as the first one. Thus, while usefulness of information structure for NL generation tasks is demonstrated, the question about how information structure generally works in texts is left unanswered.

Now, let us consider the case of generating realistically complex texts. Is the information structure readily available for each utterance? Modern NL generation systems have planning process at the level of content generation as well as surface generation, e.g., McKeown [1985] and Prevost [1995]. Since a typical planner involves propositions as a unit of processing, it may be able to determine the information structure of a complex utterance involving a subordinate clause based on how the utterance is derived in connection to the context. But, since an information-structure division generally corresponds to units smaller than a clause, a process of identifying information structure is still needed.

For the case of a NL generation module as a part of a machine translation system, it is in general impossible that an automated system can derive the 'intention' of the writer of the source text, cf. planning in NL generation. In fact, most of the currently available systems simply transfer either isolated syntactic and/or semantic structures between the corresponding utterances. Thus, while a generation module requires a solution to the Identification Problem, the current solutions to NL generation problems involving information structure do not solve the Identification problem.

Note about Evaluation Methodology

As we have seen above, evaluation is a missing component in all previous proposals for the Identification Problem. Let us briefly discuss the methodology we might use for this purpose. One possible direction is to identify a non-linguistic observable phenomena practiced in, e.g., psycholinguistics. They control referential status of physical objects and observe the relation with linguistic expression [Arnold et al., 1997]. But, since we want to evaluate identification processes, this approach does not seem to be applicable to our case. Another technique is to directly observe processing load through eye tracking [e.g., Rayner and Pollatsek, 1987 (a review)]. This seems like a promising possibility, but is beyond the scope of the current work. The present work pursues a purely linguistic way of evaluation in the remainder of this thesis.

2.5 Summary

The main conclusion of this chapter is that the Identification Problem still remains wide open. In the previous section, we identify problems specific to the computational approaches to the Identification Problem. But, more importantly, this group of work lacks the essential properties required for a solution to the problem, i.e., realistic coverage, an evaluation method, and a clear theory-procedure relation.

On the other hand, previous theories of information structure reviewed in Section 2.3 are mostly indifferent to the Identification Problem. Although various properties related to information structure have been investigated, previous theories do not delineate the properties of theme and rheme and the relation between theme and rheme as pursued in our simple hypothesis (20). This situation calls for a theory of information structure that can overcome these problems.

Chapter 3

A Theory of Information Structure

In order to address the Identification Problem, we must first characterize information structure in terms of the properties of its components and the relation between the components. We adopt the notions of 'contextual link' and 'semantic composition' as key properties to define binomial partition of information structure, and explicate these notions. In particular, contextual link is defined as bounded inference, that is characterized in terms of discourse status, domain-specific knowledge, and linguistic marking. The chapter also demonstrates that the problems observed for binomial information structure can be overcome by adopting an appropriate grammar formalism and introducing an additional degree of freedom with structured meaning.

The chapter first presents our characterization of information structure. The next section discusses contextual link. We devote a section for linguistic marking of contextual link and analysis of special constructions in English. The last two sections introduce grammatical components of the theory and structured meaning.

3.1 Main Hypothesis: Semantic Partition between Theme and Rheme

Precise Formulation of The Main Hypothesis

In the previous chapter, we have seen that neither referential status nor linguistic form alone is sufficient to identify information structure. In this chapter, we attempt to incorporate these two properties with our main hypothesis (20). Although the main hypothesis is based on Vallduví's

[1990, p. 23] idea that "information structure is a relational notion", we depart from his analysis in several points. As we discussed in the previous chapter, we stick to the classical, simpler binomial partition of information structure. Although binomial partition is not without problems, other options appear to be more problematic, as discussed in Subsection 2.3.4. Another crucial difference from Vallduví [1990] is our position that linguistic structure alone does not fix the information structure. For this reason, analysis of 'contextual link' is essential for our solution to the Identification Problem.

As has been discussed in Subsection 2.3.1, we generally consider a theme as 'contextually linked', or 'presuppositional' [Chomsky, 1971; Jackendoff, 1972] although we cannot say that a rheme is *not* presuppositional or 'new'. The least amount we can say about this situation is that a theme *must* be contextually linked, but a rheme does not need to be. We have also associated rheme with a projection of a contrast, 'contrastiveness'. But this is not a requirement for a theme. For the moment, we call semantic, binomial partition of information structure 'semantic composition' in accordance with the view that semantic components are combined to become a more complex object. Before proceeding, let us rephrase the main hypothesis in a way convenient for the current purpose.

The main hypothesis about information structure is now characterized as follows (with symbolic representations):

(48) **Main Hypothesis** (information structure)

a. The theme is necessarily contextually-linked, i.e., $\Box \mathit{linked}\,(\mathit{Theme})$.

b. The rheme is *not* necessarily contextually-linked, i.e., $\neg\Box \mathit{linked}\,(\mathit{Rheme})$.

c. The theme is *not* necessarily contrastive, i.e., $\neg\Box \mathit{contrast}\,(\mathit{Theme})$.

d. The rheme is necessarily contrastive, i.e., $\Box \mathit{contrast}\,(\mathit{Rheme})$.

e. A proposition is a semantic composition of a theme and a rheme, i.e., $\mathit{Prop} = (\mathit{Theme})\,(\mathit{Rheme})$.

What (*a*) and (*b*) convey is that a contrast between a theme and a rheme is a contrast between the polarity of the necessity on the contextual-link property. Similarly, the contrast between (*c*) and (*d*) is the contrast between the polarity of the necessity on contrastiveness. The last statement (*e*) connects the theme and the rheme, representing the binomial relation between theme and rheme

in terms of semantic operation. The modality '□' involved in the above can be interpreted as quantification over the search process. For example, "□*linked* (*Theme*)" means that for every possible choice of theme-rheme pair, the theme is a contextual link. Thus, the hypothesis can be seen as a declarative form of such an identification process. Although we do not discuss theory-process relation in detail, the above main hypothesis can be seen as the backbone of such a relation.

Let us now examine some basic properties of the main hypothesis (48). It is consistent with the question test. The element of the response that is contextually linked to the question is a theme and the complement regardless of its referential status is a rheme. Since the notion of contextual link is more general than discourse oldness, inferrable theme is also possible. The hypothesis is equally applicable to analysis of extended texts, not just question-answer pairs. It is also consistent with generation process [e.g., Prevost, 1995], by specifying theme-rheme divisions based on the contextual link status assumed by the speaker.

Before proceeding, we should note the following. Our main hypothesis (48) does not make a reference to direct information-structure marking. We do not emphasize this point in this thesis because the focus of information-structure analysis here is written English where direct information-structure marking is rather impoverished. But the information-structure identification for spoken English and other languages can definitely take advantage of such marking. For example, Steedman [1999] presents a theory of information structure that projects theme and rheme status from intonation (in English). A similar process of projecting theme/rheme status from word order (e.g., Catalan) or particles (e.g., Japanese) is quite possible. Our proposal is compatible with such analyses. When direct marking of information structure is available, its status can simply overwrite the current analysis. In this respect, the main hypothesis (48) is a general statement that applies to underspecified cases, and subsumes more specific cases.

In the rest of this chapter, we explicate the involved notions used in the main hypothesis (48), i.e., contextual link and semantic composition. A successful completion of this process coupled with reasonable evaluation will constitute a support for the hypothesis as a theory of information structure. At this point, we make a qualification about the working domain.

Working Domain: Medical Case Reports

For the development and evaluation of the theory, we concentrate on a single working domain involving medical case reports, a type of expository texts, from a journal called "The Physician and Sportsmedicine". The choice of expository texts is natural considering the range of applications we have discussed in the Introduction. While analysis of question-answer corpora is another possibility, we consider this as a special case of the Identification Problem and attempt to solve a more general case where the context is not fixed by a question. The reasons we focus on medical case reports are as follows. First, the terminology is relatively unambiguous and referents can be identified relatively easily. Second, the domain knowledge involved in the texts is relatively limited, e.g., presence of the physician (the author of the report). Finally, a sample of medical case reports has been found on-line.

In expository texts, we can safely assume that every utterance is 'informative' at the propositional level.[1] We may add this assumption in the following form:

(49) The proposition (for an utterance) is necessarily *not* contextually-linked, i.e.,
$\Box\neg \mathit{linked}\,(\mathit{Prop})$.

In a sense, the relation between the status of a rheme, $\neg\Box \mathit{linked}$, and that of an utterance, $\Box\neg \mathit{linked}$, is a more accurate characterization of saying that a rheme is 'new' found in, e.g., Jackendoff [1972]. That is, a rheme is an essential component to make the proposition 'new' regardless of its own status.

As we mentioned in Section 2.2 (p. 22), we do not elaborate on contrastiveness for the rest of this thesis mainly for practical reasons. First, an analysis of contrastiveness is difficult to implement. Second, for expository texts, the materials are predominantly discourse-new. Thus, it is more critical to identify a contextual link for a theme (see in Chapter 7). As a consequence, the identification process ignores (48c, d).

The question whether the theory and the practice in the present work generalizes to other

[1]This is in contrast to the spoken form where informationally-redundant utterances are not uncommon [Walker, 1992]. Even for this case, we may still maintain that every utterance is informative by adopting the theory of conversational implicature [Grice, 1975] and arguing that a redundant proposition actually infers something new.

domains remains to be answered. Although different types of linguistic constructions may be involved in different domains, this component seems more consistent than the difference in domain-specific knowledge and inference. Since our theory is not bound to a specific inference mechanism unlike, e.g., Hahn et al. [1996], adjustment to a new domain seems feasible.

3.2 Contextual Link

In the previous section, we have placed the notion of contextual link at a critical position for the Identification Problem. This section explores an idea that contextual link is a bounded sequence of inference. We then make a point that such a bound on inference comes from outside the logic of inference.

3.2.1 Contextual Link and Inference

In order to explore the notion of 'contextual link', let us recall the following two examples:

(50) *i.* John has a house.

ii. [The house]$_{Theme}$ [looks exotic]$_{Rheme}$.

(51) *i.* John has a house.

ii. [The door]$_{Theme}$ [looks exotic]$_{Rheme}$.

Here, "*the house*" in (50*ii*) is discourse-old and "*the door*" in (51*ii*) is discourse-new but INFERRABLE [Prince, 1981; Prince, 1992]. Despite this difference, it is natural to identify the analogous information structures, as shown above.

As we have reviewed in Subsection 2.3.1 (p. 28), the basic idea of contextual link (in different names) has been discussed in many previous proposals [Chomsky, 1971; Jackendoff, 1972; Sgall et al., 1986; Rochemont, 1986; Prince, 1992]. A common observation is that inference is involved in the case like (51*ii*) above. Such an inference mechanism can be 'open-ended' [Brown and Yule, 1983, p. 269]. Thus, as a backbone, we need to assume a general mechanism of inference.

Let us first consider that referents of various semantic types (individuals, properties, events, etc., as discussed on p. 42 in Subsection 2.3.4) are textually or situationally EVOKED at the time of utterance. For example, at the time of uttering (51*ii*), the referent corresponding to "*a house*"

is textually EVOKED and available.[2] This *base* set of available referents can be extended by an inference mechanism. As we have set out (Section 2.2), the inference mechanism itself is a big problem, and not our central concern. But, for the sake of precision, we assume the following simple, but general inference mechanism.

(52) (assumption) Inference mechanism:

a. Textually or situationally EVOKED referents are available for processing (zero inference).

b. Relations that hold for an available referent are available. In addition, the results of composing any of these relations and referent(s) are available.

c. Referents that satisfy an available property are available. In addition, the results of composing them are available.

Note that the availability of referents and relations are constrained by various factors. Here, we assume that availability is limited to those which the speaker believes that the hearer knows, i.e., 'common ground' [Clark, 1996, for discussion].

For example, at the time of uttering (51*ii*), all the relations holding for "*the house*" are available (52*b*). Among them, there is a 'part-whole' relation applicable to "*the house*". The result of composing "*the house*" and this relation yields a property "the house has (as a part) *X*", as specified by the second clause in (52*b*). The referent corresponding to "*the door*" in (51*ii*) satisfies this property, and thus is available. Although the speaker knows that "the door looks exotic", it is not in the common ground. Thus, the inference process stops here, and the entire utterance is not considered inferrable.[3]

The above inference mechanism is recursive. Therefore, the set of available referents resulting from the process is in general unbounded. This point is made to cover inference generally, and does not claim that such an unbounded set is processed automatically. In addition, not all the available referents are equally salient in a specific context [Brown and Yule, 1983, Section 7.8]. But these are issues beyond the current scope.

We now present the notion of contextual link.

[2] In the present work, we exclude intra-utterance reference for simplicity. The process may well involve both inter- and intra-utterance reference as in Strube [1998].

[3] For a related implementation, see Dahl et al. [1987] and Palmer et al. [1993].

(53) (hypothesis) **Contextual link** is a relation between a referent in the utterance under consideration and a textually or situationally EVOKED referent where the relation is a bounded (including zero) sequence of inference steps.

We may also refer to a referent available through a contextual-link relation as a 'contextual link'. For example, we can say that "*the door*" in (51*ii*) is a contextual link. This process basically covers both EVOKED and INFERRABLE.[4] We may consider a BRAND-NEW referent as those which is not available even through an unbounded sequence of inferences. The status of UNUSED referents in the current formulation is not so clear. One possibility is that they are available in some 'extended situation'. But this point is not critical because UNUSED referents are not common in our domain.

The above characterization of contextual link has some properties distinct from proposals of Bos et al. [1995] and Hahn et al. [1996]. Unlike theirs, a general inference mechanism is assumed in a modular fashion. No a priori limit on inference steps is made. Another distinction from Bos et al. [1995] is that accommodation is not unconditionally supported (see p. 30 in Section 2.3.1). We could deal with it in a way similar to the case of UNUSED referents with 'extended situation', as mentioned in the previous paragraph.

3.2.2 Logic-External Properties for Bounding Inference

In the previous section, we have only said that inference is bounded. In this section, we discuss the way such inference is bounded. Our hypothesis is as follows:

(54) (hypothesis) Bounds on inference are conditioned by properties *external* to the logic of inference.

In other words, the above statement corresponds to the view that a general logic, for the purpose of identifying contextual links, does not have a means to terminate by itself. The current proposal hypothesizes the following properties for this purpose:

(55) *a*. Linguistic marking: e.g., definiteness in English

b. Discourse status: i.e., discourse-old referent is a contextual link

c. Domain-specific knowledge: e.g., presence of a physician and a patient in medical reports

[4] Nevertheless the above definition may not exactly correspond to the intuition given in Prince's [1981].

The above classification is not exclusive. A contextual-link referent may possess multiple properties. In order for this set of specifications to be useful, they must at least be sound. While the specification may never be complete, it must be as much complete as possible.

Among the mechanical algorithms we have reviewed in the previous chapter, Hajičová et al. [1995] focus on linguistic marking (*a*) and Hahn [1995] focuses on discourse-oldness (*b*) and domain-specific knowledge (*c*). Hoffman [1996] focuses on linguistic marking (*a*) and discourse status (*b*). The current position is that all of these must be taken into consideration.

On a more linguistic side, Birner [1997] argues that inferrables are linguistically marked. Her argument is based on several distinct linguistic phenomena including topicalization and VP preposing. But this statement is too strong. There are examples of indefinite inferrables that appear as a contextual link although this is not always the case (see Chapter 7).

In the following, we discuss the last two properties. Linguistic marking for contextual link is discussed in the next section as it requires more space.

Discourse Status

The notion of discourse status that we are talking about is basically the same as Prince [1992] (see Subsection 2.3.1). But there are two points to note. First, we deal with discourse referents [Karttunen, 1976] of a general kind, ranging over various semantic types (p. 42 in Section 2.3.4). That is, discourse statuses of not only individual types but also properties, propositions, etc. are also considered.

Second, we assume a simple notion of context that is compatible with the idea of general discourse referents. Each successfully interpreted referent is simply added to the context (if it is not already there). As we do not assume intra-utterance reference, the addition of new referents can be done once for each utterance. The context is then a heterogeneous set of discourse referents, monotonically extended as utterances are processed.[5] This is a generalization of Stalnaker's [1978, p. 321] 'context set', which is a set of propositions. As we have mentioned in Section 2.2, we do not focus on the process of reference resolution. Thus, there may be cases where (actually) identical referents are present in the context set at the same time without being resolved. Our

[5]Monotonic models of contexts are in general too simplistic, but the problem with monotonicity is left for future work.

assumption is that such a case is linguistically marked and can be analyzed as contextually-linked.

The idea of discourse-oldness is characterized as the identity relation between a referent in the current utterance and another referent in the context. A more formal representation of discourse status is described in Sections 3.4 and 4.2, after the grammatical component is discussed.

In one respect, the above idea is a cruder picture than various theories of discourse, e.g., File Change Semantics (FCS) [Heim, 1982] and Discourse Representation Theory (DRT) [Kamp, 1981]. It is because no hierarchical structure among referents is assumed. It is tempting to consider some kind of structure among referents, e.g., partial ordering by 'informativeness' relation [van Eijck, 1996, p. 89]. This may also be relevant to disambiguation of information structure. But it is beyond the scope of the current work.

Domain-Specific Knowledge

Inference may also be bounded by limited use of domain-specific knowledge. While discourse-oldness is an identity relation to a referent in the discourse, we consider a type of domain-specific knowledge that is an identity relation to a referent in the situation. Domain-specific knowledge is a prerequisite for logical inference, but the point here is that a logic does not define domain-specific knowledge. By assuming such referents in the initial situation, the inference process involving them can be effectively bounded by checking the identity relation. Such situationally-available referents also constitute the context along with the discourse referents (as discussed above).

The only domain-specific knowledge currently considered for our domain is the situational availability of physicians (e.g., *physician(s)*, *clinician(s)*) and patients (i.e., *patient(s)).* This kind of domain-specific knowledge is justifiable because each domain has its own *typical* situational setting. If such a setting is applicable to every text in the domain, it is acceptable to apply the knowledge.

3.3 Linguistic Marking in English

This section specifies linguistic marking for contextual links, and then examines several special constructions in English where we observe subtle distinctions between the linguistic marking for contextual link and that for information structure.

3.3.1 Linguistic Marking for Contextual Links

Assignment and Projection of Contextual-Link Status

A representative case of linguistic marking for contextual link is definite determiners [e.g., Heim, 1982; Poesio et al., 1997]. In Subsection 2.3.3, we have pointed out that direct linguistic marking of information structure is available only at the matrix level and non-recursive. Thus, there is no projection problem. On the other hand, linguistic marking for contextual links can appear recursively at all levels of linguistic structure. Accordingly, we need a systematic way to analyze projection of a contextual link for an arbitrary linguistic structure. This is in a sense response to Levinson's [1983, p. x] question about the projection problem for information structure in an indirect way.

For analysis of presupposition, Karttunen [1973, p. 173] introduced the ideas of 'hole' and 'plug' for presupposition projection. Informally, presupposition survives a hole, e.g., a verb *know*, but not a plug, e.g., a verb *say*. The problem of contrast projection (see Subsection 2.3.2) may also be analyzed in terms of survival of projection under various conditions.

We extend this survival-or-no classification to a more general one involving contextual links, as shown below.

(56) *a*. **Assignment**: The contextual-link status of a phrase is set/reset by one of its components.

b. **Projection**: The contextual-link status of a phrase is projected from one of its components.

For example, assignment is typically done by a function word such as a definite or indefinite determiner. Projection is typically done from a content word through a composition with certain function words. By studying contextual-link status for different linguistic structures, we can tell the consequence compositionally.

Now, there remains the main task of identifying whether a certain linguistic form is a contextual link or not. That is, we must judge whether the phrase requires a bounded sequence of inferences from an available referent. This requires linguistic analyses for various constructions. Fortunately, this is a well-studied area, e.g., Heim [1982] for definite/indefinite NP's. In the following, we examine various linguistic structures with respect to assignment/projection of contextual links. This includes contextual-link assignment by definite determiner and utterance-initial modifiers;

non-contextual-link assignment by indefinite determiner; and projection of contextual link through nominal pre-modifiers and coordinators.

Before proceeding, we must make a few remarks. The present work is incomplete in that we could not examine all the possible linguistic structures. But, even though the description can be as complex as a complete grammatical description (and thus generative), the description is bounded by the complexity of the grammar and thus presumably finite. The current coverage focuses on the constructions commonly found in medical reports in English. We observe that the coverage for our training data generalizes fairly well to reserved test data (see Chapter 7).

Definite Determiner

First, we need to clarify that we use the term 'definite' as a formal property [Prince, 1992, Section 2.1]. For example, a noun phrase "the social cost" is definite because it has the definite determiner, *the*. This is distinct from Chafe [1976, p. 39], who considers definiteness as a conceptual notion.

The role of definite determiners with respect to referential status has been investigated for a long time. For example, Brown and Yule [1983, p.170] cite an analysis that goes back to 1751 about the relation between known/unknown and definite/indefinite articles. For the present purpose, we follow more recent work [e.g., Hawkins, 1978; Heim, 1982; Quirk et al., 1985] and consider definiteness as a source of contextual-link status.

The assignment mechanism by definite determiner can be seen below. Here, a contextual link and a non-contextual link are abbreviated as *CL* and *NL*, respectively.

(57)

	Definite determiner	Noun
Example:	*the*	*door*
Contextual-link status:	–	*CL* or *NL*
Contextual-link status:	$\boxed{CL}$	

The contextual-link status of the definite determiner, *the*, itself is not critical here. The point is that it assigns a contextual-link status to the NP, shown as $\boxed{CL}$, regardless of the status of the noun, *door*.

Now, suppose that some kind of *door* that is uniquely identified is already in the discourse, it is a contextual link through discourse-oldness. The definite determiner carries on the status to the NP. If such unique identity is not guaranteed, the NP would fail to refer to a particular referent.

This position does not reject the idea that the definite determiner assigns a contextual link because the reference failure can be explained as a result of this (impossible) assignment.

On the other hand, suppose that no *door* is in the discourse or in the situation. The noun *door* is a non-contextual link. But the definite determiner still assigns a contextual-link status to the NP. This is where inference is called for, as discussed in Heim [1982]. Definite reference with a non-contextual-link noun is acceptable only when the referent corresponding to the NP is inferrable from the context. If not, reference failure may occur. This point contrasts with Bos et al. [1995], who propose that 'accommodation' always saves the reference process. In either case, a definite expression often becomes a theme, especially at the matrix level, due to its strong property to be a contextual link.

The same analysis holds for the case where the involved noun is complex, e.g., post-modified by a PP or a relative clause. Thus, nested instances of definite determiners assign contextual link status for each time, but the assignment by the embedded definite determiner does not affect the assignment of the outer definite determiner.

Other types of definite determiners include demonstrative and possessive. Demonstratives do not allow inferrables as referents, but assigns a contextual link status to the noun phrase in a manner similar to the above case. For possessive, I attempt a slightly different analysis later in this subsection.

While definite expressions are almost always contextually-linked, it is not completely so. There are cases where definite expressions express non-contextual links as follows:[6]

(58) *i.* Both buses and trolleys are operating here.

ii. Take the first bus. (a non-contextual link)

This contrasts with the corresponding contextual-link case as follows:

(59) *i.* You see three buses and a trolley over there.

ii. Take the first bus. (a contextual link)

In (58*ii*), the definite determiner, *the*, is required for the logical reason encoded in the phrase [Quirk et al., 1985, p. 270]. Thus, the expression "*the first bus*" is ambiguous between a logical use of definite determiner (58*ii*) and a contextual-link assignment (59*ii*). But this class of expressions involves a linguistic cue such as *first* or *next*, and thus can be separated from other definite

[6]Related examples are also found in Brown [1995].

expressions. In our experiment data, there is no instance of this type that affects identification of information structure. Quirk et al. [1985, p. 271] also states that body parts generally require *the*. We will come back to this case when we discuss indefinite article.

Quirk et al. [1985, p. 269] discuss yet another case of 'sporadic' referents. The situation seems idiosyncratic and differences between British and American English have also been reported. We do not discuss this case any further.

Utterance-initial Modifiers

Although English has a relatively fixed word order, there are cases where word order is flexible. We consider two such cases. One is sentential adverbials and the other is subordinate clauses. The following two examples are taken from our experiment data, and shown with the alternative word order.

(60) *a*. Until the early 1980s, tuberculosis was considered a minor, controllable public health problem.

b. Tuberculosis was considered a minor, controllable public health problem until the early 1980s.

(61) *a*. As it is used here, the term "injury" means any cheerleading injury that forces the person to miss at least 1 day of participation.

b. The term "injury" means any cheerleading injury that forces the person to miss at least 1 day of participation as it is used here.

For this matter, de Swart [1999, p. 359] analyzes temporal adverbs and argues that preposed time adverbials are themes (but postposed ones are not necessarily rhemes). The present work regards de Swart's [1999] analysis as evidence for the *contextual-link* status of preposed time adverbials, but not for theme marking. This is because adverbials can be freely preposed in an embedded clause and do not meet our requirement for direct theme marking.

The argument of de Swart is natural: preposed time adverbials set the time reference. We may extend the analysis to other situation-setting adverbs. Recall that Hoffman's [1996] topic algorithm (44) has the following condition: "when no anaphor is available in the previous utterance, choose

situation-setting adverb as the theme". This seems too strong. We also conjecture that utterance-initial modifiers are all theme, but, at this point, I am not aware of further backing in the literature.[7]

The contextual-assignment mechanism of utterance-initial modifiers are shown below. Note that the assignment of the *CL* status does not depend on the status of the argument.

(62)

	Modifier		Main clause
	Functor	Argument	
Example:	*Until*	*the 1980s,*	*tuberculosis....*
Contextual-link status:		*CL*	
Contextual-link status:	$\boxed{CL}$		

Unlike the case of the definite determiner, which is purely lexical, the above assignment is also structural in that the effect also depends on the position of the involved modifier relative to the main clause. We expect that a theory must be able to specify such structural specification in a systematic manner, which is not possible with partial parsing of Hahn [1995].

Indefinite Article

Next, let us consider the case of resetting a contextual-link status, i.e., assignment of non-contextual link to the phrase. The indefinite article, *a/an*, falls into this category. Negative also resets a contextual-link status (it does not specify a referent). The mechanism of assignment is shown below.

(63)

	Indefinite article	Noun
Example:	*a*	*door*
Contextual-link status:	–	(CL) or NL
Contextual-link status:	$\boxed{NL}$	

Typically, the noun is a non-contextual link. If the noun is a contextual link, the indefinite article still assigns non-contextual link status to the NP. This can confuse the hearer because *some* door is already in the context and the speaker insists on a 'new' door. If the speaker's intention is to refer to a new door that is distinct from what is already in the context, another determiner, e.g., *another*, may be more suitable. But there is another possibility. Let us take a look at the following example from our experiment data:

[7] Bonnie Webber [p.c., 1999] raised the following question. Not all utterance-initial modifiers behave in the same way. For example, *when* may well be a contextual-link assigner, *until* may actually not.

(64) *i.* Don't Miss Gastrointestinal Disorders in Athletes

ii. Gastrointestinal (GI) problems are common among athletes.

(three utterances omitted)

vi. so an athlete may ignore symptoms and seek medical care only when they become severe enough to interfere with performance.

Here, the noun *athlete* in (*ii*) is discourse-old. A possible analysis is that the indefinite article is used for generic reference. At this point, I conjecture that indefinite with a contextual-link noun is generic and that it exceptionally assigns a contextual-link status to the NP. This point needs further investigation, and we will come back to the consequence of this conjecture in Chapter 7.

While both countable NP's with *a/an* and uncountable NP's with no article are considered indefinite (by lacking a definite determiner), there is a semantic distinction. The indefinite article, *a/an*, in general (conversationally) implies that there are no more than one [e.g., Hawkins, 1978, p. 179; Hawkins, 1991, p. 417]. This use of the indefinite article is thus often in contrast with other determiners, e.g., *some*, *many*, *all*. On the other hand, uncountable indefinites do not have this property. Possibly for this reason, we observe more problems with identifying contextual links for uncountable indefinites (see Chapter 7).

While the majority of indefinite NP's are non-contextual links, some case assigns a contextual-link status even when the associated noun is a non-contextual link. Let us examine the following examples:

(65) *a.* I met some students before class. *A student* came to see me after class as well. [Hawkins, 1991, (11), p. 418]

b. I picked up that book I bought and *a page* fell out. [Prince, 1992, (19b)]

c. Miss Murchison,' said Mr. Urquhart, with an expression of considerable annoyance, 'do you know that you have left out *a whole paragraph*.' [Gundel, 1996, (7), p. 143]

"*A student*" in (65*a*) must be considered EVOKED because the referent is already available in the discourse.[8] "*A page*" in (65*b*) and "*a whole paragraph*" in (65*c*) are INFERRABLE. We must consider these cases as contextually linked.

[8] Contrary to a previous example (64), this instance of indefinite with a contextual link is not generic. But we will see a condition applicable to this case below.

thus, indefinite marking (at least in simple referential NPs) cannot in general separate EVOKED, INFERRABLE, and NEW. But a closer look at the involved nouns shows that there is something more to say. The first point is the lexical distinction between nouns like *page*/*paragraph*, and nouns like *student*. As observed by Prince [1992], '*page*-type' nouns are associated with another entity, say, "*a book*". In other words, this type of noun is **two-place** (or *n*-place in general), unlike *student*. We can elaborate this point as follows. First, only two-place nouns are typically defined in terms of an *of* relation in dictionaries, e.g., "page (definition 1): one side *of a leaf of* something printed or written, as a book, manuscript, or letter" [Random House, 1993]. Second, two-place nouns cannot introduce a new referent without reference to the associated referent. We can see this effect in the following test: "OK, let's start. Here is *#a page*/*a book*." using *book* as an example of one-place noun. In this regard, two-place nouns are always INFERRABLE and never NEW, while one-place nouns may correspond to any of the three statuses.A preliminary corpus check on a two-place noun *uncle* shows 47 out of 48 instances in New York Times 1995 data from Linguistic Data Consortium (LDC) are associated with an explicitly introduced referent. The case without an associated referent seems to be metaphorical. A similar result has been observed for another two-place noun *leg*. This explains why body parts usually require the definite determiner *the* Quirk et al. [1985, p. 271] (see p. 66). It must be associated with the person it belongs. On the other hand, for a set of body parts, it is also common to use the indefinite article *a/an* to indicate that only one of them is under discussion (in many cases, it does not matter which one of them).

Since the distinction between the two types of nouns is specified in the lexicon and does not require further information, we can say, for two-place nouns, linguistic information is sufficient to invoke the necessary inference. Naturally, there may be cases where a noun is ambiguous between one-place and two-place.

In example (65*a*), the process to identify the referent of "*a student*", EVOKED, is a resolution process (i.e., identity check) and not an inference. If a one-place noun that is not NEW is always EVOKED and never INFERRABLE, we can still avoid the complexity involved in an inference process. In addition, the EVOKED status of "*a student*" is strongly affected by the use of the adverbial phrase "*as well*". If we drop "*as well*" in (65*a*), the interpretation of "*a student*" is likely to be NEW rather than EVOKED, or could even be a generic. Thus, the process that invokes resolution here seems to be in the domain of semantics and not world knowledge.

Therefore, for the above cases, we have certain linguistic cues that an indefinite expression is INFERRABLE. Although I do not claim that every indefinite INFERRABLE is linguistically marked, the above presentation shows that there still are some linguistic tools to pick up a number of indefinite INFERRABLES.

Projection of Contextual-link Status

We now turn to the discussion of projection of the contextual-link status. Included in this category are non-definite determiners, certain restrictive post-nominal modification, function words, argument-taking adverbs (not at the utterance-initial position), subordinators, and coordinators.

We have seen that definite determiners and indefinite articles *assign* contextual-link and non-contextual-link statuses, respectively. In between these two classes, other determiners are treated as projectors of contextual-link status. For example, the contextual-link status of a noun phrase "*many researchers*" depends on that of *researchers*, as shown below.

(66)

	Determiner	Noun
Example:	*many*	*researchers*
Contextual-link status:	–	*X*
Contextual-link status:	$\boxed{X}$	

Here *X* is either a contextual link or a non-contextual link.

Restrictive post-nominal modifiers project the contextual-link status of the argument. For example, when *tuberculosis* is a contextual-link through discourse status, "*cases of tuberculosis*" is a contextual link due to the projection of the status from *of*-PP. For this reason, many such cases are attached with the definite determiner. The phrase "*cases of tuberculosis*" is not definite, but can be considered structurally-signaled INFERRABLE from "*of tuberculosis*".

The next case involves function words such as prepositions and auxiliary verbs. Our position is to consider them in the same class as non-definite determiners. For example, in a verb phrase "*function at a high level*", the preposition *at* projects the contextual-link status of "*a high level*". Similarly, for the case of "*is estimated*", the auxiliary *is* projects the contextual-link status of *estimated*. Assuming the same specification as non-definite determiners in (66), these function words project the contextual-link status of the argument: an NP for the case of preposition, and a main verb or another auxiliary verb for the case of auxiliary verb.

Yet another case of contextual-link projection involves coordinators. In this case, it is two-place (*n*-place in general case) rather than one-place as above. For example, the projection mechanism for a phrase "*proprioceptive training and proprioceptive rehabilitation*" is shown below.

(67)

	Conjunct 1	Coordinator	Conjunct 2
Example:	*proprioceptive training*	*and*	*proprioceptive rehabilitation*
Contextual-link status:	*X*	–	*Y*

Contextual-link status:		
	Contextual link	if *both X* and *Y* are contextual links
	Non-contextual link	otherwise

This is slightly different from the previous cases of projection because coordination in general requires that the conjuncts are *like* categories.

There is possible support for this case. When multiple individuals are coordinated, e.g., "*John and Mary*", there may be 'collective' and 'distributive' readings [Landman, 1996, p. 425 (citing several earlier papers); Palmer, 1990 (for an implementation)]. The situation can be exemplified as follows (modified from Landman):

(68) *a*. John and Mary carried the piano upstairs. (collective)

b. John and Mary signed the application. (distributive)

c. John and Mary visited their friends. (ambiguous)

The point is the existence of collective reading suggests the availability of a contextual link covering both individuals. But, even for the distributive case, e.g., (*b*) above, it is in general possible to refer to both *John* and *Mary* collectively as *they*.

Nominal Pre-modifier

Nominal pre-modification can be very complex [Quirk et al., 1985 (for an analysis and examples)]. Here, we only consider two types of nominal pre-modifiers: adjective and noun (for noun-noun compound), which are most common in our experiment data. Between these, noun-noun compounds pose a great challenge because in general, either noun can be the head of the compound [e.g., Marcus, 1980; McDonald, 1981; Sparck Jones, 1983] and this may cause distinct interpretations about the relation between the two components.

Probably, the only currently available technique to analyze the structure of noun-noun compounds is to identify the semantic relation from lexical information as has been done in the above-mentioned literature. This could be done automatically to some extent [McDonald, 1981 (applying semantic network)], but other factors including pragmatic aspects may also affect this process [Sparck Jones, 1983]. Considering such difficulties and observing the experiment data, we take a position that the contextual-link status of the first noun is projected to the noun-noun compound. This assumption needs to be re-examined for other domains because this may well depend on the current domain.

Thus, the distinct cases of contextual-link projection are hypothesized as follows: (i) modification by a noun or a denominal adjective, and (ii) modification by a non-denominal adjective. Denominal adjectives, e.g., *medical*, are closely related to nouns and usually restricted to attributive (i.e., pre-nominal) positions [Quirk et al., 1985, p. 432].

The first case, noun or denominal adjective modification carries some nominal meaning. This type of modification projects its contextual-link status, as shown below.

(69)

	Noun/Denominal Adjective	Noun
Example:	*exercise*	*program*
Contextual-link status:	*X*	*CL* or *NL*
Contextual-link status:	$\boxed{X}$	

Here, "*exercise program*" may correspond to "*program for exercise*". The modification provides a cue for the inference process to make the noun INFERRABLE. Note that the above status may still be set/reset by a determiner.

On the other hand, modification by a regular adjective projects the contextual-link status from the noun as follows:

(70)

	Common Adjective	Noun
Example:	*active*	*woman*
Contextual-link status:	*CL* or *NL*	*X*
Contextual-link status:	$\boxed{X}$	

In this case, the adjective is an additional property for the referent. Here, "*active woman*" corresponds to "*woman is active*". Thus, the contextual-link status of the adjective does not affect the result status in the same way as the first case.

Possessive

Although possessive is usually considered definite, it does not seem as strong as a definite determiner in terms of contextual-link assignment. We assume a slightly complicated contextual-link projection for possessive NPs.

(71)

	Possessor	Possessive	Possessee
Example:	*a patient*	's	*capacity*
Contextual-link status:	*X*	–	–
Contextual-link status:		*X*	
Contextual-link status:		$\boxed{X}$	

In the above, the contextual-link status of the possessor is projected to the entire NP.

Pronoun

Pronouns must be subclassified into the following three types:[9]

(72) *a*. Definite: contextual link, e.g., *these*

b. Indefinite: non-contextual link, e.g., *anyone*

c. Argument-taking: project the contextual-link status of the argument, e.g., "*many of X*"

The first case sets a contextual-link status, and the second case resets one. The third case is the same as a non-definite determiner.

Summary

As we have seen so far, linguistic marking of contextual link is rich and complex in English. In addition to linguistic marking, contextual-link status can be identified through discourse status and domain-specific knowledge. Thus, it is also possible that the contextual-link status of an discourse-old element may be projected through a complex linguistic structure guided by linguistic marking.

Before proceeding, let us make a remark on where contextual-link assignment/projection is found. Contextual-link assignment/projection is generally associated with linguistic structure where

[9] A pronoun has complex properties including the cases of discourse deixes [Webber, 1991] and the fact that a single pronoun can refer to different types of referents [Webber, 1983]. But for the purpose of analyzing contextual-link status, these kinds of subtlety do not seem critical.

extraction is not possible, e.g., NP and adverbial phrase. In these phrase types, a theme-rheme partition cannot occur because such a partition cannot be the semantic composition that results in a proposition.

On the other hand, between a verb and its arguments or between a clausal modifier and the modified clause, a contextual-link can give rise to a theme with the complement, a rheme. Thus, in general, assignment and projection of contextual-link status is not observed for these types of combinations. The resulting phrase may thus involve a mixture of contextual-link statuses. We discuss a systematic way to deal with such a case using 'structured meaning' at the end of this chapter.

3.3.2 Special Constructions

This section analyzes various constructions in English and investigates whether the construction marks information structure and/or a contextual-link status.

Topicalization, Left Dislocation, and Focus Movement

Prince [1984] discusses the pragmatic functions of topicalization and left dislocation. For example, an unmarked sentence form "*John saw Mary yesterday*" corresponds to the following two examples [Prince, 1984, (2), p. 213]:

(73) *a*. Mary John saw yesterday. (topicalization)

b. Mary, John saw her yesterday. (left dislocation)

Topicalization involves a 'gap' in the main clause, but left dislocation does not. Prince's analysis goes as follows. For topicalization (TOP), the topicalize/dislocated NP must be referential and either evoked or in a salient set relation to an evoked referent (special case of inferrable). It also signals a 'narrow' rheme within the main clause corresponding to a pitch accent. Dislocation can be classified into two subcases. The first case (LD-1) is similar to topicalization except that the 'narrow' rheme requirement does not apply. For the second case (LD-2), none of these requirements is observed. But the dislocated NP must be a rheme (Prince's 'focus').

Prince [1984, p. 220] argues that one function of TOP is to set up an open proposition in contrast to the rheme (her 'focus'). The information structure may look like the following:

(74) [This dream] [I've had *t*] [maybe three, four times]]

Theme *Rheme*

The above analysis also depends on whether the interpretation for "*this dream I've had*" can be considered a contextual link or not. This seems to be the case because in (19) on p. 218 [Prince, 1984], the preceding utterance includes "*I have a recurring dream in which...*".

But the unmarked order can be associated with the same (even more straightforward) information structure: "[I've had this dream]$_{Theme}$ [maybe three, four times]$_{Rheme}$". Then, the TOP counterpart may be used to *contrast* "*this dream*" with some other dream and still keeps the original information structure (contrastive topic as in Büring [1997b]). On the other hand, if the gap is at the end of the utterance, the unmarked form has a discontiguous information structure, but the topicalized form has a binomial *Theme – Rheme* partition as follows.

(75) *a*. [Felix]$_{Theme}$ [**praised**]$_{Rheme}$ [Donald]$_{Theme}$. (unmarked)

b. [Donald, Felix]$_{Theme}$ [**praised**]$_{Rheme}$. (topicalized)

In addition, specification of a theme requires that the theme in the above be a contextual link.

Prince [1984, fn c. on p. 214] also analyzes 'focus movement', which is structurally identical to topicalization (at least superficially) but with distinct *Rheme – Theme* pattern as follows:

(76) [**A bite**] [he wouldn't eat *t*]

Rheme *Theme*

As the example shows, the moved NP can (but does not need to) be a BRAND-NEW referent. There is no assignment of contextual-link status by focus movement. Thus, this construction only marks information-structure.

If we consolidate the preposing phenomenon common to topicalization and focus movement, the construction *either* (i) retains the original information structure (topicalization from in the middle), (ii) sets up *Theme – Rheme* information structure (topicalization from the rightmost position), or (iii) sets up *Rheme – Theme* information structure (focus movement). It is a weak condition in that the construction does not determine an information structure, but it licenses a set of information-structure patterns. Since this is a structural condition, it must be specified in the grammar and interfaced to the information-structure unit, not possible in Hahn's [1995] partial-parsing approach.

Left dislocation is structurally different from topicalization/focus movement due to the absence of the gap. LD-1 is like topicalization. But the function of LD-2 seems less certain. One possibility is that it shares the weak information-structure condition of the combination of topicalization and focus movement. That is, all of these may be a weak information-structure marker.

Finally, let us return to the hypothesis (30). Topicalization, focus movement, and left dislocation are basically all root phenomena and cannot be embedded. Thus, we can say, these constructions are partially and weakly information-structure marking. We will be comparing this situation with cleft in English shortly and with long-distance fronting in Japanese in Chapter 5.

Cleft and Pseudocleft

The traditional view about cleft (*it*-cleft) is that utterance (77*a*) below presupposes (77*b*) [Delin, 1995, p. 98, citing earlier work].

(77) *a*. It was **John** who left. (cleft)

b. Somebody left. (presupposition)

But Prince [1978, p. 898] points out that a large number of cases (called informative-presupposition *it*-cleft) do not fit into this pattern. The following is an example from Delin [1995, (7), p. 104].

(78) *i.* Joe Wright you mean

ii. Yes yes

iii. I thought it was Joe Wright who'd walked in at **first**

The information structure for the clefted part appears as follows (*a*), cf. (*b*) for (77a).[10]

(79) *a*. it was [Joe Wright]$_{Theme}$ [who'd walked in at **first**]$_{Rheme}$

b. It was [**John**]$_{Rheme}$ [who left]$_{Theme}$.

Thus, the cleft construction does not assign rheme or theme status on the clefted NP. The only possibility is that it separates theme and rheme.

Collins [1991, p. 111] presents data (Table 3.1) regarding the distribution of referential and contrastive status on the components of cleft sentences (based on a modern British English corpus). This shows that the construction does not assign contextual-link status either.

[10]The information-structure analysis for the element "*it was*" is ignored here because it is not critical for the current purpose.

	Clefted element	Complement	%
Unmarked	NEW/Contrastive	EVOKED/INFERRABLE	36.0
Marked	EVOKED/INFERRABLE	NEW/Contrastive	34.6
	NEW/Contrastive	NEW/Contrastive	29.4

Table 3.1: Corpus Analysis of Clefting [Collins 1991]

In addition, the cleft construction can be embedded, as shown in the following example [Delin, 1995, (24a), p. 111]:

(80) If it was John that ate **beans**, Bill will be disappointed.

Thus, following the hypothesis (30) that linguistic marking of information structure is matrix-level, it is not inherently an information-structure marker.

In summary, the cleft construction seems to serve various functions, including information structure (indirectly), contextual link, and contrastiveness, in a rather heterogeneous way. Thus, we could not reliably identify the involved information structure simply from the form. This contrasts with the case of topicalization/focus movement/left dislocation.

Let us now turn to the pseudocleft construction. Although pseudocleft has been once considered interchangeable with *it*-cleft as shown below, Prince [1978, (1), p. 883] argues that they are quite different.

(81) *a*. What John lost was his keys. (pseudocleft)

b. It was his keys that John lost. (*it*-cleft)

Structurally, the pseudocleft construction simply includes a 'free relative' (also 'headless' relative) at the subject position [Higgins, 1979, p. 1].[11]

Empirically, Collins [1991, p. 133] shows data (modern written British English) that the free relative of pseudoclefts are either EVOKED (64.6%) or INFERRABLE (35.4%). Note that his definition of 'free relative' includes the form such as "*the thing that...*", "*the place where...*", and "*all that...*". Collins [1991, p. 145] also shows that in 'reverse pseudoclefts', i.e., of the form "*that's what...*", the free relative is not new.[12] Then, the free relative part of a pseudocleft must be a contextual link.

[11]The definition of free relative varies. We may generally consider any *wh*-word without the head noun as free relative, e.g., *what*, *where*, *when*, *why*, *how*.

[12]He states that this type of utterance adds little information. But this point needs to be explored further.

In summary, the free relative involved in a pseudocleft marks a contextual-link status. As a free relative can appear basically in any NP slot, it works much like a definite determiner. As in the case of definite determiners, free relatives can indirectly mark a theme through the main hypothesis (48). This is quite distinct from the case of cleft in agreement with Prince's [1978] argument.

VP Preposing and Inversion

Ward [1990, p. 760, citing his 1985 thesis] argues that VP preposing "marks the entity represented by the preposed constituent as being anaphorically related to other discourse entities via a salient (partially ordered) set relation" and makes the complement as rheme ('focus').[13] The following is an example of VP preposing from Ward [1990, (1), p. 742].

(82) At the end of the term I took my first schools; it was necessary to pass, if I was to stay at Oxford, and pass I did, after a week... (the preposed VP is underlined)

He also states that the anaphoric relation is *explicit*. This suggests that VP preposing sets *Theme – Rheme* information structure.

Birner [1994, p. 251] argues that the preposed element of inversion (see below from Birner [1994, (1a), p. 233]) is either discourse-old or INFERRABLE (counting 99.77% of 1290 utterances), corresponding to our contextual link.

(83) Labor savings are achieved because the crew is put to better use than cleaning belts manually; also eliminated is the expense of buying costly chemicals. (the inverted elements are underlined)

In addition, for NPs, the preposed elements are 90% out of 1485 tokens definite, while 51% of the postposed tokens are definite. This again suggests the *Theme – Rheme* pattern.

Let us now turn to an observation that neither VP preposing nor inversion seem to be embedded. Thus, both VP preposing and inversion can be considered information-structure marking, following the hypothesis (30) that linguistic marking of information structure is matrix level. Neither VP preposing nor Inversion is very common in expository texts, but we do have one instance of inversion in our experiment data.

[13] A more recent survey is found in Birner and Ward [1999].

Heavy NP Shift

The situation with heavy NP shift (see an example below) seems less clear than previous cases.

(84) *a*. Max put all the boxes of home furnishings in his car. (canonical order)

b. Max put in his car all the boxes of home furnishings. (shifted form; Zubizarreta [1998, (145), p. 148])

Hawkins [1994] argues that the primary factor is constituent weight. On the other hand, Arnold et al. [1997] argues that the construction is conditioned by both referential status (newness) and grammatical complexity. It seems inconclusive to determine the status of heavy NP shift as either a marker of information-structure or contextual-link.

Since and Because

While both *since* and *because* can be used for a subordinate reason clause, their pragmatic function appears different. I personally have never paid close attention to any distinction until recently. I also observed that a Dutch linguist used *since* and *because* interchangeably in her examples. When I asked her about her intuition, she told me that they are the same.

Now, the observation is as follows. In response to a *why* question, only *because* clause, but not *since* clause, can be used [Lambrecht, 1994, p. 69]. Quirk et al. [1985, p. 1071] also observes that only *because* clauses can be placed in various 'focus'-related positions such as clefted position, focus of negation, and association with *only*. In addition, Moser and Moore [1995, p. 133] present a corpus-based analysis showing that 22 out of 23 occurrences of *since* precede the main clause while 13 out of 13 occurrences of *because* follow the main clause. These observations indicate that *since* cannot be a rheme, but do not restrict the status of *because*. This suggests that *since* is a theme marker.

There is a potential problem with the above analysis. Our hypothesis about information-structure marking (30) on p. 37 predicts that *since* (as a theme marker) cannot appear in embedded environments. But the following examples show the contrary.

(85) *a*. We know the story unfolds in the not-too-distant future <u>because</u> <u>since</u> there's no land to grow tobacco, they must have salvaged their cigarettes from somewhere. (New York Times 07-28-95 from LDC NYT95 at position 45048430)

b. This is the point we are seeking, for since the lengths of the subintervals tend to zero, the point *P* is also near the sequence *Q* of endpoints from the set *B*. (from a textbook on Topology)

An alternative view is that *since* is a contextual-link marker. This can explain why *since* can be a theme at the matrix level, but cannot explain why it cannot be a rheme. The situation is analogous to the case of definite expression. A definite expression at the matrix level can be a theme, but it can also be a rheme depending on the statuses of the other elements of the utterance.

At the moment, we consider the examples (85) exceptional, retain the idea that *since* as a theme marker. Further investigation is called for.

Summary

The special constructions in English are complex with respect to their pragmatic functions. The above analysis to identify marking for information structure and contextual link can provide fresh insight into this situation.

3.4 Grammatical Components

In the previous section, we have observed that lexical and structural information is crucial for identifying contextual links. To access these properties, we take a grammatical approach. In this section, we develop our grammar to capture the other major component of the main hypothesis (48), i.e., 'semantic composition'. In the first subsection, we define the notion of semantic composition along the line of Montague [1974]. This approach allows us to relate a semantic structure tightly with a surface syntactic structure. The second subsection is a partial solution to the problem with binomial information structure. By choosing an appropriate grammar formalism, we can analyze so-called 'non-traditional' constituents without loosing the precision of Montague's idea.

3.4.1 Syntax-Semantics Interface

Our starting point is the tradition of Montague [1974], also discussed in more recent textbooks [Chierchia and McConnell-Ginet, 1990; Gamut, 1991]. The semantic process can then be represented as follows (slightly modified from Gamut [1991, p. 149]):[14]

(87) Linguistic structure $\xrightarrow[\text{Translation}]{}$ Semantic representation $\xrightarrow[\text{Interpretation}]{}$ Semantic value

While it is possible to directly interpret linguistic structure (bypassing semantic representation), we opt for the above two-step approach for expository and practical reasons. For much of the discussion about formalization, we use semantic representation rather than semantic value (full interpretation).[15] In addition, our implementation solely deals with semantic representations for practicality. One additional note is that in the above figure, 'linguistic structure' is a result of parsing a linguistic expression (a string of tokens with no structure).

For semantic representation, we use the following notations:

(88) *a*. Variable: upper case, e.g., X

b. Constant: lower case

Individual: e.g., a

Property: e.g., f or $\lambda X.\lambda Y.f(X)(Y)$ (in a lambda notation)

c. Functor-argument structure: e.g., $f(a)(b)$ where the argument b is least oblique [16]

d. Modification structure: e.g., $a//b$ where a is modified by b

In many cases, a predicate may also specify an event argument. In this thesis, we consistently omit such an argument although we discuss some issues related to event.

Next, the process of translation and interpretation is represented as follows [Gamut, 1991, p. 160]:

[14]The representation in Gamut [1991] is as follows:

(1) natural language $\xrightarrow[\text{translation}]{}$ logical language $\xrightarrow[\text{interpretation}]{}$ models

[15]Semantic representation is also called logical form (LF).

[16]In this notation, "*Felix praised Donald*" is translated into $praise'(donald')(felix')$. The other argument ordering $praise'(felix')(donald')$ with the subject and object appearing according to the surface order is probably more common. The reason for the present choice of notation is that the basic operation of functional application closely corresponds to 'concatenation' or 'juxtaposition'. In addition, there is another advantage in relation to binding phenomenon discussed in Steedman [1996].

(89) *a*. Translation: $x \longmapsto x'$ (some upper-to-lower case conversion may be involved)

b. Interpretation: $[\![\varphi]\!]_{M,g} = \langle$semantic value$\rangle$

Note: M and g are the model and the assignment of variables.

For example, the translation of [*Felix praised*] [*Donald*] is shown as follows:

(90) *a*. *Felix praised* $\longmapsto \lambda X.praise'(X)(felix')$

b. *Donald* $\longmapsto donald'$

This in turn can be interpreted in a model M_1 with an arbitrary assignment g_2 as follows:[17]

(91) *a*. $[\![\lambda X.praise'(X)(felix')]\!]_{M_1,g_2} = \mathsf{property}_{123}$

b. $[\![donald']\!]_{M_1,g_2} = \mathsf{individual}_{456}$

The next step of combining elements is **semantic composition**. At the level of semantic representation, semantic composition is a relation applied to two input representations and one result representation. We consider the following two cases for semantic composition:

(92) *a*. Functional application for a functor M and an argument N: MN or $[M](N)$

β-reduction: e.g., $\left[\lambda X.f(X)\right](a) \longrightarrow_\beta f(a)$

Note: The distinct sets of parentheses in the form "$[M](N)$" is used as a visual cue of functional application.

b. Functional composition: $\left[\lambda X.f(X)\right] \circ \left[\lambda Y.g(Y)\right] = \lambda Y.f\left(g(Y)\right)$

Continuing with the earlier case, the semantic composition of "$\lambda X.praise'(X)(felix')$" and "$donald'$" can be achieved by functional application with the result "$\left[\lambda X.praise'(X)(felix')\right](donald')$". After application of β-reduction, we obtain "$praise'(donald')(felix')$". Its interpretation is "$[\![praise'(donald')(felix')]\!]_{M_1,g_2} = \mathsf{true}$" (in a certain model M_1).

At the level of semantic value, the semantic composition of (91*a*) and (91*b*) is obtained by applying the set membership "$\mathsf{individual}_{456} \in \mathsf{property}_{123}$" where $\mathsf{property}_{123}$ is a set of individuals. This should yield the same truth value as the above. The process of semantic interpretation shown above can be associated with surface syntactic structure, as shown in Fig. 3.1.

[17] A model is roughly a specification about how symbols are interpreted in the world. An assignment is a mapping from a free variable to a referent. In the shown example, there is no free variable, thus the assignment is irrelevant. For more detail, see the above-mentioned textbooks.

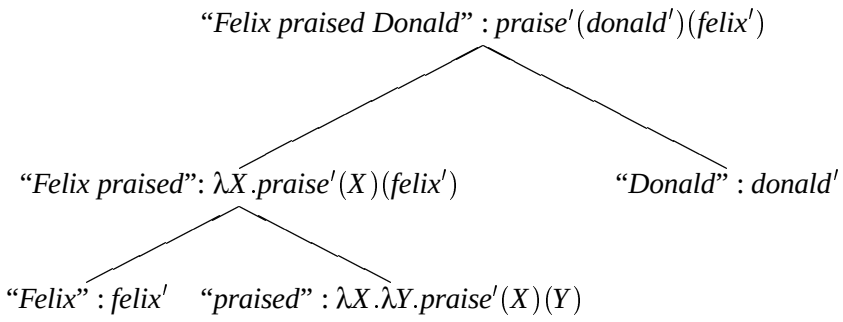


Figure 3.1: Syntax and Semantics along Linguistic Structure

In Subsection 3.2.2, we have discussed the notion of context and discourse status. With the semantics assumed here, we define the **context** as a set of semantic values, corresponding to various semantic types. Then, a semantic value is **discourse-old** if the identical one is already in the context. Note that distinct linguistic expressions or even distinct semantic representations may be interpreted into a single semantic value. For example, the following situation is possible:

(93) *a*. $[\![felix']\!]_{M_1,g_2} = \mathsf{individual}_{456}$

b. $[\![dr.\_katz']\!]_{M_1,g_2} = \mathsf{individual}_{456}$

As long as we analyze discourse status at the level of semantic value, reference can be correctly resolved even for a case like this (reference resolution is not our focus, though).

Let us now see how the main hypothesis (48) can be applied to identify information structure. Suppose that a question "*Who did Felix praise?*" has already introduced a representation "$\lambda X.praise'(X)(felix')$" into the context. The last semantic composition of the response "*Felix praised Donald*" is "$\left[\lambda X.praise'(X)(felix')\right](donald')$".[18] The component "$\lambda X.praise'(X)(felix')$" is discourse-old, and thus a contextual link. Then, the main hypothesis (48) can be applied to identify the theme, "$\lambda X.praise'(X)(felix')$", and the rheme "$donald'$".[19]

3.4.2 Flexible Constituency

Any grammar compatible with this type of semantics may be a candidate as a grammar formalism of choice. But there are a few other issues. Earlier in Subsection 3.2.2, we have considered semantic representations of various types as a source of interpretation (i.e., to obtain discourse

[18] This is not the only derivation, but we will come back to this point later.

[19] Prideaux [1979] had an idea of deriving information structure from surface structure via semantics.

referent). But most traditional grammars do not recognize a linguistic unit, i.e., a constituent, of the type "*Felix praised*", i.e., non-traditional constituent. Another problem is discontiguous information structure of the pattern such as "$Theme - Rheme - Theme$". A solution to the latter problem is possible by extending the notion of semantic representation and semantic composition, and is discussed in the next section. A solution to the former problem is possible by adopting an appropriate grammar formalism such as Combinatory Categorial Grammar (CCG) [Ades and Steedman, 1982; Steedman, 1991a].

CCG is motivated for syntactic reasons as well, with respect to coordination, extraction, and phonological structure in English [Dowty, 1988; Steedman, 1991a]. In this section, we will briefly describe some ideas about CCG and about how such non-traditional constituents can be recognized. The detailed discussion of CCG is given in Chapter 4, and some practical points in Chapter 6.

In CCG, each linguistic expression is associated with a 'category'. A category is a pair of 'syntactic types', e.g., *NP* and *S*, and the corresponding 'semantic representation', e.g., $john'$ and $clever'(john')$. Surface structure is derived through the combination of categories, i.e., both syntactic type and semantic representation. Such a combinatory process involves two types (in the current work): 'functional application' and 'functional composition'. Roughly speaking, use of functional application alone results in a system closely corresponding to context-free grammar. But, with functional composition, we have more flexibility in the way categories are combined. Now, let us represent functional composition as $f \circ g$, as in mathematics. Then, combination of $f \circ g$ and a is equivalent to combination of f and $g(a)$, i.e., "$[f \circ g](a) = [f](g(a))$". Thus, if subject-verb-object sequence can be represented as "$f - g - a$" sequence, both bracketing "$f - [g - a]$" and "$[f - g] - a$" are possible. For the earlier example, $[f - g]$ corresponds to "*Felix praised*". Now, the standard technique to analyze a NP as a function f in the Montague tradition is 'type raising'. For example, the individual type a can be type raised to $\lambda P.P(a)$, a function that takes a property as an argument. Type raising was originally motivated for coordination of an individual and quantified NPs, e.g., "*John and most students*". The associativity observed here is the source of flexibility in CCG (and other categorial grammars).

By adopting CCG, we can recognize surface constituency more flexibly than traditionally considered. This can provide a theoretical background for relating surface structure and semantic interpretation. In an earlier section, we have reviewed several cases of information-structure marking

in terms of linguistic structure. The framework allows us to describe such relations in a straightforward manner. In addition, if we process information structure in close connection to semantic representation, the framework allows parallel processing of surface structure, semantic interpretation, and information-structure processing.

3.5 Discontiguous Information Structure

In the previous section, we have seen that Combinatory Categorial Grammar (CCG) is a solution to non-traditional constituency. But we also have observed another problem for binomial information structure, i.e., discontiguous information structure. This problem has not yet received full attention, except for Krifka [1992] and Steedman [1999, Section 5.5]. This section presents a solution to this problem based on their insight and techniques, focusing on the concept underlying the solution. A more formal presentation will be covered in Section 4.3.

Motivation

We have adopted a binomial information structure to model the informational contrast between theme and rheme. But, as discussed in Subsection 2.3.4, other types of partitions have been proposed as well. One (but not the only) motivation for such a move is to account for discontiguous information structure such as in the form of "*Theme – Rheme – Theme*", as can be seen in the following example [Steedman, 1999, (35)]:

(94) *Q*: I know which team Mary **expects** to **lose**. But which one does she **want** to **win**?

A: [Mary **wants**]$_{Theme}$ [**Ipswich**]$_{Rheme}$ [to **win**]$_{Theme}$.

The following is a still more complicated example with the pattern of "*Theme – Rheme – Theme – Rheme*" [p.c., Mark Steedman, 1998].

(95) *Q*: I know what team Fred wants to win the Cup, but which team does Alice want to lose which contest?

A: [Alice wants]$_{Theme}$ [**Australia**]$_{Rheme}$ [to lose]$_{Theme}$ [the **Ashes**]$_{Rheme}$.[20]

Although CCG can accept constituents more flexibly than traditional grammars do, discontiguous information structures do not correspond to constituents recognized even by CCG.

[20] With or without L+H* on the themes.

Analysis

By observing the examples (94, 95), we might consider a possibility that the discontiguity is a result of syntactic restrictions on realization of information structure. That is, in English, the word order is basically fixed and the information structure is separated due to that factor. If this is the case, we should be able to analyze and predict occurrences of discontiguous information structure simply through syntax. But this is not the case.

Let us consider an example in Japanese (grammatical labels: TOPic, ACCusative, NOMinaLizer, COPula, and Question).

(96) *Q*: Ken-wa nani-o tabeta-no?
Ken-TOP what-ACC ate-Q
"What did Ken eat?"

A: [Ken-wa]$_{Theme}$ [**banana**-o]$_{Rheme}$ [tabeta]$_{Theme}$.
Ken-TOP banana-ACC ate
"Ken ate a banana."[21]

The strict verb-final property is one thing that causes the discontiguous information structure. But that is not the only factor. Either of the following responses may be uttered in place as well.

(97) *a*. [**Banana**-o]$_{Rheme}$ [Ken-wa tabeta]$_{Theme}$.
banana-ACC Ken-TOP ate
"It was a banana that Ken ate."

b. [Ken-ga tabeta-no-wa]$_{Theme}$ [**banana**-da]$_{Rheme}$
Ken-NOM ate-NML-TOP banana-COP
"What Ken ate was a banana."

Note that the above two are grammatically more marked forms than the SOV in (96*A*) and that there are forms of questions that correspond to these marked forms. But, in any case, the form of question does not seem to restrict the form of response.

Thus, we cannot say that discontiguous information structure is a result of syntactic constraints. We need to accept that there are various factors that cause discontiguous information structure. For whatever reasons, once a particular construction is chosen, information structure must be realized even if discontiguity results.

[21] Depending on the situation, the definite article *the* may also be applicable.

Even for the discontiguous case, there are a few properties that stay as in the contiguous case. First, the surface syntax does not violate the grammaticality. Second, discontiguous theme (rheme) elements can be combined into a single theme (rheme) semantic unit, and then the theme and the rheme can compose and derive the proposition corresponding to the utterance. For example, consider the utterance (95*A*) repeated below.

(98) [Alice wants]$_{Theme}$ [**Australia**]$_{Rheme}$ [to lose]$_{Theme}$ [the **Ashes**]$_{Rheme}$.

Each theme/rheme component may be semantically represented as follows:

(99) *a*. "Alice wants": $\lambda X.\lambda Y.want'(X)(Y)(alice')$

b. "Australia": $australia' = \lambda P.P(australia')$

Note: The right-hand side is a 'type-raised' semantic representation of the individual.

c. "to lose": $\lambda X.lose'(X)(pro)$

d. "the Ashes": $ashes' = \lambda P.P(ashes')$

Here, the treatment of control structure has been simplified [Steedman, 1996, for more detail]. The semantic representations for the combined theme and rheme are as follows:

(100) *a*. $Theme: [\lambda X.\lambda Y.want'(X)(Y)(alice')](\lambda X.lose'(X)(pro))$
$= \lambda X.\lambda Y.want'(X)(lose'(Y)(pro))(alice')$

b. $Rheme: [\lambda P.P(ashes')] \circ [\lambda P.P(australia')] = \lambda P.P(australia')(ashes')$

Informally, this corresponds to a pair of (ordered) individuals that would satisfy a certain property.

The proposition can now be derived as follows:

(101) $Proposition: [\lambda P.P(australia')(ashes')](\lambda X.\lambda Y.want'(X)(lose'(Y)(pro))(alice')) =$
$[\lambda X.\lambda Y.want'(X)(lose'(Y)(pro))(alice')](australia')(ashes') =$
$want'(australia')(lose'(ashes')(pro))(alice')$

The correct semantic analysis of discontiguous information structure and thus must correspond to the *usual* semantic analysis of utterance.

Therefore, while semantic derivation of discontiguous information structure does not directly correspond to the surface derivation, it must be semantically in concordance with the surface derivation. We propose an analysis of discontiguous information structure, which can be used to account for the semantic derivation we have just seen above.

Structured Meaning Approach: Introduction

In order to allow the discontiguous patterns, we need to accept an additional degree of freedom in linguistic analysis. For this purpose, we adopt the 'structured meaning' approach [von Stechow, 1991; Krifka, 1992] (both cite earlier work of Klein and von Stechow and that of Jacobs).

The point of the structured-meaning analysis is as follows. The traditional semantic representation as a value corresponding to a constituent is not sufficient to analyze the correct 'focus' projection, i.e., the focus scope. We use the term 'focus' here following the literature (but it really is our 'contrast'). This problem can be solved if, as a semantic representation, we associate with a constituent a 'structure', rather than a value. For a sentence "John only introduced Bill to Sue.", the following three distinct focus scopes are possible [e.g., von Stechow, 1991] (the index is used to indicate the association).

(102) *a*. John only$_1$ introduced **Bill**$_1$ to Sue.

b. John only$_1$ introduced Bill to **Sue**$_1$.

c. John only$_1$ introduced [**Bill** to **Sue**]$_1$.

Purely syntactic approaches [e.g., Chomsky, 1971; Culicover and Rochemont, 1983] assume that a focus feature $[+F]$ on a phrase is projected from a pitch accent at a specific position, e.g., rightmost head of the phrase. But these approaches would assign the same syntactic structures for the above cases. Thus, the above distinction cannot be accounted for.

Structured meaning is proposed to solve this problem by deriving structured semantic representation to capture the underlying contrast between 'background' and 'focus' (their terminology). Combined with a semantic analysis such as Rooth [1996], this approach can provide correct semantics for the examples in (102). The standard representation used in the literature for structured meaning is $\left\langle Background, Focus \right\rangle$. The structured meanings corresponding to the verb phrases in (102) are shown as follows:

(103) *a*. John only$_1$ introduced **Bill**$_1$ to Sue.

$\left\langle \lambda X.\lambda Z.introduce'(X)(sue')(Z), bill' \right\rangle$

b. John only$_1$ introduced Bill to **Sue**$_1$.

$\left\langle \lambda Y.\lambda Z.introduce'(bill')(Y)(Z), sue' \right\rangle$

c. John only$_1$ introduced [**Bill** to **Sue**]$_1$.

Analysis 1: $\left\langle \lambda X.\lambda Y.\lambda Z.introduce'(X)(Y)(Z), bill', sue' \right\rangle$ (multiple foci as a list [von Stechow, 1991, p. 43])

Analysis 2: $\left\langle \lambda X.\lambda Y.\lambda Z.introduce'(X)(Y)(Z), bill' \bullet sue' \right\rangle$ (multiple foci as a product [Krifka, 1992, p. 21])

In order to justify the structured-meaning approach, let us discuss a few more applications. Structured meaning is also used for an analysis of propositional attitude [Cresswell, 1985]. The point is that the argument of propositional-attitude verbs, e.g., *think*, is not a semantic representation as a value but its structure. Another application is to an analysis of thematic role [Chierchia, 1989]. He shows that this move can provide an appropriate analysis of control structure.

But there is a limitation with the previous work. The general case of semantic composition is not discussed in von Stechow [1991]. Krifka [1992] defines four cases of functional application of two structured meanings, depending on how the two components of structured meanings are applied. But his analysis is also too limited for our purposes. The only case of composing two structured meanings results in a 'product' $(bill' \bullet sue')$, as can be seen in (103). We need a more general approach that is applicable to an arbitrary semantic type. Since CCG involves both functional application and functional composition as a means of semantic composition, we also need to consider both of these.

Since the 'structured meaning' approach is occasionally compared with the 'alternative semantics' approach [Rooth, 1985], it seems beneficial to briefly discuss their relation. Structured meaning is one way of semantic representation and alternative semantics is one way of interpreting semantic representations. Researchers who focus on structured meaning assume certain semantic interpretations [Krifka, 1992, p. 21]. Those who focus on Alternative Semantics assume certain syntactic mechanisms to deliver a desirable semantic representation [Rooth, 1996]. Therefore, it is rather pointless to compare both approaches in terms of expressibility, and argues that structured meaning is more expressive than Alternative Semantics as in von Stechow [1991, p. 73]. He seems to consider alternative semantics too simplistically. Partee [1999] also emphasizes the difference that structured meaning and Alternative Semantics are a 'grammaticalized' and a 'non-grammaticalized' approach. But these approaches must be syntactic and semantic sides of a single coin.

Application to the Current Theory

In the current work, we adopt structured meaning for the contrast between a contextual link and a non-contextual link. The intuition behind this move is that for each constituent, the semantic representation may keep such a contrast rather than reducing it to a simple semantic value, unlike assignment/projection of contextual-link status (Subsection 3.3.1). This enables us to 'carry' a binomial internal structure of constituents to the next level of semantic composition. The use of contextual-link status is feasible because it can be identified in terms of discourse status, linguistic form, and domain-specific knowledge.

The structured meaning approach adopted in this section allows us to analyze discontiguous information structure within a binomial model of information structure. This is important for several reasons. First, we can analyze realistic linguistic data with a simple model of information structure. Second, by avoiding multiple partitions of information structure, we can focus on a small number of properties that characterize information structure more precisely. By integrating with a Montague-style analysis, congruent relations between syntax, semantics, and information structure are possible. It facilitates the connection between linguistic marking of information structure and contextual link to the grammatical components of phonology, syntax, and semantics. The relation to processing can be improved as well by allowing parallel processing of contextual link and information structure along parsing. Potentially, it can also provide semantic representations for Alternative Semantics analysis. In the next chapter, we will also discuss formalization of the proposed approach and an application to an analysis of 'gapping'.

3.6 Summary

In the theory of information structure developed in this chapter, we emphasize the following two points. Themes are necessarily 'contextually-linked' and a proposition is a 'semantic composition' of a theme and a rheme. The notion of contextual link is further characterized by discourse status, domain-specific knowledge, and linguistic marking. We also observe that a number of linguistic analyses provide support for contextual-link marking.

Semantic composition is captured within a framework of CCG, which can recognize surface constituents corresponding to units of information structure. We also address another potential

problem for binomial partition and propose a solution using structured meaning. The chapter argues that the proposed theory can be used for analyzing information structure in texts and is thus a key to the Identification Problem.

We have left two main components of the theory for the following two chapters, i.e., formalization within CCG and analysis of linguistic marking of information structure in Japanese. Once these are explored, we can proceed to implementation and evaluation of the theory.

Chapter 4

Formalization of the Theory with Combinatory Categorial Grammar

In the previous chapter, we have mentioned that two potential problems for binomial partition of information structure, i.e., non-traditional constituency and discontiguous information structure, can be solved by adopting Combinatory Categorial Grammar (CCG) and by integrating structured meaning, respectively. This chapter demonstrates that how these two points can be achieved within a variant of CCG formalism. We also show that the characterization of contextual links can be specified within the same framework. In the present work, we use the term 'formalization' in the sense of 'specification' within a grammar formalism as a basis for implementation. Thus, it is distinct from the level of formalization commonly pursued by formal semanticists.

Section 4.1 introduces and discusses CCG. Topics include a review of several motivating cases, derivations of a simple sentence, a summary of the standard framework, and some extensions of the framework. We also discuss computational properties of CCGs in Subsection 4.1.5. Then, Section 4.2 discusses specification of contextual links. Finally, the idea of structured meaning is integrated with the framework (Section 4.3).

4.1 Combinatory Categorial Grammar

This section introduces the CCG framework. We start from simple examples of derivations in CCG. Then, a summary of standard CCG and two types of extensions are presented. Finally,

generative power and theoretical parsing efficiency are discussed.

4.1.1 Motivation

In Section 3.4, we observe that tight syntax-semantics relation in the Montagovian tradition [Montague, 1974] can simplify the analysis of information structure. We have also argued that this direction can be extended to include 'non-traditional' constituency such as subject-verb sequence, e.g., "*Felix praised*" as in (18). These points can be captured by a group of extended Categorial Grammars including Combinatory Categorial Grammar (CCG) [Ades and Steedman, 1982; Steedman, 1985; Dowty, 1988].[1]

Let us first explore several motivating cases involving 'non-traditional' constituency. Probably the most discussed aspect of non-traditional constituency is in association with coordination. For example, the following pattern [Steedman, 1996, (86), p. 37] poses a problem to most traditional grammar formalisms because subject-verb sequence cannot be readily recognized.

(104) { Keats steals and Chapman eats } apples.

Similar situations are observed in other languages as well. The following is a coordination of NP sequences in Japanese [Komagata, 1997a, (1)].

(105) John-ga Mary-o , Ken-ga Naomi-o tazuneta.
{ John-NOM Mary-ACC (and) Ken-NOM Naomi-ACC } visited
"John visited Mary and Ken, Naomi."

Again, this is a problem for most grammar formalisms.

Many constructions involving 'extraction' are often handled with the help of empty categories, i.e., 'trace'. Let us now turn to the following example [Steedman, 1996, modified from (34), p.59]:

(106) the apples which I think Keats likes

A textbook-style analysis [Haegeman, 1991, p. 370] of such a case may look like the following:

(107) [whom$_i$ [I think Keats likes t_i]]

But this type of analysis is not re-usable for Right Node Raising (RNR) [Steedman, 1996, (35), p. 59].

(108) { I think Keats likes, but you say he detests }, the man in the grey flannel suit.

[1] Wood [1993] is a good overview of Categorial Grammars in general.

The traditional work often assumes that the above two phenomena require separate analyses. But they are parallel with respect to both surface structure and interpretation. Thus, it is desirable to have a grammar that can demonstrate this point [Steedman, 1996, p. 59].

Non-traditional constituency is also observed in relation to prosodic structure in English [Steedman, 1991a, (49), p. 282].

(109) *Q*: I know what Fred **cooked**. But then, what did he **eat**?

A: [Fred $\underset{\text{L+H* LH\%}}{\textbf{a-ate}}$] [the $\underset{\text{H* LL\%}}{\textbf{beans}}$].

The symbols below (*A*) indicate intonation. L+H* and H* tones are argued to be theme and rheme markers, respectively [Steedman, 1991a]. The traditional approach is forced to take a position that prosodic structure is independent of syntactic structure (this is in fact the line taken by many recent researchers, see the discussion and references in Steedman [1999, Chapter 5]). But Steedman [1999] argues that it is more intuitive if the prosodic structure is close to syntactic structure.

In addition, there is an interesting observation about prosodic structure in Japanese. Kubozono [1993, p. 3] analyzes Japanese prosody in detail and discovers that right-branching cases, but not left-branching ones, are marked. This suggests that the prosodic structure in Japanese has a left-branching structure as in the case of English. This is striking from the view point that Japanese syntax is strictly head-final, i.e., right-branching [Kubozono, 1993, p. 158]. This situation for a simple Subj-Obj-Verb pattern is shown below.

(110) *a*. Prosodic phrasing: [[Subj Obj] Verb]

b. Syntactic structure: [Subj [Obj Verb]] (as assumed by Kubozono)

Kubozono [1993, p. 222] proposes a solution to adjust prosodic structure to match right-branching syntactic structure. Although discussion on this point is beyond the scope of the current work, we can also address the problem from the syntactic side. Namely, the assumption that syntactic structure in Japanese is categorically right-branching may not be correct. In fact, we have already observed in (105) that Subj-Obj sequence (NP sequence) can be a constituent for the coordination purpose. Thus, it may well be the case that the observed prosodic phrasing directly corresponds to the syntactic structure recognized by CCG.

There is yet another point from psycholinguistic view point, i.e., incremental processing. If a human processes utterances in the left-to-right order, the string consists of the subject and the

verb in an utterance in English must have been (at least partially) processed before the object is encountered [Ades and Steedman, 1982].

Among the family of Categorial Grammars that naturally capture non-traditional constituency, we adopt CCG for the following reasons. Various linguistic analyses have been undertaken within the framework, e.g., coordination/extraction [Steedman, 1985; Dowty, 1988; Steedman, 1996], interface to prosody (in English) and information structure [Steedman, 1991a; Prevost, 1995; Hoffman, 1995]. The standard version of CCG [Steedman, 1996] has a desirable generative capacity, i.e., mildly context-sensitive and weakly equivalent to Lexicalized Tree-Adjoining Grammar (LTAG), [Vijay-Shanker and Weir, 1994] and is polynomially parsable [Vijay-Shanker and Weir, 1993]. Several forms of extensions have been proposed and their generative power and parsing efficiency have been analyzed [Hoffman, 1995; Komagata, 1997a]. Yet another area is relation to quantifier scope [e.g., Park, 1996]. In addition, a practical parser has been constructed [Wittenburg, 1986; Komagata, 1997a]. Not all of these aspects have been explored in other related extended Categorial Grammars such as Lambek Calculus [Lambek, 1988, originally published in 1958] and Unification Categorial Grammar (UCG) [Zeevat, 1988].

4.1.2 Derivation Examples

Traditional Case

In this subsection, we first introduce the basics of CCG through a 'traditional' derivation of "Felix praised Donald". Then, the second half presents a 'non-traditional' derivation of the same sentence.

First, the lexical entry for each word is specified in the following manner:

(111) Lexicon: ⟨phonological form⟩ $\underset{\text{(assignment)}}{:=}$ ⟨category⟩
where ⟨category⟩ := ⟨syntactic type⟩ : ⟨semantic representation⟩
For example,

a. Felix := $NP : felix'$

b. praised := $(S\backslash NP)/NP : \lambda X.\lambda Y.praise'(X)(Y)$

c. Donald := $NP : donald'$

For simplicity, we may also refer to syntactic type as category where no confusion arises. The complex category $(S\backslash NP)/NP$ can be read that it first takes an NP category to the right and then

another NP category to the left ('result-leftmost' representation).[2] We assume left-associativity for the slash symbols '/' and '\'. Thus, we may abbreviate $(S\backslash NP)/NP$ as $S\backslash NP/NP$ without parentheses. Each category may be associated with a finite, non-recursive set of features such as $NP_{[agr=(3pers,sing,nom)]}$ although not shown in this chapter to avoid complexity. Steedman [1996, Section 2.1] discusses the use of features for agreement and binding. Features are used extensively in the implementation (Chapter 6).

We first see the derivation of the VP "*praised Donald*". Syntactically, this process can be seen as a result of **functional application** to the two categories (informally, a cancelation of the outermost argument) as follows:

(112) *a*. Rule: Functional Application (in categorial form)

$$X/Y \qquad Y \quad \Longrightarrow \quad X$$

b. Instance of rule application

praised	Donald		praised Donald
$S\backslash NP/\underline{NP}$	$\underline{NP}$	$\Longrightarrow$	$S\backslash NP$

Note: Underline may be used to indicate the cancelation of the involved categories.

Semantically, the process is an instance of functional application (β-reduction in the lambda-calculus term) as follows:

(113) *a*. Rule: functional application (β-reduction)

$$\lambda X.f(X) \qquad a \quad \Longrightarrow \quad f(a)$$

b. Instance of rule application

praised	Donald		praised Donald
$\lambda X.\lambda Y.praise'(X)(Y)$	$donald'$	$\Longrightarrow$	$\lambda Y.praise'(donald')(Y)$

The next step of deriving a sentence from the subject and the VP is analogous except for the directionality of the functor.

(114)

	Felix	praised Donald		Felix praised Donald
Syntactically:	$\underline{NP}$	$S\backslash\underline{NP}$	$\Longrightarrow$	S
Semantically:	$felix'$	$\lambda Y.praise'(donald')(Y)$	$\Longrightarrow$	$praise'(donald')(felix')$

[2]The result-leftmost representation is seen in contrast to the European tradition [e.g., Morrill, 1994], where argument categories are placed either to the left or right depending on the slash direction as in $NP\backslash S/NP$. It is more difficult to read off the type in this notation.

This way, surface structure and semantic representation can be associated in a straightforward manner following the Montagovian tradition.

Non-traditional Case

For the non-traditional derivation, we need two more rules: *type raising* and *functional composition*. Intuitively, **type raising** is an operation of transforming, say, an NP into a functor category that takes a VP as its argument. This shift was originally motivated to capture the property of quantified NPs whose quantifier scopes over a VP [Montague, 1974]. Type-raising the individual type such as *Felix* also allows us to coordinate it with a quantified NP, as in "*Felix and some dogs*". The following example illustrates the application of type raising to an NP, *Felix*:

(115) *a*. Syntactically: $X \Longrightarrow S/(S\backslash X)$

e.g., $NP \Longrightarrow S/(S\backslash NP)$

Note: We may abbreviate $S/(S\backslash NP)$ as $NP^{\uparrow}$.

b. Semantically: $a \Longrightarrow \lambda F.F(a)$

e.g., $felix' \Longrightarrow \lambda F.F(felix')$

Functional composition (for CCG) is basically the same as its mathematical counterpart. In mathematics, functional composition provides a means of analyzing function application in an associative way: e.g., $f(g(X)) = [f \circ g](X)$. In CCG, functional composition enables the grammar to recognize subject-verb sequence as a constituent, still looking for an object. Assuming that type raising is applied to *Felix*, we describe the next step involving functional composition as follows:

(116) *a*. Rule: Functional Composition (in categorial form)

$X/Y \quad Y/Z \Longrightarrow X/Z$

b. Instance of rule application

Felix	praised	Felix praised
$S/\underline{(S\backslash NP)}$	$\underline{(S\backslash NP)}/NP \Longrightarrow$	S/NP

Semantically, the process is as follows:

(117) *a*. Rule: functional composition (in mathematical term)

$\lambda Y.f(Y) \quad \lambda X.g(X) \Longrightarrow \lambda X.f(g(X))$

$f \quad g \quad f \circ g$ (in another notation)

b. Instance of rule application[3]

Felix	praised		Felix praised
$\lambda f.f(felix')$	$\lambda X.\lambda Y.praise'(X)(Y)$	$\Longrightarrow$	$\lambda X.praise'(X)(felix')$

The resulting category, "$S/NP : \lambda X.praise'(X)(felix')$", is the CCG representation of the non-traditional constituent we are concerned with. Another step of functional application leads to derive exactly the same category including the semantic representation.

(118)

	Felix praised	Donald		Felix praised Donald
Syntactically:	$S/\underline{NP}$	$\underline{NP}$	$\Longrightarrow$	S
Semantically:	$\lambda X.praise'(X)(felix')$	$donald'$	$\Longrightarrow$	$praise'(donald')(felix')$

This demonstrates that CCG is capable of recognizing non-traditional constituents needed for our analysis of information structure. Staying with the Montagovian tradition, the grammar still tightly interfaces surface syntactic structure and semantic representation. This allows us to provide an interface not only between linguistic expression and semantic representation, but also between semantic representation and our notion of referent, a unit of information structure, as we will see shortly.

4.1.3 Standard CCG: A Summary

In CCG, like other lexicalized formalisms such as Lexicalized Tree-Adjoining Grammar (LTAG) [Schabes, 1990], much of the syntactic information is stored in the lexicon. But a great deal of syntactic generality comes from the use of a small number of combinatory rules introduced in the previous section. Mostly following the framework outlined in Steedman [1996], we summarize our combinatory rules as follows:

(119) Functional application:

					Rule symbol
a.	$X/Y : f$	$Y : a$	$\Longrightarrow$	$X : f(a)$	$(>)$
b.	$Y : a$	$X\backslash Y : f$	$\Longrightarrow$	$X : f(a)$	$(<)$

(120) Functional composition:

[3] $[\lambda f.f(felix')](\lambda x.\lambda y.praise'(x)(y)) = [\lambda x.\lambda y.praise'(x)(y)](felix') = \lambda x.praise'(x)(felix')$

$$a.\quad X/Y:f \qquad Y/Z:g \quad\Longrightarrow\quad X/Z:\lambda X.f(g(X)) \qquad (>B)$$
$$b.\quad Y\backslash Z:g \qquad X\backslash Y:f \quad\Longrightarrow\quad X\backslash Z:\lambda X.f(g(X)) \qquad (<B)$$

To be precise, the combinatory rules are rule schemata. The meta-variables X, Y, etc. need to be instantiated for a particular choice of categories. We will also call the application of these rule schemata **combination**.

There are directional variations for functional composition as follows:

(121) Functional composition (crossing variation):

$$a.\quad X/Y:f \qquad Y\backslash Z:g \quad\Longrightarrow\quad X\backslash Z:\lambda X.f(g(X)) \qquad (>B\times)$$
$$b.\quad Y/Z:g \qquad X\backslash Y:f \quad\Longrightarrow\quad X/Z:\lambda X.f(g(X)) \qquad (<B\times)$$

The backward variety (*b*) is used for combining certain verb arguments in a non-traditional way, as in the following example:

(122) John put on the table "John put on the table"

$$\underline{S/PP}/NP \qquad S\backslash\underline{(S/PP)} \quad\Longrightarrow\quad S/NP \qquad (<B\times)$$

(with PP ⇓ $S\backslash(S/PP)$)

The resulting phrase may be a part of a relative clause "*the book John put on the table*" or a heavy NP shift "*John put on the table that incredibly-heavy dictionary*". The forward variety is not used in English [Steedman, 1996, p. 53], but is a basis for long-distance fronting in Japanese.

Generalized versions of functional composition with multiple arguments, e.g., "$X/Y \quad Y/Z_1/Z_2 \Longrightarrow X/Z_1/Z_2$" are also used in Steedman [1996, p. 35]. These can be labeled as $>B^k$ and $<B^k$ where k corresponds to the number of 'passed' arguments $Z_1,\ldots,Z_k$. In this regard, the basic functional composition above may be written as $>B^1$ and $<B^1$, and functional application as $>B^0$ and $<B^0$.

We repeat the basic form of type raising as follows:

(123) Type raising:

$$a.\quad X:a \quad\Longrightarrow\quad S/(S\backslash X):\lambda f.f(a) \qquad (>T)$$
$$b.\quad X:a \quad\Longrightarrow\quad S\backslash(S/X):\lambda f.f(a) \qquad (<T)$$

Here, we assume syntactic type raising (dynamically applied during derivation). But it is also possible to type raise categories in the lexicon [Steedman, 1991b, p. 75].

Finally, we give the coordination rule schemata as follows:

(124) Coordination:

$$X:f \qquad Coord:c \qquad X:g$$

$$\Longrightarrow \quad X:\begin{cases} c(f)(g) & (<\Phi^0>) \\ \lambda X.c(f(X))(g(X)) & (<\Phi^1>) \\ \lambda X.\lambda Y.c(f(X)(Y))(g(X)(Y)) & (<\Phi^2>) \end{cases}$$

Separate semantic cases are needed for different arities.

In addition, there is another type of combinatory rule called 'substitution', "$(X/Y)/Z \quad Y/Z \Longrightarrow X/Z$" (one direction), used for the analysis of parasitic gap, also a part of the CCG framework [Steedman, 1996, p. 39], but not used in the current work.

4.1.4 Extensions of CCG

While the standard CCG is capable of dealing with a wide range of linguistic constructions, there are cases where some extensions are called for. We present two such cases in this section, namely Multiset-CCG [Hoffman, 1995] and CCG-GTRC [Komagata, 1997c; Komagata, 1997a].

Multiset-CCG

Languages like German and Turkish are known for their extremely flexible word order. Becker et al. [1991] observe that German long-distance fronting (scrambling) involved in this phenomenon has the following properties: (i) there is no bound on the distance of movement and (ii) there is no bound on the number of constituents that are moved. The same situation is also observed in Turkish. Hoffman [1995, (11), p. 46] follows Becker et al. [1991] and represent the phenomenon in the following form:

(125) $(NP_1 ... NP_m)_{scrambled}\ V_m ... V_1$

Hoffman [1995] then argues that the competence grammar must be able to capture the set of all of these scrambled strings. She points out that standard CCG does not have this property and that a more powerful grammar is called for. Hoffman [1995] develops a formalism called 'Multiset-CCG'. The idea behind Multiset-CCG is that the 'bags' of arguments of different verbs can mix freely. The term 'bag' is used here to indicate that they are multisets, allowing duplicate entries

without order, and neither sets (non-redundant) nor lists (ordered).

In support for her choice of Multiset-CCG, Hoffman [1995, Section 2.4] discusses several types of extensions of standard CCG. One of such extension is to use type raising and backward crossing composition (see the previous subsection) to partially cover the case of (125). The following example shows the case where the NP arguments of the inner verb V_1 are fronted to the sentence-initial position.

(126)

NP_{1b}	NP_1	NP_{2a}	V_1	V_2	
$S/\underline{(S\backslash NP_{ACC})}$	$S/\underline{(S\backslash NP_{NOM1})}$	$S/\underline{(S\backslash NP_{NOM2})}$	$\underline{S}\backslash NP_{NOM1}\backslash NP_{ACC}$	$S\backslash NP_{NOM2}\backslash\underline{S}$	$<B^2$
			$S\backslash\underline{NP_{NOM2}}\backslash NP_{NOM1}\backslash NP_{ACC}$		$>B^2_\times$
			$S\backslash\underline{NP_{NOM1}}\backslash NP_{ACC}$		$>B_\times$
			$S\backslash\underline{NP_{ACC}}$		$>$
			S		

But Hoffman [1995, p. 34] points out that this approach cannot deal with scrambled coordination such as the following:[4]

(127)

NP_{ACC}	NP_{NOM}	&	NP_{ACC}	NP_{NOM}	V	
$S/(S\backslash NP_{ACC})$	$S/(S\backslash NP_{NOM})$		$S/(S\backslash NP_{ACC})$	$S/(S\backslash NP_{NOM})$	$S\backslash NP_{NOM1}\backslash NP_{ACC}$	
	$>B$			$>B$		
$S/(S\backslash NP_{ACC}\backslash NP_{NOM})$			$S/(S\backslash NP_{ACC}\backslash NP_{NOM})$			
					$<\&>$	
						*

But there is a simple solution. We can admit that local scrambling is a reflection of the ambiguous verb categories between $S\backslash NP_{NOM}\backslash NP_{ACC}$ and $S\backslash NP_{ACC}\backslash NP_{NOM}$ [Baldridge, 1998, Section 3.2]. Then, the above situation can be handled within the framework of standard CCG.[5]

[4] This situation is the same in Japanese.

[5] But there is an even worse possibility. The following example is acceptable in Japanese (or Korean).

(1) Donald-o Felix-ga , Mickey-ga Roger-o hometa.
{ Donald-ACC Felix-NOM } CONJ { Micky-NOM Roger-ACC } praised
"Felix praised Donald, and Mickey [praised] Roger."

We cannot discuss this situation any further in the current work.

There also is a warning against the power of Multiset-CCG. Joshi et al. [1994] question the property (125) and point out that two levels of argument mixture can be covered within the framework of TAG. Their argument is that if a competence grammar can characterize the practical bound on a phenomenon, it is more appropriate to assume such a grammar for description. In either case, since we do not readily encounter this situation in our English and Japanese data, we are not committed to Multiset-CCG.

CCG-GTRC

There is another extension of CCG, which involves the use of variables in type raising [Komagata, 1997c; Komagata, 1997a]. This extension is motivated by the constituency of NP sequences in Japanese.[6] A sequence of NPs can form a non-traditional constituent with respect to coordination, as seen in (105) repeated below.

(129) John-ga Mary-o , Ken-ga Naomi-o tazuneta.
{ John-NOM Mary-ACC (and) Ken-NOM Naomi-ACC } visited
"John visited Mary and Ken, Naomi."

This is a very common construction frequently found in real text (there is an example (1) on p. 211). Unfortunately, this case has been neglected from legitimate analyses. By type-raising the NPs, CCG can provide a straightforward analysis of NP sequences corresponding to (129), as shown below.

(130)

John- ga	Mary- o	
NOM	ACC	
NP	NP	
$\Downarrow$	$\Downarrow$	type raising
$S/\underline{(S\backslash NP)}$	$\underline{(S\backslash NP)}/((S\backslash NP)\backslash NP)$	
$S/((S\backslash NP)\backslash NP)$		functional composition

The NPs are assigned type-raised categories associated with the basic category NP, and these functions can compose to derive another function category, which represents the NP-NP sequence. The two instances of such a category can then be coordinated and/or take the transitive verb category, $(S\backslash NP)\backslash NP$, as the argument to derive the sentence category S.

[6] Earlier analyses of Japanese using Categorial Grammar include [Kurahone, 1983 (focus on verb semantics)] and [Whitelock [1988] (focus on morphology)].

If the length of NP sequence is not bounded, we are forced to extend the formalism with a mechanism that can handle NP sequences of potentially infinite length. A natural move is to use variables in type raising as follows:

(131) Variable type raising:

$$a.\quad X : a \implies \mathsf{T}/(\mathsf{T}\backslash X) : \lambda F.F(a)$$

$$b.\quad X : a \implies \mathsf{T}\backslash(\mathsf{T}/X) : \lambda F.F(a)$$

Note: T is a variable over categories.

We may also abbreviate the above two with corresponding directionality as follows:

(132) $$X : a \implies \mathsf{T}\Big\langle\Big(\mathsf{T}\Big\rangle NP_1\Big) : \lambda F.F(a)$$

Then, unbounded length of NP sequence can be analyzed as a constituent, e.g., $\mathsf{T}\Big\langle\Big(\mathsf{T}\Big\rangle X_1 \ldots \Big\rangle X_k\Big)$. The resulting category may be called 'generalized type-raised categories' (GTRC). We abbreviate the extension of CCG with GTRC as CCG-GTRC.

4.1.5 Generative Power and Theoretical Parsing Efficiency

The most notable milestone regarding the generative power of CCG in relation to other formalisms including Tree-Adjoining Grammar (TAG) [Joshi et al., 1975; Joshi, 1985] and Linear Index Grammar (LIG) is Vijay-Shanker et al. [1986] and Weir and Joshi [1988], also published as Joshi et al. [1991], and more recently reworked as Vijay-Shanker and Weir [1994]. These formalisms are among the class called 'mildly context-sensitive grammars'. The finding of these papers is that these formalisms are all weakly equivalent (i.e., with respect to string generation capacity but not structural isomorphism). This finding is also important in relation to the processor. There is a class of automata called Embedded Push-down Automata that processes exactly the class of these grammars [Vijay-Shanker, 1988, Chapter 3].

All three variants of Multiset-CCG developed by Hoffman retain desirable formal properties: they are mildly context-sensitive.[7]

As for CCG-GTRC, one might be concerned about the use of variables that may introduce unexpected effects. It is not apparent whether the resulting formalism retains the same computational

[7] One variant of Multiset-CCG (Curried Multiset-CCG) is more powerful than the standard CCG, but the other two (Pure and Prioritized Multiset CCG) are incomparable to the standard CCG in this respect Hoffman [1995, Section 4.1.2].

properties as before. A general use of variables in a variant of categorial grammar makes it difficult even to demonstrate decidability [Emms, 1993]. But, since the use of variables in CCC-GTRC is fairly limited, an intuition is that it does not much increase the power of the formalism. My earlier paper [Komagata, 1997c] investigated all the possible occurrences of GTRCs in combinatory rules and argued that, with certain conditions, CCG-GTRC is weakly equivalent to the standard CCG. The main idea of the weak equivalence between CCG-GTRC and the standard CCG is that every derivation in CCG-GTRC can be simulated in the way the languages generated by the two grammar instances are exactly the same. The simulation uses the idea related to 'wrapping' [Bach, 1979; Dowty, 1979]. The propositions are proved by an extensive use of mathematical induction on the structure of derivation.

Another issue is theoretical parsing efficiency . Naturally, it is highly desirable that our grammar exhibits some polynomial parsing algorithm, as in the case of Context-Free Grammar (CFG), where CKY-style parsing algorithm has the $O\left(n^3\right)$ worst-case performance [Aho and Ullman, 1972, p. 317, for analysis]. But, since the number of categories in a CKY table cell is not bounded for CCG, a naive CKY-style algorithm for CCG does not have a polynomial bound.

There is a potential computational problem with accepting a wider variety of constituents. As seen above, multiple derivations may derive multiple instances of a single category (e.g., through a traditional and a non-traditional derivations). This situation is often called **spurious ambiguity** [Wittenburg, 1986]. If this kind of ambiguity is left untreated in the process, the number of categories being processed can easily explode in an exponential manner. Theoretical and practical solutions to this situation are discussed in Subsection 6.2.2.

Through a careful study of the properties possessed by CCG categories, Vijay-Shanker and Weir [1990] present a worst-case polynomial parsing algorithm for CCG. Later, they presented a more general algorithm covering several mildly context-sensitive grammar formalisms [Vijay-Shanker and Weir, 1993]. Their polynomial parsing algorithm employs a structure sharing technique [Billot and Lang, 1989; Dymetman, 1997] for efficient storage of potentially unboundedly-long categories. Crucially, the proposed structure sharing does not suffer from the existence of spurious ambiguities. Although this result alone does not demonstrate the practicality of the formalism, one without this property is unlikely to be practical.

The CCG formalism ('standard' CCG) used in the above comparison consists of combinatory rules: functional application and functional composition of fixed *k*. The coordination is not included as a rule but basically the same effect can be achieved by categories such as $S\backslash S/S$.[8]

All three variants of Multiset-CCG developed by Hoffman [1995] are polynomially parsable. Similarly, my earlier paper [Komagata, 1997a] shows that CCG-GTRC is polynomially parsable. The worst-case polynomial algorithm for CCG-GTRC is an extension of the polynomial algorithm for the standard CCG. The algorithm for CCG-GTRC also utilizes the idea of structure sharing for efficient storage and retrieval of GTRCs. A more detailed discussion of CCG-GTRC including formal and computational properties is found in Appendix A.

4.2 Specification of Contextual-Link Status

In this section, we confirm that the specification for contextual-link status can be formalized with the CCG framework. The discussion includes: discourse status, domain-specific knowledge, and linguistic marking for contextual link and information structure.

Discourse Status

For each CCG-constituent recognized by the grammar, there are corresponding semantic representations. These semantic representations can be used as discourse referents in a general sense (Subsection 3.2.2).

In order to formalize the notion of discourse-oldness, we need the following: (i) a mechanism to store all these objects and (ii) a mechanism to search through the storage for redundancy. Identification of discourse-old status checks the applicability of the identity relation on semantic representations between the semantic representation under consideration and one in the storage.

As we have mentioned earlier (Section 2.2), lack of exact reference resolution is not very crucial to information-structure analysis, especially for the case where the expression linguistically marks discourse-oldness. If the contextual-link status can be determined only through discourse-oldness and this depends on exact reference resolution, the formalization based on the use of

[8] While the coordination *schema* can deal with any category, coordination *category* can deal with only a closed set of categories. But this is not a limitation in practice. Another point is that the coordination category can compose with a functor. For example, $NP\backslash NP/NP$ may compose with a determiner NP/N if there is no further restriction.

semantic representations (rather than semantic values) would fail to recognize the correct discourse status. But, as we will see in Chapter 7, such cases rarely occur in our domain. The difficulty associated with INFERRABLE is far more common.

Domain-specific Knowledge

For the domain-specific knowledge, we only assume that physician(s) and patient(s) are available in the initial context regardless of the discourse. We can formalize this by simply asserting properties *physician* and *patient* in the initial context. Then, when these nouns are used in the text, they appear as if they were discourse-old, and can be identified as a contextual link.

Linguistic Marking

Linguistic marking is the case where the grammatical information is required. The mechanism of contextual-link assignment and projection is straightforward. For example, definite determiners of a category NP/N can assign a contextual-link status as specified on the result category, NP in this case, as shown below.

(133)

	Definite determiner	Noun
Example:	*the*	*door*
Syntactic type:	$\underset{CL}{NP}/\underset{-}{N}$	$\underset{CL \text{ or } NL}{N}$
Syntactic type:	$\underset{\boxed{CL}}{NP}$	

The specification of the contextual-link status, which may be realized as a feature, is shown below the result category, NP. The indefinite article is analogous, but it assigns a non-contextual-link status instead.

In Section 3.3, we have also discussed special cases with definite and indefinite articles. This introduces more complication to the above story. First, definite expressions with a special pre-nominal modifier such as *first* and *last* may be non-contextual links (p. 65). Thus, to be precise, contextual-link assignment of a definite determiner must check the lexical instantiation of the argument (e.g., through semantics). It should not categorically assign a contextual-link status if the semantics involves one of the special pre-nominal modifiers. But we do not formalize this particular aspect because this is not critical in our experiment data (Chapter 7).

Another exceptional case is indefinite INFERRABLES (p. 68). The point was that the main class of indefinite INFERRABLES are lexically marked, i.e., as two-place common nouns. Thus, these nouns can be marked as a contextual link. As in the case of "*a page (of a book)*" (65*b*), a countable two-place noun may be attached with an indefinite article. According to the description of indefinite article above, it is a non-contextual-link assigner. But there may be another type of indefinite article that contrasts with other quantifiers, but does not assign non-contextual-link status to these two-place nouns.[9]

Utterance-initial modifiers are also similar to the definite determiner except that only the utterance-initial variety, i.e., S/S, assigns a contextual-link status. The post-modifier type $S\backslash S$ does not have any special function.

Projection of contextual-link status can be done by using variable unification, as shown for a non-definite determiner below.

(134)

	Determiner	Noun
Example:	*many*	*researchers*
Syntactic type:	$\frac{NP}{X}/\frac{N}{X}$	$\frac{N}{Status}$
Syntactic type:	$\boxed{\frac{NP}{Status}}$	

This class includes auxiliary verbs and coordinators (for multiple arguments).

The other case of projection from the functor is similar. But the contextual-link information is carried over from its own contextual-link status as follows:

(135)

	Pre-modifier	Noun
Example:	*exercise*	*program*
Syntactic type:	$\frac{N}{Status}/\frac{N}{-}$	$\frac{N}{CL \text{ or } NL}$
Syntactic type:	$\boxed{\frac{N}{Status}}$	

As we have discussed in Subsection 2.3.3, direct information-structure marking is a matrix-level phenomenon. Thus, our grammar must be able to distinguish between the matrix and embedded environment. One way to do this is to assume utterance boundary categories, say $\$/S$ and/or

[9]Combined with the analysis of indefinite generics (p. 68), there is another possibility that an indefinite article actually projects the contextual-link status of the argument. If this is the case, the distinction between indefinite articles and other non-definite determiners disappears.

$\backslash S$, and assign the matrix feature only to the immediately composed *S*. We may associate the latter category with the period for the case of written text. A possible semantics for such a category is 'assertion' of the proposition corresponding to the category *S*. This point naturally connects to dynamic semantics (Subsection 2.3.2).

The special constructions in English discussed in Section 3.3.2 involves linguistic marking of both contextual link and information structure. A simpler case is pseudocleft. We analyze it simply as a contextual-link assigner, as in the case of a definite determiner. The subordinator *since* as a theme marker can be specified for the status with a feature, which may be checked at the time the information structure is identified.

VP preposing and inversion mark the "*Theme – Rheme*" partition. These cases require special syntactic types that license these constructions. For example, inversion of a *PP* may need a special *PP* category such as "$S/\underset{\text{subj}}{NP}/(S\backslash\underset{\text{verb}}{NP}/PP)/\underset{\text{arg of } PP}{NP}$". There are different ways of characterizing such a construction, but this category assumes that the exceptional behavior comes from the fact that *PP* is preposed and not from the verb or the subject. Since this is a matrix-level phenomena, we may require that this category is available only immediately to the right of utterance-boundary category $\$/S$. This can be done by, e.g., making the preposed *PP* "$S/\underset{\text{subj}}{NP}/(S\backslash\underset{\text{verb}}{NP}/PP)/\underset{\text{arg of } PP}{NP}\backslash(\$\backslash S)$". Once inversion is available only at the matrix level, we only need to mark the *PP* as a contextual link. As the *PP* is involved in the last semantic composition, *PP* is identified as a theme.

The conditions involved in topicalization, focus movement, and left dislocation are rather complex. Three different cases of information-structure marking must be considered. The category for the preposed NP are (i) $S/(S/NP)$ for the case of topicalization and focus movement and (ii) S/S for the case of left dislocation.[10] Since these are matrix-level phenomena, we can use contextual-link assignment for specifying information structure. We have seen that these constructions weakly partition theme and rheme. Thus, the preposed NP for all three cases may either set or reset the contextual-link status on itself. The contextual-link status of the remaining part of the utterance is determined in relation to that of the preposed NP.

For the case of topicalization, the topicalized NP is a contextual link and must be a part of the theme. The remaining part must contain a rheme due to the assumption (49) in Section 2.2. But it may also contain a part of the theme. This is consistent with the analysis (74) repeated below.

[10]Other cases which prepose non-NPs are analogous.

(136) [This dream] [I've had t] [maybe three, four times]]

Theme *Rheme*

For the case of focus movement, the preposed NP is assigned a non-contextual-link status. Thus, it must be a part of the rheme. If the remaining part is a contextual link, a "*Rheme – Theme*" pattern emerges as shown in (76) repeated below.

(137) [**A bite**] [he wouldn't eat t]

Rheme *Theme*

Summary

While the inference behind a contextual link can be complex, the three conditions for identifying a contextual link can be formalized within a grammar in a fairly straightforward manner.

4.3 Integration of Structured Meaning

One way in which we have departed from standard Montagovian, and from the standard CCG as well, is the use of structured meaning as semantic representation. As has been discussed in Section 3.5, this approach has the advantage of keeping the simple binomial information structure and being systematic.

The description in Section 3.5 demonstrates that structured meaning *can* be used for representing discontiguous information structure but does not show how it can be done. Further investigation of the way structured meanings are derived shows that it is not a simple issue.

In the following subsection, we demonstrate that the structured-meaning analysis approach presented in Section 3.5 can be integrated within CCG in a precise manner. In the second subsection, we demonstrate that the same mechanism can be applied to the analysis of 'gapping', i.e., a construction of the form "*Harry will buy bread, and Barry, potatoes*".

4.3.1 Composition of Structured Meanings

In our case, the structured meaning is adopted to capture the contrast between a contextual link and a non-contextual link. Its representation is as follows: $\langle Contextual\text{-}link, Non\text{-}contextual\text{-}link \rangle$ or $\langle C, N \rangle$ for short. We now assume that each constituent is associated with a structured meaning

in addition to a non-structured representation. The semantic composition in the general case is to combine $\langle C_1, N_1 \rangle$ and $\langle C_2, N_2 \rangle$ to obtain $\langle C', N' \rangle$ where C' and N' must be determined from the input components depending on the condition. In the following, we check distinct cases depending on the type of input. We denote semantic composition of structured meaning (and also its component) as "$\langle C_1, N_1 \rangle + \langle C_2, N_2 \rangle$" (here, '$+$' is used to separate the two categories). Two special cases are $\langle C, - \rangle$ and $\langle -, N \rangle$ where the entire semantic representation is either a contextual link or a non-contextual link. '$-$' here indicates a null component. Although this approach is naturally more complicated than the case without structured meaning, all possibilities can be completely specified.

Composition Type: $\langle -, N_1 \rangle + \langle -, N_2 \rangle$

This case is exactly like the usual semantic composition. We can simply operate on the non-link field as the following example shows.

(138)
$$\frac{\text{Felix} \quad \left\langle -, \underset{\text{non-link}}{\lambda P.P\,(felix')} \right\rangle \qquad \text{praised} \quad \left\langle -, \underset{\text{non-link}}{\lambda X.\lambda Y.praise'\,(X)\,(Y)} \right\rangle}{\left\langle -, \underset{\text{non-link}}{\lambda X.praise'\,(X)\,(felix')} \right\rangle}$$

In this case, the component $\lambda X.praise'\,(X)\,(felix')$ is obtained by functional composition. But, in general, either functional application or composition may apply.

Composition Type: $\langle C_1, - \rangle + \langle -, N_2 \rangle$

This is a representative case of forming a structured meaning.

(139)
$$\frac{\text{Felix praised} \quad \left\langle \lambda X.praise'\,(X)\,(felix'), - \right\rangle \qquad \textbf{Donald} \quad \left\langle -, donald' \right\rangle}{\left\langle \underset{\text{contextual link}}{\lambda X.praise'\,(X)\,(felix')}, \underset{\text{non-link}}{donald'} \right\rangle}$$

Further composition involving this type of structured meaning gets more complicated. The analysis for the mirror image, $\langle -, N_2 \rangle + \langle C_1, - \rangle$, is analogous.

Composition Type: $\langle C_1, N_1 \rangle + \langle -, N_2 \rangle$

First, we should note that the surface order of the components C_1 and N_1 for $\langle C_1, N_1 \rangle$ can be either $C_1 - N_1$, or $N_1 - C_1$. The following is an example for the $N_1 - C_1$ ordering (e.g., as a response to "Who praised who?"):

(140) **Felix** praised **Donald**

a. $\dfrac{\left\langle -, \lambda P.P\left(felix'\right)\right\rangle \quad \left\langle \lambda X.\lambda Y.praise'\left(X\right)\left(Y\right), -\right\rangle}{\left\langle \lambda X.\lambda Y.praise'\left(X\right)\left(Y\right), \lambda P.P\left(felix'\right)\right\rangle} \quad \left\langle -, \lambda P.P\left(donald'\right)\right\rangle$

b. $\dfrac{\left\langle \lambda X.\lambda Y.praise'\left(X\right)\left(Y\right), \lambda P.P\left(felix'\right)\right\rangle}{\left\langle \lambda X.\lambda Y.praise'\left(X\right)\left(Y\right), \lambda P.P\left(donald'\right)\left(felix'\right)\right\rangle}$

Let us focus on the second semantic composition (*b*). The question here is how we can obtain from $\lambda P.P\left(felix'\right)$ and $\lambda P.P\left(donald'\right)$ the correct $\lambda P.P\left(donald'\right)\left(felix'\right)$, but not $\lambda P.P\left(felix'\right)\left(donald'\right)$. The answer is that the corresponding syntactic composition is complete with the correct semantic representation. Any non-link in the above derivation that cannot derive the correct semantics after composing with $\lambda X.\lambda Y.praise'\left(X\right)\left(Y\right)$ should be discarded. Naturally, only $\left[\lambda P.P\left(donald'\right)\left(felix'\right)\right]$ $\left(\lambda X.\lambda Y.praise'\left(X\right)\left(Y\right)\right)$ can result in the correct semantics $praise'\left(donald'\right)\left(felix'\right)$, and not $\lambda P.P\left(felix'\right)\left(donald'\right)$. Thus, $\lambda P.P\left(felix'\right)\left(donald'\right)$ should be rejected. Similarly, the other ordering of composition $\left[\lambda X.\lambda Y.praise'\left(X\right)\left(Y\right)\right]$ $\left(\lambda P.P\left(donald'\right)\left(felix'\right)\right)$ should be rejected because it does not result in the correct semantics.

If structured meaning is used only for identifying a pair of contextual-status, the above-mentioned semantic check may be sufficient. In the next subsection on 'gapping', we observe a possibility that this process may also involve syntactic types.

The following is for the other surface ordering, $C_1 - N_1$.

(141) Felix **praised** **Donald**

a. $\dfrac{\left\langle \lambda P.P\left(felix'\right), -\right\rangle \quad \left\langle -, \lambda X.\lambda Y.praise'\left(X\right)\left(Y\right)\right\rangle}{\left\langle \lambda P.P\left(felix'\right), \lambda X.\lambda Y.praise'\left(X\right)\left(Y\right)\right\rangle} \quad \left\langle -, \lambda P.P\left(donald'\right)\right\rangle$

b. (see below)

The focus is again (*b*). In this case, the other derivation, "[Felix] [praised Donald]", is more favorable because both non-links are combined together without complication. Thus, if an alternative derivation is available, we do not need to consider this case. If the alternative derivation is not available for some reason, we are forced to derive $\left\langle -, praise'\,(donald')\,(felix')\right\rangle$ because there is no specification that can upgrade a non-contextual-link material to a contextual link.

Composition Type: $\langle C_1, -\rangle + \langle C_2, -\rangle$

When two contextual links are composed, the resulting phrase must be identified as contextual link by one of the three properties. This case happens when a complex phrase is discourse-old or a special linguistic marking is present. For example, if "*Felix praised*" is already in the context, the following derivation is possible.

(142)
$$\frac{\begin{array}{cc}\text{Felix} & \text{praised}\\ \left\langle \lambda P.P\,(felix'), -\right\rangle & \left\langle \lambda X.\lambda Y.praise'\,(X)\,(Y), -\right\rangle\end{array}}{\left\langle \lambda X.praise'\,(X)\,(felix'), -\right\rangle}$$

If the resulting unit is not contextually-linked, "$\langle C_1, -\rangle + \langle C_2, -\rangle$" can only result in either $\langle C_1, C_2\rangle$ or $\langle C_2, C_1\rangle$. That is, only one of them can remain as a contextual link and the other is considered a non-contextual link. This pattern is the source of a contextually-linked rheme. But the proposition is not a contextual-link. In this case, the rheme (either C_1 or C_2) must be contrastive as we discussed in Chapter 2, but we do not go into this point in our formalization or implementation.

Composition Type: $\langle C_1, N_1\rangle + \langle C_2, -\rangle$

This case is in a sense a combination of the previous two cases. Let us only consider the subcase where the component ordering of $\langle C_1, N_1\rangle$ is $N_1 - C_1$. For C_1 and C_2 to be combined, the resulting unit must be a contextual link through one of the three possibilities. If the combination of C_1 and C_2 is not a contextual link as a whole, only the full contextual link would survive in the result, as shown below.

(143) **Felix** praised Donald

a. $\dfrac{\left\langle -, \lambda P.P\left(felix'\right)\right\rangle \quad \left\langle \lambda X.\lambda Y.praise'\left(X\right)\left(Y\right), -\right\rangle}{\left\langle \lambda X.\lambda Y.praise'\left(X\right)\left(Y\right), \lambda P.P\left(felix'\right)\right\rangle} \quad \left\langle \lambda P.P\left(donald'\right), -\right\rangle$

b. $\dfrac{\left\langle \lambda X.\lambda Y.praise'\left(X\right)\left(Y\right), \lambda P.P\left(felix'\right)\right\rangle}{\left\langle \lambda P.P\left(donald'\right), \lambda X.praise'\left(X\right)\left(donald'\right)\right\rangle}$

Composition Type: $\langle C_1, N_1\rangle + \langle C_2, N_2\rangle$

Here, we consider the ordering $C_1 - N_1 - C_2 - N_2$. This case could end up with $\langle C', N'\rangle$, $\langle C_1, N''\rangle$, $\langle C_2, N'''\rangle$, or $\langle -, N''''\rangle$. The condition for $\langle C', N'\rangle$ is that $C_1 + C_2$ is a contextual link for its own reason and $N_1 + N_2$ can compose to result in a legitimate category. These two intermediate results must compose to the category corresponding to the entire phrase. If $C_1 + C_2$ is not a contextual link, other cases may still apply. For the case where the result is $\langle C_1, N''\rangle$, $\langle -, N_1\rangle + \langle C_2, N_2\rangle$ should not be available. If so, the bracketing $C_1 - [N_1 - C_2 - N_2]$ is available for a simpler derivation. The case for $\langle C_2, N'''\rangle$ is analogous. The last case applies when the previous three fail.

The following is an analysis of (95) on p. 85 assuming that $\lambda X.\lambda Y.want'\left(X\right)\left(lose'\left(Y\right)\left(pro\right)\right)\left(alice'\right)$ corresponding to "*Alice wants – to loose*" is a contextual link.

(144)
$$\dfrac{\dfrac{\text{[Alice wants]}_{Theme} \quad \text{[}\textbf{Australia}\text{]}_{Rheme}}{\left\langle \lambda X.\lambda Y.want'\left(X\right)\left(Y\right)\left(alice'\right), -\right\rangle \quad \left\langle -, \lambda P.P\left(australia'\right)\right\rangle} \quad \dfrac{\text{[to lose]}_{Theme} \quad \text{[the }\textbf{Ashes}\text{]}_{Rheme}}{\left\langle \lambda X.lose'\left(X\right)\left(pro\right), -\right\rangle \quad \left\langle -, \lambda P.P\left(ashes'\right)\right\rangle}}{\left\langle \lambda X.\lambda Y.want'\left(X\right)\left(lose'\left(Y\right)\left(pro\right)\right)\left(alice'\right), \lambda P.P\left(australia'\right)\left(ashes'\right)\right\rangle}$$

Composition Type: Coordination

One additional case is coordination. For a coordination of the type "$\langle C_1, -\rangle + \& + \langle C_2, -\rangle$", we adopt the following condition:

(145) *a.* $\langle C', -\rangle$ if the coordination of C_1 and C_2, i.e., C', is a contextual link (where $C' = C_1 + C_2$)

b. $\langle -, N''\rangle$ otherwise (N' is the semantic representation for the entire phrase)

While a more fine-grained analysis is possible, e.g., coordination of $\langle C_1, N_1\rangle$ and $\langle C_2, N_2\rangle$, we only consider the above simple analysis.

Discontiguous Components

Up to here, we have been assuming that the components of a structured meaning are contiguous. But this is not always the case. For example, the composition of $\langle C_1, N_1 \rangle$ and $\langle -, N_2 \rangle$ (with the $N_1 - C_1 - N_2$ surface ordering) may end up with $\langle C_1, N' \rangle$ where the component N' is discontiguous. If there is a further composition of $\langle C_1, N' \rangle$ with another structured meaning, we cannot use the same condition for the contiguous case because the boundaries of $\langle C_1, N' \rangle$ is both N while the boundaries of some $\langle C_2, N_2 \rangle$ with contiguous components C_2 and N_2 are C_2 and N_2 (in either order). In order to close the operation of composition on structured meanings, we can only consider a finite number of subcases. One way to do this is to set up four possible boundary types, $N-N$, $N-C$, $C-N$, and $C-C$, and define the condition for these four subcases. We omit the actual conditions as it is tedious (commonly observed cases have been implemented and described in Chapter 6).

Complexity of Structured Meaning Representation

Naturally, the complexity introduced by the use of structured meaning is a concern. Here, we investigate the complexity of structured meaning and that of composing structured meanings.

First, the structural variation of structured meanings is limited to a pair of semantic representation with two additional cases where either of them is null. But each component can be discontiguous, as has been seen above. For a string of n lexical categories, each lexical category may belong to either C or N of $\langle C, N \rangle$. Thus, for the span of this n lexical categories, in theory, there are at most 2^n distinct structured meaning.[11] But, in practice, structured meanings with an internal division more complex than $C-N-C-N$ or $N-C-N-C$ are extremely rare. This is because in many cases, assignment or projection of a contextual-link status results in either $\langle C, - \rangle$ or $\langle -, N \rangle$ reducing the internal structure. Following the discussion on page 73 (in the last paragraph of the Summary), the two main sources of structured meanings are predicate-argument structure involving a main verb and modification structure involving a clausal modifier.

If the most complicated internal structure for a structured meaning is in practice 4-way, as in $C-N-C-N$, the practical bound on the number of distinct structured meanings for a single category is no more than the number of structure meanings for a 4-category sequence, i.e., $2^4 = 16$. This applies at every step of derivation. As a consequence, the overall increase of complexity due to

[11] This does not include various kinds of ambiguities.

introduction of structured meaning is in practice at most 16 times that of the case without structured meanings.

Next, let us discuss the complexity of composition involving structured meanings. The operation is closed because in addition to the structured meaning itself, we only recognize the contextual-link status of the boundary categories. As stated earlier, there are four boundary status pairs, $C-C$, $C-N$, $N-C$, $N-N$, for a structured meaning $\langle C,N\rangle$. There are two special cases $\langle C,-\rangle$ and $\langle -,N\rangle$ with no partition of the contextual-link status. For a composition of "$\langle C_1,N_1\rangle + \langle C_2,N_2\rangle$" resulting in $\langle C',N'\rangle$, we consider these 6 patterns for each input and result. A simplistic bound on all the possible combinations of the 6 patterns is $6\times 6\times 6 = 216$. This is a large number, but many of these patterns are not necessary in practice. For the current purpose, it is sufficient to show that there is a bound.

In conclusion, processing structured meanings is in practice multiplicative rather than exponential. This property is very important for the practicality of the use of structure meaning.

Summary

This subsection shows that composition of structured meaning can be done precisely and in practice does not increase asymptotic complexity. The conditions for composing structured meaning have been discussed, and summarized as follows:

(146) Conditions for semantic composition of structured meanings:

a. The semantic composition (either through functional application or functional composition) of the two components must be consistent with the semantic representation of the entire phrase.

b. The contextual-link component of a structured meaning (after composition) must satisfy the requirements for a contextual link (through discourse status, domain-specific knowledge, and/or linguistic marking).

Although the present application of structured meaning is for contextual-link status, the technique discussed in this section is applicable to structured meaning for contrast and other purposes.

4.3.2 Identification of Information Structure

Assuming that intermediate steps of compositions of structured meanings go well, identification of information structure is almost trivial. For the final structured meaning $\langle C, N \rangle$, we identify $Theme = C$ and $Rheme = N$. For example, again consider "Felix praised **Donald**." in response to "Who did Felix praise?" The last semantic composition is as follows:

(147)
$$\frac{\text{Felix praised} \quad \textbf{Donald}}{\left\langle -, \lambda X.praise'(X)(felix') \right\rangle \quad \left\langle -, donald' \right\rangle}$$
$$\left\langle \lambda X.praise'(X)(felix'), donald' \right\rangle$$
$$\text{contextual link} \downarrow Theme \qquad \text{non-link} \downarrow Rheme$$

There may be some discontiguity within the theme and/or the rheme. But the information structure can be identified exactly the same way as before. The present approach is an improvement over that in Komagata [1998a]. In that paper, I characterized information structure as the last step of semantic composition. But this approach without structured meaning cannot cover discontiguous information structure in a general way as the present formulation does.

If multiple structured meanings are available at the end, the current theory accepts all the available information structures. For our implementation (Chapter 6), though, we have a few heuristics to choose more likely information structures.

4.3.3 Analysis of Gapping

We have seen that the problem of discontiguous information structure for the binomial-partition analyses can be solved by adopting the structured-meaning approach. We have also noted that the structured meaning approach is applicable to other areas including the analyses of contrast, propositional attitude, and thematic role. In this subsection, we apply structured meaning to yet another phenomenon of 'gapping'. In particular, we recast Steedman's [1990] 'decomposition' analysis in terms of structured meaning.[12]

Gapping in English has a form shown below [Steedman, 1990, (85), p. 242].

(148) Harry will buy bread, and Barry, potatoes.

[12]The analysis presented here is not compatible with Steedman's [1999, Chapter 7] more recent analysis.

It has received much attention for its peculiar construction. While earlier analyses were purely syntactic, Kuno [1976] pointed out pragmatic factors involved in the construction (for an extensive review, see Steedman [1999, Chapter 7]). Since this is a phenomenon involving discontinuity and potentially information structure, it would be a good demonstration if the structured-meaning approach can be applied to it.

Steedman's [1990, (85), p. 242] analysis is shown below. Note that there is a information-structure condition for decomposition that the gap must be 'known' [p. 250].

(149)
$$\dfrac{\dfrac{\text{Harry will buy bread,}}{S} \quad \text{and} \quad \dfrac{\text{Barry,} \quad \overset{S/(S\backslash NP)}{\text{potatoes}}}{S\backslash(S\backslash NP/NP)}}{\dfrac{S\backslash NP/NP \quad S\backslash(S\backslash NP/NP)}{S\backslash(S\backslash NP/NP)}\,{<}decomp\quad}\,{<}\&{>}$$

In the following, we apply our structured-meaning approach to the above case. The derivation of the category *S* results in a structured meaning such that the verb and the arguments split as $\left\langle \underset{Contextual\text{-}link}{Verb}, \underset{Non\text{-}link}{Arguments} \right\rangle$, and the non-contextual-link component is coordinated with the right conjunct. Let us assume that the split with respect to the contextual-link status is a source of gapping corresponding to Kuno's [1976] intuition and Steedman's [1990] condition on the decomposition. Then, the present approach can provide the following analysis for the left conjunct.

(150)
$$\dfrac{\begin{array}{ccc}\text{Harry} & \text{will buy} & \text{bread,}\\ NP & S\backslash NP/NP & NP\\ \left\langle -, harry' \right\rangle & \left\langle \lambda X.\lambda Y.buy'(X)(Y), - \right\rangle & \left\langle -, bread' \right\rangle\end{array}}{\left\langle \lambda X.\lambda Y.buy'(X)(Y), \lambda P.P(bread')(harry') \right\rangle}$$

At this point, let us hypothesize that the semantic unit $\lambda P.P(bread')(harry')$ is available as a part of a 'virtual category'. It is not a real category because *Harry* and *bread* are discontiguous. If we only deal with semantic information, this might be enough (as we have been doing up to this point in this section). But, in order to proceed with the coordination with the right conjunct, we need to analyze the syntactic type of the virtual category as well. Assuming that both NP's have

a syntactic type $\mathsf{T}\backslash\left(\mathsf{T}/NP\right)$,[13] this virtual category might correspond to several distinct syntactic types as shown below.

(151)

	Harry	bread		*Harry, bread* [virtual category]	
a.	$\mathsf{T}/(\mathsf{T}\backslash NP_1)$	$\mathsf{T}/(\mathsf{T}\backslash NP_2)$	$\Longrightarrow$	$\mathsf{T}/(\mathsf{T}\backslash NP_1\backslash NP_2)$	$(>B)$
b_1.	$\mathsf{T}/(\mathsf{T}\backslash NP_1)$	$\mathsf{T}\backslash(\mathsf{T}/NP_2)$	$\Longrightarrow$	$\mathsf{T}\backslash(\mathsf{T}\backslash NP_1/NP_2)$	$(>B\times)$
b_2.	$\mathsf{T}/(\mathsf{T}\backslash NP_1)$	$\mathsf{T}\backslash(\mathsf{T}/NP_2)$	$\Longrightarrow$	$\mathsf{T}/(\mathsf{T}\backslash NP_2\backslash NP_1)$	$(<B\times)$
c.	$\mathsf{T}\backslash(\mathsf{T}/NP_1)$	$\mathsf{T}/(\mathsf{T}\backslash NP_2)$	$\Longrightarrow$	fail	
d.	$\mathsf{T}\backslash(\mathsf{T}/NP_1)$	$\mathsf{T}\backslash(\mathsf{T}/NP_2)$	$\Longrightarrow$	$\mathsf{T}\backslash(\mathsf{T}/NP_2/NP_1)$	$(<B)$

In the above, we have used forward crossing composition '$>B\times$', which is not generally assumed for English [Steedman, 1996, p. 53]. We will come back to this point shortly. First, the result (b_2) and (*d*) are excluded because of the semantic condition for composing structured meanings (146*b*). That is, the argument order of subject and object are incorrect, and thus cannot result in the correct semantic representation. Now, we extend this condition to syntactic type as well. This requires that the syntactic types of the components must be composed into the resulting syntactic type. Both the possibilities of having the virtual category to the left and right of the verb category are considered below.

(152)

a.	$\mathsf{T}/(\mathsf{T}\backslash NP_1\backslash NP_2)$	$S\backslash NP/NP$	$\Longrightarrow$	fail	
	$S\backslash NP/NP$	$\mathsf{T}/(\mathsf{T}\backslash NP_1\backslash NP_2)$	$\Longrightarrow$	fail	
b_1.	$\mathsf{T}\backslash(\mathsf{T}\backslash NP_1/NP_2)$	$S\backslash NP/NP$	$\Longrightarrow$	fail	
	$S\backslash NP/NP$	$\mathsf{T}\backslash(\mathsf{T}\backslash NP_1/NP_2)$	$\Longrightarrow$	S	$(<)$

Thus, the only possibility is that the virtual category is to the right of the verb and the variable T is instantiated as *S* with the correct argument order. This 'virtual' derivation is shown below.

(153)

will buy	*Harry, bread* [virtual category]
$S\backslash NP/NP$	$S\backslash(S\backslash NP/NP)$
$\left\langle \lambda X.\lambda Y.buy'(X)(Y), - \right\rangle$	$\left\langle -, \lambda P.P(bread')(harry') \right\rangle$
S	
$\left\langle \lambda X.\lambda Y.buy'(X)(Y), \lambda P.P(bread')(harry') \right\rangle$	

[13] We use the variable notation for conciseness and generality.

Although this scheme appears like decomposition, it is not exactly the type of decomposition proposed in Steedman [1990]. This is because the identification of the virtual category is done constructively at the same time as the category *S* is derived.

The syntactic type, semantic type, and relative position of the virtual category license the following coordination with the right conjunct.

(154) *Harry, bread* and Barry, potatoes

[virtual category]

$$\frac{\mathsf{T}\backslash\left(\mathsf{T}/NP_1\right) \qquad \mathsf{T}\backslash\left(\mathsf{T}/NP_2\right)}{\begin{cases} \mathsf{T}/(\mathsf{T}\backslash NP_1\backslash NP_2) & \text{fail} \\ \mathsf{T}\backslash(\mathsf{T}\backslash NP_1/NP_2) & \text{success} \\ \mathsf{T}\backslash(\mathsf{T}/NP_2/NP_1) & \text{fail} \\ \mathsf{T}/(\mathsf{T}/NP_2\backslash NP_1) & \text{fail} \end{cases}}$$

$$S\backslash(S\backslash NP/NP)$$

$$\frac{\left\langle -, \lambda P.P\,(br)\,(h)\right\rangle \qquad \left\langle -, \lambda P.P\,(p)\,(ba)\right\rangle}{\left\langle -, \lambda P.and\left(P\,(br)\,(h)\right)\left(P\,(p)\,(ba)\right)\right\rangle}\,\text{<\&>}$$

$$S\backslash(S\backslash NP/NP)$$

Then, this can compose with the verb to derive the desired *S* category with the intended semantic representation.

There is one more point we should address. Forward crossing composition $(>B\times)$, which is crucial to the derivation of the virtual category and the right conjunct, is not normally allowed in the surface grammar of English [Steedman, 1996, p. 53]. In a sense, this would incorrectly predict 'scrambling' of arguments at the left of a verb. The current position to compromise the demand for forward crossing composition in the above analysis and this constraint is the following. Forward crossing composition is available even in English (for both surface and virtual cases). But the result of this process is available only for compositions and coordination involving a virtual category.

The above analysis of gapping in terms of structured meaning demonstrates usefulness of structured meaning beyond the current project. It also suggests that the structured meaning may involve syntactic types, as well as semantic representation.

4.4 Summary

This chapter demonstrates CCG's advantages in (i) recognizing non-traditional constituents in accordance to units of information structure, (ii) capturing the properties for contextual links, and (iii) integrating structured meaning for analysis of discontiguous information structure. The formalization congruently integrates syntax, semantics, and discourse status, and provides a basis for bridging the theory (Chapter 3) and the implementation (Chapter 6) in a straightforward manner.

Chapter 5

Realization of Information Structure in Japanese

The goal of this chapter is to justify the use of linguistic marking of information structure in Japanese for evaluation purposes. While much has been said about Japanese particles and scrambling, there are few analyses made from the view point of modern information-structure analysis.

Since the object language, Japanese, is quite different from English in many respects, the first section makes an introduction to the language. In Sections 5.2 and 5.3, we present analyses of two most crucial elements: functions of particle *wa* and long-distance fronting, respectively. Based on these analyses, Section 5.4 analyzes linguistic marking of information structure as a result of these elements, and presents a procedure to predict *wa* or *ga* from information structure.

5.1 Introduction

This section briefly presents some background on the Japanese language, introduces the relevant linguistic properties, and previews the arguments explored in the following sections.

Before moving on to the focal issues, let us make a brief note about the Japanese language.[1] Japanese is a strictly head-final, SOV language. It is sometimes classified as an agglutinative language due to its morphological generativity, especially the verb morphology involving aspect, negation, voice, causativity, and even politeness. NPs are usually marked with particles including

[1] Shibatani [1990] is an excellent introduction to the language for non-Japanese-speaking readers.

case particles and adverbial particles.[2] Japanese does not have a determiner system corresponding to the one in English. In particular, formal definite/indefinite distinction is not in general available in Japanese. This brings an interesting contrast with English, which does not have an extensive system of direct information-structure marking in the written form. At the matrix level, the definite/indefinite distinction of the subject in English closely corresponds to the use of morphological particles *wa*/*ga* (respectively) on the subject in Japanese. But this observation is limited to the matrix level, and does not extend to embedded environments. But, since our theory of information structure is based on the notion of contextual link (Section 3.1), we suspect that the relation between contextual link and information structure might be roughly the relation between definiteness in English and morphological marking in Japanese.

While a lot of work has been done in this area and a great deal of discovery has been made, there are still many remaining issues. Unfortunately, the previous work are not necessarily as precise nor as accurate as we require for the current purposes including computational implementation. One general problem is that the literature tends to have narrow viewpoints. Approaches from theoretical syntax take up the topic of our interest but critical elements in pragmatics are often ignored [e.g., Tateishi, 1994]. On the other hand, discourse/pragmatic analyses tend to focus on the description of phenomena and do not provide us with theories useful for our purposes [e.g., Watanabe, 1989; Shimojo, 1995; Noda, 1996]. Formal and computational analyses typically start from assumptions too simplistic to cover realistic data [e.g., Uetake, 1992; Porter and Yabushita, 1998].[3]

Let us briefly look at the case of the adverbial particle *wa*.[4] This particle is often associated with 'thematic' and 'contrastive' functions [e.g., Kuno, 1972]. But the situation surrounding this particle is rather complicated. First, the nature of the functions is not entirely clear, reflecting a difficulty with many related notions. For example, we cannot assume that the 'thematic' function of Kuno [1972] coincides with our 'theme'. In addition, we need to distinguish the notions of referential status and information structure as we have been doing so far. Second, the distribution of these functions is not sufficiently explored. Assuming that they have distinct roles for these

[2]Nominal constructions suffixed with particle(s) are called either NP [Shibatani, 1990] or PP (postpositional phrase) [Gunji, 1987]. Some work distinguishes between these two [Sadakane and Koizumi, 1995]. A recent analysis on various particles can be found in Siegel [1999].

[3]Uetake [1992]; Porter and Yabushita [1998] do not consider contrastive *wa*, which we will cover in the next section.

[4]I follow Shibatani [1990] in using the term 'adverbial particle' but other terms are also used (esp. in the Japanese linguistics literature written in Japanese)

types, we need to distinguish these functions. Furthermore, the relation between these functions is a theoretically interesting issue on its own. Another critical aspect is the relation between the adverbial particle *wa* and case particles. For example, the choice between an adverbial particle *wa* and a nominative case particle *ga* is often completely pragmatic,[5] and can pose a great problem for a NL generation system. This point was mentioned but not explored at all in Nagao [1989]. The only other description known to the author is a generation system of Matthiessen and Bateman [1991].

Another well-discussed aspect about Japanese is 'scrambling'. Scrambling is often classified as local (clause-bounded) and long-distance (unbounded) varieties [Gunji, 1987, p. 219-220]. In the current work, we call them 'local scrambling' and 'long-distance fronting' (or fronting for short), respectively. Since long-distance fronting is more closely related to information structure, we will focus on this type. The function of local scrambling is not very clear and is left out in the current work [cf. Miyagawa, 1997]. A simplistic idea about long-distance fronting is that it is 'topicalization', i.e., to separate a theme [e.g., Kiss, 1981]. But this construction can also serve fronting constituents for emphatic purpose [Gunji, 1987, p. 218]. We will explore a solution in Section 5.3.

In relation to the functions of particle *wa* and long-distance fronting, we should note one more phenomenon, which we do not discuss any further in this thesis. It is an outermost *wa*-marked constituent (often called 'major subject') that does not appear to be an argument of the main predicate, as shown below (the following grammatical labels are used: TOP = topic, NOM = nominative; the complete list of grammatical functions is on p. xiv).[6]

(155) Sakana-wa tai-ga ii.
fish-TOP red snapper-NOM excellent
"As for fish, a red snapper is excellent."

The utterance is propositionally complete without the *wa*-marked phrase. Thus, it is not obvious how the *wa*-marked phrase is grammatically related to the proposition although the connection is not unreasonable at the knowledge level. Among many analyses of this type, Tateishi [1994, p. 28] argues that a major subject is at Spec of CP, and Gunji [1987, p. 171] argues that it is an

[5] Although *ga*-marking is possible on some objects, e.g., "Ken-wa Naomi-ga sukida" (*Ken likes Naomi*), such a case is excluded from the current work.

[6] This is an often-discussed example in the literature. See Noda [1996, p. 54] for more details.

adjunct. Before closing this introductory section, let us discuss a few more points. The first one is that the previous literature mostly ignores the importance of phonological prominence (except for a relatively old paper [Finn, 1984]). In order to take advantage of the effect of phonological prominence, this chapter primarily focuses on the spoken form. On the other hand, we discuss little phonological aspects themselves. One assumption in this chapter is that phonological prominence is observable in Japanese.[7] For text analysis, unfortunately, we cannot access this information, and we will need to deal with underspecified cases.

Second, in Japanese, a sequence of NPs can form a constituent in a fairly general manner. The situation can be observed in relation to coordination and information structure as follows:[8]

(156) *a*. { Ken-wa banana-o , Naomi-wa mango-o } tabeta.
Ken-TOP banana-ACC (and) Naomi-TOP mango-ACC ate
"Ken ate a/the banana, and Naomi [ate] a/the mango."

b. { Ken$_i$-wa banana$_j$-o }
Ken-TOP banana-ACC

[t_i [Sara-ga t_j tabeta] -to omotta]. [9]
Sara-NOM ate -COMP thought
"Ken thought that Sara ate a/the banana."

Note: the fronted non-traditional constituent "*Ken-wa banana-o*" can be coordinated with another phrase of the same category.

While these are problems for most grammars, they can be accounted for in a general way in Combinatory Categorial Grammar (CCG) [Ades and Steedman, 1982]. A formal and computational analysis of the involvement of NP sequences in a general form is given in Appendix A. Many of the syntactic and semantic elements discussed in this chapter have been implemented in an earlier version of the CCG parser [Komagata, 1997a].Finally, we note that a closely related situation about particle use and long-distance fronting is observed in Korean.[10] We will take advantage of this situation and cite related work about Korean as well.

[7]It has been argued that a certain notion of 'prominence' in English can be identified computationally [Maghbouleh, 1996].

[8]The following grammatical labels are used: TOP = topic, NOM = nominative, ACC = accusative, and COMP = complementizer.

[9]The traces t_i/t_j are shown only for presentation uprposes. Our theory of grammar, based on Combinatory Categorial Grammar does not assume the notion of empty categories.

[10]The genetic relation between Korean and Japanese is still actively debated [e.g., Shibatani, 1990, Chapter 5].

Towards the end of this chapter, we will observe a distribution of functions such as the following table:

		Information structure	
		Matrix clause	Embedded clause
wa (adverbial particle)	**Prominent**	Theme or Rheme	Unspecified
	Non-prominent	Theme	Not available
ga (case particle)	**Prominent**	Rheme	Unspecified
	Non-prominent	Rheme	Unspecified
o, *ni* (case particle)	**Prominent**	Rheme	Unspecified
	Non-prominent	Theme or Rheme	Unspecified

Table 5.1: Realization of Information Structure in Japanese (preliminary)

This is a rather messy array of data, and more complicated than many previous analyses. While a result like this is still useful for computational applications, we must have a theoretical justification for it.

In the subsequent sections, we will make the following points for the present analysis of the linguistic marking of information structure:

1. The basic function of *wa* is a 'strong' contrastiveness, always associated with phonological prominence.

2. The thematic function of *wa* is available only as a result of long-distance fronting. Thematic *wa* need not be prominent.

3. Long-distance fronting in Japanese is a general-purpose constituent re-ordering mechanism. It typically sets up an information structure at the matrix level.

4. The linguistic marking of information structure in Japanese is a result of complex interaction of functions of particles and long-distance scrambling.

5.2 Functions of Particle *wa*

This section is divided into three subsections: introduction to the two functions of *wa*, and more details on contrastive and thematic functions.

5.2.1 Two Functions of *wa*

Kuno [1972], among others, argues that the particle *wa* has thematic and contrastive functions. This point can be seen in the following short discourses. As before, **boldface** indicates phonological prominence.[11]

(157) Thematic *wa*:

i. "Ken behaved strangely yesterday."

ii. Ken-wa **banana**-o **tabeta**.
Ken-TOP banana-ACC ate
"Ken ate a/the banana."

(158) Contrastive *wa*:

Q: "Among those people, who ate bananas?"

A: **Ken**-wa banana-o tabeta.
Ken-CONT banana-ACC ate
"Ken ate a banana (someone else didn't eat a banana)."

In (157), the first utterance introduces a person whose name is *Ken*, and the second utterance provides new information about *Ken*. In (158), the question sets a context. The response not only answers the question but also carries a presupposition indicated in '(...)'.[12] Although Kuno's description is that these two functions are exclusive and we frequently use the terms 'thematic *wa*' and 'contrastive *wa*', we do not mean that there are two distinct types of *wa*.

We continue to consider the same notion of theme (Section 3.1) and contrast (Section 2.3.2), and that particle *wa* exhibits both of these properties under certain circumstances (more on these points later). Thus, when we say thematic (contrastive) *wa* in this thesis, it means that the instance of *wa* is a part of a theme (has a contrastive interpretation). Since the theme property, i.e., information structure, and contrastiveness are basically independent, there is a case where both properties co-exist. This situation is suggested in Shibatani [1990, p. 265], and is described more explicitly for the Korean counterpart, *(n)un* in Han [1998, p. 2] and Wee [1995, Section 2.2]. The following example shows the overlapping case.

[11]The following grammatical labels are used: TOP = topic, CONT = contrastive, NOM = nominative, and ACC = accusative, DAT = dative, and COMP = complementizer.

[12]For an extensive review about presupposition, see [Beaver, 1997].

(159) Thematic/contrastive *wa*:

Q: "What did these people eat?"

A: **Ken**-wa **banana**-o tabeta.
Ken-TOP/CONT banana-ACC ate
"Ken ate a banana (someone else didn't eat a banana)."

Now, there are various different views about the relation between these two functions: (i) the two functions are independent [Tateishi, 1994, p. 175], (ii) the contrastive function is derivable from the thematic one [Miyagawa, 1987, p. 197; Noda, 1996, suggested in earlier chapters], (iii) the thematic function is derivable from the contrastive one [Shibatani, 1990, p. 265; Teramura, 1991, p. 41; Choi, 1997, p. 548], and (iv) both functions can be derived from a single basic function [Han, 1998, p. 1; Wee, 1995, Section 2.1 (both for Korean)].

As the way to analyze the particle *wa* depends on this issue, let us assume the position (iii) above and provide some justification as follows. The position (i) is not attractive because of the existence of the overlap. For example, the distinction in Tateishi [1994, Chapter 6], i.e., thematic *wa* as a determiner and contrastive *wa* as a modifier, is not applicable to the overlapping case. The position (ii) is not attractive from the distributional and historical points. While the distribution of thematic *wa* is limited to the utterance-initial position, that of contrastive *wa* is cross-categorial (including positions after another particle, verb, and adverb) [Aoki, 1992; Tateishi, 1994; Noda, 1996], much like English *only*. It is more natural to think that the narrower distribution is due to some restriction rather the opposite. Furthermore, historically speaking, thematic *wa* is believed to have developed much later than contrastive *wa* [Ueno, 1987, p. 242; De Wolf, 1987, p. 281]. The position (iv) is an attractive approach but also more difficult because we need to posit an abstract unified level, which tends to escape directly observable phenomena for evaluation.

We thus proceed by assuming that contrastive function is basic and relate the thematic function under special conditions.

5.2.2 Contrastive Function

This subsection shows that contrastive function is associated with phonological prominence and that it has a presupposition stronger than the case without *wa*-marking, and that the phenomenon can be analyzed in terms of Alternative Semantics [Rooth, 1985, and later work]. A more detailed

version of this subsection including a formalization is found in Komagata [1998b].

One immediate problem with most of the previous work is ignorance of phonological prominence. In addition, most of the previous work simply assumes the domain of contrastive *wa* is the preceding noun. But such an analysis would face a problem accounting for distinct presuppositions in the following example:

(160) *a*. Ken-wa [**Naomi**-no banana] -wa tabeta.
Ken-TOP Naomi-GEN banana -CONT ate
"Ken ate Naomi's banana."

Presupposition: "Ken didn't eat *someone else*' banana."

b. Ken-wa [Naomi-no **banana**] -wa tabeta.
Ken-TOP Naomi-GEN banana -CONT ate
"Ken ate Naomi's banana."

Presupposition: "Ken didn't eat *something else* of Naomi."

Only one paper came to my attention in this respect. Huruta [1982] considers *wa* suffixing on a complex NP such as the one shown above. But he ignores phonological prominence and is forced to accept the ambiguous situation.

Next, the studies primarily concerned with the contrast between the adverbial particle *wa* and case particle *ga* tend to overlook the cross-categorial distribution of contrastive *wa* [e.g., Kuno, 1972]. Discussion on contrastive *wa* is often limited to the individual-type NPs, but not extended to the case of *wa*-suffixing to the universal quantifier [Han, 1998, for a related example (10), p. 8].[13]

(161) *Q*: "Did Ken praise Naomi?"

A: Ken-wa **minna**-o/*wa hometa.
Ken-TOP everyone-ACC/CONT praised
"Ken praised everyone (in contrast to just Naomi)."

While "everyone" in (*A*) is in contrast to *Naomi* in (*Q*) and the accusative marker is possible, contrastive *wa* cannot be used in this utterance. This asymmetry is independent of the grammatical relations, the underlying case marking (on the *wa*-marked phrase), and scrambling of the *wa*-marked phrase. A correct analysis of contrastive *wa* and a contrast without *wa* must be able to capture this asymmetry.

[13]The thematic *wa* can follow a universally-quantified phrase Han [1998, p. 8].

Many previous analyses are not accurate either. For example, many assume that the presupposition associated with contrastive *wa* is that there is another element in the context in contrast to the one marked with *wa* [e.g., Miyagawa, 1987, p. 190; Shibatani, 1990, p. 265; Han, 1998, p. 2]. But this presupposition is too weak, as can be seen in the following example:

(162) *i.* "Here are a banana and a mango."

ii. Ken-wa **banana**-o/#wa tabe, **mango**-mo tabeta.
Ken-TOP banana-ACC/CONT ate (and) mango-too ate
"Ken ate the banana, and ate the mango too."

The *wa*-marking is infelicitous in this context even though 'mere contrast' requirement is satisfied.

Another group of analyses assumes a presupposition that considers contrasts with and without *wa* basically identically [Teramura, 1991, p. 66; Noda, 1996, p. 7], also in some respect in Choi [1997, p. 549]. Their analyses share the basic idea shown in the following example:

(163) Ken-wa **Peru**-de-wa **banana**-o tabeta.
Ken-TOP Peru-in-CONT banana-ACC ate
"Ken ate bananas in Peru."

Presupposition: "Ken ate something else somewhere else."

In their analysis, the contrast relations between *Peru* and *somewhere else* and between *banana* and *something else* are identical, disregarding the presence of *wa*-marking. One immediate problem with this approach is that it automatically fails to account for the asymmetry in conjunction with the universal quantifier in (161).

There is a relatively old, but impressive work by Huruta [1982]. The analysis is more accurate than most other work including many newer ones. One problem with this analysis is rather ad hoc selections of contrast 'relations' for distinct syntactic types. For example, the individual type, e.g., ken', is contrasted with $\lambda P.\exists Y\,[(Y \neq ken') \wedge P(Y)]$, i.e., a set of properties that holds for someone other than ken', but a property type, e.g., $\lambda X.child'(X)$, is contrasted with $\lambda P.\exists Y\,[\neg child(Y) \wedge P(Y)]$, i.e., a set of properties that holds for some non-child (but not $Y \neq child'$), and so on. He needs to set up a referent and its contrastive relation case-by-case depending on the phrase type. We would prefer a more general relation to capture the notion of contrastiveness.

Let us first discuss the relation between an element X in the utterance and another element X^c

in contrast in the presupposition. I argue that this can be uniformly captured by a relation involving the notion of 'alternatives' in relation to the phonological prominence, following Alternative Semantics [Rooth, 1985; Rooth, 1992; Rooth, 1996]. This generalizes the case of [Huruta, 1982] where distinct relations are used for different phrase types.[14] The presupposition for the two types of contrasts is as follows:

(164) *a*. Contrast without *wa* (**weak**): The presupposition is that there is some distinct X^c (or, something else is involved).

b. Contrast with *wa* (**strong**): The presupposition is that there is some X^c that does not hold in the current situation. X^c is necessarily distinct from X in this case.

We first observe that the presupposition for contrast without *wa* involves conventional (non-cancellable) and conversational (cancellable) implicatures [Grice, 1975; Karttunen and Peters, 1979]. In fact, the following situation seems identical to English.

(165) *a*. Ken-wa **banana**-o tabeta.
Ken-TOP banana-ACC ate
"Ken ate a/the banana."

Presupposition: (i) "Something else is involved." (conventional, non-cancellable)
(ii) "Ken didn't eat something else." (conversational, cancellable)

b. Ken-wa **banana**-o tabenakatta.
Ken-TOP banana-ACC didn't eat
"Ken didn't eat a/the banana."

Presupposition: (i) "Something else is involved." (conventional, non-cancellable)
(ii) "Ken ate something else." (conversational, cancellable)

McGloin [1987, p. 166] observed that the case like (165*b*) is ambiguous between the scope of negation. Here, we consider the same ambiguity in terms of the applicability of the conversational implicature (ii), while (i) is always available with the phonological prominence.

We now examine the case with contrastive *wa*, which is again always accompanied with prominence.

(166) *a*. Ken-wa **banana**-wa tabeta.
Ken-TOP banana-CONT ate

[14] It is also possible to apply Alternative Semantics even to the higher-order contrast between the functions of *wa* or *ga*. Such a case can occur when *wa* or *ga* itself, and not an element in the phrase, receives prominence.

"Ken ate a/the banana."

Presupposition: "Ken didn't eat something else." (conventional)

b. Ken-wa **banana**-wa tabenakatta.
Ken-TOP banana-CONT didn't eat
"Ken didn't eat a/the banana."

Presupposition: "Ken ate something else." (conventional)

The presuppositions have propositional forms identical to the (*ii*) versions of (165). But it is now conventionalized, or grammaticalized. This distinction can be observed in (162). The utterance (162*ii*) cannot be felicitous if the contrast without *wa* has the same presupposition as the case with *wa*. We say this presupposition with *wa* in (166) is **stronger** than that without *wa* in (165*a*). The situation can be summarized as follows:

		Contrastiveness (conventional implicature)
Phrase without *wa*	Non-prominent	None
	Prominent	Weak (possibility of conversationally strong)
Phrase with *wa*	Non-prominent	Not available (as contrastive *wa*)
	Prominent	Strong

Table 5.2: Contrastive Function of *wa*

The following example shows the case where both types of contrasts are involved, as in Teramura's analysis for (163).

(167) *a*. Ken-wa **Peru**-de-wa **banana**-o tabeta.
Ken-TOP Peru-in-CONT banana-ACC ate
"Ken ate bananas in Peru."

Presupposition: (i) "Ken didn't eat bananas somewhere else." (from **Peru**-de-wa)
(ii) "Something other than banana is involved." (from **banana**-o)

b. Ken-wa **Peru**-de-wa **banana**-o tabenakatta.
Ken-TOP Peru-in-CONT banana-ACC didn't eat
"Ken didn't eat bananas in Peru."

Presupposition: (i) "Ken ate bananas somewhere else." (from **Peru**-de-wa)
(ii) "Something other than banana is involved." (from **banana**-o)

The analysis is that both types of presuppositions simply *co-exist*. It is also possible that, for example in (167*a*), there is a conversational implicature such as "Ken ate something else somewhere

else", as in Teramura's analysis for (163). It is not easy to show that such presupposition is only conversational (cancellable). But the following example seems to provide a support for the current position.

(168) *i.* "Ken ate neither bananas nor mangos in Montana."

ii. Ken-wa **Peru**-de-wa **banana**-o tabeta.
Ken-TOP Peru-in-CONT banana-ACC ate
"Ken ate bananas in Peru."

Presupposition: (i) "Ken didn't eat bananas somewhere else." (from **Peru**-de-wa)
(ii) "Something other than banana is involved." (from **banana**-o)

But the strong presupposition "Ken ate something else somewhere else" cannot mean "Ken ate mangos in Montana", which is contradictory, even though the components are available in the previous utterance.

We now show that the above analysis provides a solution to the problems we discussed earlier. First, as soon as we consider phonological prominence and the Alternative Semantics approach, we obtain a solution to the problem of 'association with contrast' (160). Next, let us consider the 'asymmetry' problem repeated below:

(169) Ken-wa **minna**-o/*wa hometa.
Ken-TOP everyone-ACC/CONT praised
"Ken praised everyone (in contrast to just Naomi)."

The basic idea is that the universally-quantified NP is in contrast to various kinds of quantified NPs [Büring, 1997b, p. 40]. The weak contrastiveness associated with prominence without *wa* is easily satisfied because the universally-quantified NP can contrast with virtually anything. On the other hand, the strong contrastiveness associated with the contrastive *wa* can only contrast with *nobody* because any positive set would result in a contradiction, e.g., "not somebody praised" is equivalent to "nobody praised". But, as long as an alternatives set involves some element other than *nobody*, that element must be a positive one and thus the alternatives set is contradictory. Therefore, no alternatives analysis is possible for contrastive *wa* in this case.

Although we did not discuss above, there is an issue in relation to the pragmatic function without *wa*-marking. As we have briefly seen in Subsection 2.3.4, [Kuno, 1973, p. 49 (citing Kuroda)] argues that many instances of *ga* result in exhaustive interpretation. But Shibatani [1990, (14), p. 271] presents the following example, and argues against Kuno that it is epiphenomenal.

(170) *a*. Nani-ga siroi?
what-NOM whilte
"What is white?"

b. Yuki-ga siroi. Sorekara, usagi-mo siroi.
snow-NOM white then rabbit-too white
"Snow is white. And the rabbit is white too."

This is consistent with Vallduví's [1990, Section 7.1] view that exhaustivity is conversational implicature [Grice, 1975]. Thus, it can be separated from the contrastiveness we are discussing.

In summary, contrastive *wa* is always associated with phonological prominence within *wa*-marked the phrase, and has presupposition stronger than just case particles.

5.2.3 Thematic Function

This subsection shows that thematic *wa* (i) is a matrix-level (root) phenomenon associated with long-distance fronting, (ii) does not require prominence, and (iii) signals a contextual link. A contextual link at the matrix level is a key element that give rise to a theme, as we have seen in Chapter 3.

We first confirm Kuno's [1973] argument that thematic *wa* does not appear in embedded environments, and then examine the thematic function at the matrix level.

Distribution of Thematic *WA*

Kuno's [1973, p. 56] argument that no thematic *wa* can appear in an embedded clause seems natural to accept. But there are arguments against this position [Tateishi, 1994; Noda, 1996]. In the following, we first review some arguments in support of Kuno's position, and then rejects Tateishi [1994] and Noda [1996] with respect to this point.

The distribution of thematic *wa*, especially in relation to the nominative case marker *ga*, has been observed well before Kuno [1973]. For example, Shibatani [1990, p. 272] cites Yamada (1908) for the following pair of sentences:

(171) *a*. Tori-ga tobu-toki naku.
bird-NOM fly-when sing/cry
"When a bird flies, someone cries."

b. Tori-wa tobu-toki naku.
bird-TOP fly-when sing/cry
"Birds sing when they fly."

Yamada's point was that depending on the particle, the word *tori* (bird) is interpreted as the subject of the embedded or the matrix clause. Although this is intuitively appealing, we need to be more specific about the syntactic structure and, more importantly, the context. We also need to clarify the definition of embedding.

The subject of the embedded clause:	Occurrences	%
a. Shared with the matrix-level subject	3	2
b. Shared with the matrix-level subject (separated by a comma)	9	7
c. Shared with a matrix-level non-subject (e.g., object)	2	2
d. Dropped (unspecified)	45	36
e. Relativized	30	24
f. *ga*-marked (nominative)	23	18
g. *mo*-marked (*too*)	2	2
h. *wa*-marked (contrastive)	4	3
i. *wa*-marked (non-contrastive)	0	0
j. Inside a direct quote	8	6
Total	126	100

Table 5.3: Subject Marking in Embedded Environments

In order to confirm Kuno's statement, I conducted a small-scale corpus analysis. The data is from "Asahi Newspaper top stories" (on-line version)[15] on Mar. 2, 1999. In the data, there are 137 sentences with 129 occurrences of *wa* and 74 occurrences of *ga*. First, the following types of embedded clauses are collected: (i) relative clause, (ii) complement clause, and (iii) subordinate clause.[16] There are 126 such occurrences. The distribution of subject marking in these embedded clauses is shown in Table 5.3. In summary, the only obvious occurrences of *wa* in an embedded environment are those in the category *h*, i.e., contrastive *wa*.

Since we are concerned with the semantic property of contrastiveness, let us consider the English translation (mine) for the four occurrences of *wa*-marking in the category *h*. The first example is as follows:

[15]The web site is "`http://www.asahi.com/paper/front.html`". The data is available through "`http://www.cis.upenn.edu/~komagata/thesis.html`".

[16]There is a case whose status is not very clear between subordinate or coordinate structures, are excluded from the count. This involves a clause linking particle *te* at the end of the first clause (see Hasegawa [1996] for our analysis).

(172) *i.* (description of a tight financial situation about a Japanese company)

ii. Since *the temporary money for this summer* will be drawn from this year's budget, they are planning to reduce the $1.7billion-administrative costs through no raise and wage cut.

The phase "*the temporary money for this summer*" can be considered to be in contrast with the fixed budget. The remaining three examples are found in another text shown below.

(173) *i.* (description of a young person who stopped breathing after drowning)

ii. They judged that *the hope of resuscitation* is completely out.

iii. (a few more utterances following the above)

iv. The physician in charge, Dr. Wada, said that *the parents* agreed but *the siblings* objected.[17]

"*the hope of resuscitation*" contrasts with the situation the young person is dying, and *the parents* and *the siblings* are explicitly contrasted.

While these three are the only clearly embedded instances of *wa*, we should briefly comment on the categories *a.* and *b*, also related to the example (171). The following is a simplified example of the category *b.*

(174) Sentaa$_i$-wa, [$\varnothing_i$ kamoku-o kimeru] -to mirareru.
center-TOP subject-ACC decide -COMP expected
"The center is expected that [it] decides on the subjects."

The comma after *sentaa* (*center*) indicates that it is the subject of the matrix clause. The subject of the embedded clause (shown as $\varnothing_i$) is dropped and coincides with the matrix-level subject. Thus, it is safe to say that the *wa*-marking is for the matrix clause and not for the embedded clause.

The following is a slightly simplified example of the category *a*.

(175) Seifu-wa kihon rinen-ni sot-te kihon keikaku-o sadameru.
government-TOP basic principle-DAT follow-as basic plan-ACC fix
"The government fixes the basic plan as it follows the basic principles."

This case is formally distinct from the category *b.* due to the absence of a comma. The question here is whether *seifu-wa* (*government*) is the subject of the matrix clause or that of the embedded

[17] Only one subject per embedding has been counted.

clause. For the above case, we can move the matrix-level object before the embedded clause as follows:

(176)	Seifu-wa	kihon	keikaku$_i$-o	kihon	rinen-ni	sot-te	t_i	sadameru.
	government-TOP	basic	plan-ACC	basic	principle-DAT	follow-as		fix
	same translation							

Since the matrix-level object cannot presumably enter into the embedded clause, *seifu-wa* (*government*) in the above case can be considered to be at the matrix level. Although this does not show that the utterance (175) must have the same structure, it still supports the possibility. In addition, it is more natural to place a pause after *seifu-wa* (*government*) when it is read aloud. Therefore, the data do not contain counterexamples to Kuno's statement that thematic *wa* does not occur in embedded environment.

Some theoretical analyses are also in support of Kuno's statement. Han [1998] applies the 'mapping hypothesis' of Diesing [1992] to Korean counterpart *(n)un*.[18] Han's [1998, p. 1] argument is that 'topic' reading, corresponding to a type of presupposition, is available only at a VP-external position (with or without contrast) as a result of quantificational force associated with the position, and VP-internal position is limited to contrastive focus. Kawashima [1989, p. 64] supports Kuno's statement from the point of view that a *wa*-marked phrase always scopes over both matrix-level and embedded clauses.

Let us now turn to the arguments that thematic *wa* can appear within an embedded clause. First, Tateishi [1994, p. 153] argues that thematic *wa* (his 'topic') can be embedded arbitrarily deep. He uses "*ano hon*" (*that book*) and explicitly provides a context where the book is anaphoric. The problem here is that anaphoricity is not sufficient for themehood. He misses this point because very little attention is paid to contrastive *wa*. All of his embedded *wa* are felicitous if pronounced with prominence and in a context where the book is contrasted with something else. They do not stand as counterexamples to Kuno's hypothesis.

Noda [1996, p. 171] argues that thematic *wa* can appear in parallel clause, 'weak' reason clause, and quotation. First, Noda's parallel clause [p. 176] are coordinate structure, and should be excluded from what we call embedding. His 'weak' reason clause is non-rhematic subordinate

[18]The mapping hypothesis says that the material from IP and the material from VP correspond to the restrictive clause and the nuclear scope of the tripartite quantification structure, respectively, as in the following example: $\forall X \left[\underset{\text{restrictive}}{man(X)} \Rightarrow \underset{\text{nuclear}}{die(X)} \right]$.

clause. A few examples of this type actually contains contrastive *wa* [p. 177]. Noda's [1996, p. 179] example of quotation is a *direct* quotation, which can be shown by the use of pronoun. We focus on expository texts, and exclude direct quotes from analysis.

We thus conclude that thematic *wa* cannot appear in embedded environment. The subject of a complement clause can be fronted relatively easily. But this is structurally different from the cases we have been looking at. Before investigating the function of fronting, let us next turn to the thematic function of *wa*.

Thematic Function at the Matrix Level

Now, we know that thematic *wa* is limited to the matrix or fronted position. In this section, we confirm the following two points: (A) instances of thematic *wa* are a part of a theme and (B) any *wa*-marked phrase is a contextual link (either thematic or contrastive). For the following discussion, let us assume that the matrix elements are vacuously fronted. Thus, when we say 'matrix level', that includes fronted cases as well.

Instances of *wa* at this position can be thematic (non-contrastive), as in the example (157) or thematic and contrastive, as in the example (159), or rhematic *and* contrastive, as in the example (158). This situation is shown in Table 5.4.

Prominence/Contrastiveness	Information structure
Prominent/Contrastive	Rhematic
	Thematic
Non-prominent/Non-contrastive	

Table 5.4: Contrastiveness and Information Structure for *wa* at the Matrix Level

Thus, the distinction between thematic and rhematic is not phonological. As long as the main hypothesis of information structure (48) are satisfied, either choice is possible. On the other hand, we can weakly relate prominence and information structure. Non-prominent *wa*-marked phrase, available only at the matrix/fronted position is thematic. Thus, this is the only case we can identify a theme based on the *wa*-marking.

Non-prominent matrix-level *wa* is 'thematic' for the following reasons. First, it cannot be used to respond to a *wh*-question.

(177) *Q*: "Who ate the banana?"

A: # Ken-wa banana-o tabeta.
Ken-TOP banana-ACC ate
"Ken ate the banana."

Second, when the context is sufficiently restricted, it can be dropped. This is not possible for a rheme.

(178) *Q*: "What did Ken eat?"

A: ∅ **banana**-o tabeta.
banana-ACC ate
"(he) ate the banana."

An instance of contrastive, thematic *wa* cannot be dropped for the contrastive reason.

While thematic *wa* is necessarily a contextual link, it is not a contextual-link marker. Because if it were, it should be able to appear in an embedded environment due to the hypothesis (30). Thematic *wa* is not for the absolute notion of referential status but for the relative notion in contrast to a rheme. Although Hinds [1987, p. 87] attempts to characterize the choice between *wa* and *ga* based on Prince's [1981] taxonomy, his argument cannot be correct. For example, he cannot explain the case where an EVOKED referent can be *ga*-marked when it is a rheme.

The special status of thematic *wa* seems to be a result of multiple factors. Originating with the contrastive function, thematic *wa* may have evolved as it loses prominence.[19] This development is possible only at the matrix level. There, loss of prominence is coupled with contextual link status. According to our theory, a contextual link is the only source of a theme. Such a development could not make sense in an embedded environment because no information-structure division is possible within an embedded clause (except for extracted constituents, which we consider 'matrix level').

The distinction between *wa* and *ga* and other case particles in an embedded environment is that of degree of contrast between strong, weak, and none, i.e., absolute semantic status in relation to referents in the context as shown in Table 5.5.

	Prominent	Non-prominent
wa	Strong contrastive	N/A
ga and other case particles	Weak contrastive	Non-contrastive

Table 5.5: *wa* vs. *ga* at Embedded Environments

[19] Historic development was briefly mentioned on page 127.

At the matrix level, the focus is placed more on the relation between distinct constituents (Table 5.6).

wa	Prominent	Non-prominent
Embedded	Theme/Rheme depending on the clause	n/a
Matrix/Fronted	Theme/Rheme	Theme

Table 5.6: *wa* vs. *ga* at the Matrix Level

So far we have noted the connection between thematic *wa* and contextual link. But is *wa* inherently contextual link including non-thematic ones? Many researchers have argued in this position as follows. Although described in different ways, they all share the basic idea, e.g., *wa* is used for 'known' [Yoshimoto, 1992, p. 2]; *wa* is 'identifiable' [Iwasaki, 1987, p. 108]; *wa* is 'set anaphoric' [Miyagawa, 1987, p. 190]; the Korean counterpart *(n)un* presupposes a 'non-empty set' [Han, 1998, p. 5].

Some borderline cases have been reported in Hinds [1987, p. 87]. These involve use of *wa* for UNUSED and anchored BRAND-NEW referents (in the sense of Prince [1981]). Anchored BRAND-NEW referent is a type of BRAND-NEW referent with some linguistic link called 'anchor' (see Table 2.1). An UNUSED referent is inferrable from the context in a wider sense. If anchored BRAND-NEW can be marked with *wa* as Hinds says, that is potentially an evidence for non-contextual-link use of *wa* (presumably contrastive). But his argument is weak because no examples are shown. For the moment, let us consider that all the instance of *wa* regardless of thematic or contrastive is a contextual link.

A conjecture here is that the contextual-link status of contrastive *wa* is not an extension of that of thematic *wa*, but that the strong contrastiveness requires the contextual-link status. Let us recall the strong presupposition: "there is something else which can fail the proposition". For this presupposition to hold, the speaker and the listener must know 'something else' (even though one of them do not know the referent of the *wa*-marked phrase), and it is likely that the referent of the *wa*-marked phrase can be inferred from this 'something else'.

There is one other point introduced by Kuno. That is, thematic *wa* is either anaphoric or generic as follows Kuno [1973, (17), p. 44]:

(179) *a*. John-wa watakusi-no tomodati desu. (anaphoric)
John-TOP my friend COP

"John is my friend."

b. Kuzira-wa honyuu-doobutu desu. (generic)
whale-TOP mammal COP
"A whale is a mammal."

While we cannot go into the issue of 'genericity' in detail, this is a separate aspect. Since we consider discourse referent of arbitrary semantic types, a generic referent can be EVOKED (anaphoric) or INFERRABLE (not anaphoric).

Summary

We have started with the contrastive function of *wa* as the basic function, and argued that its strong contrastiveness is associated with phonological prominence. This semantic/pragmatic function is available basically everywhere, distinguished from the non-contrastive and weak contrastiveness (prominence without *wa*) cases. Particle *wa* always signals contextual link through the thematic function or the strong contrastive function.

The thematic function of *wa* is a result of long-distance fronting to a matrix position. The function can co-exist with contrastiveness, but the interesting part is the non-contrastive/non-prominent use, which cannot appear in embedded clauses where no information-structure partition is possible.

5.3 Function of Long-Distance Fronting

It has been proposed that long-distance fronting makes *wa* thematic [Choi, 1997, p. 548 (for Korean)]. But we must explore this statement more thoroughly. Long-distance fronting is necessary for thematic *wa*, but it is not sufficient. Contrastive *wa* can stand at a fronted position without thematic function. In this section, we explore the idea that long-distance fronting is a general-purpose re-ordering device.

In Japanese, two types of 'movement' have been observed: local scrambling and long-distance fronting [e.g., Miyagawa, 1997].[20] Local scrambling is a movement within a clause, as seen in the following example:[21]

[20]This distinction may not be *necessary*. In the end, a single theory might be able to account for both cases.

[21]The following grammatical labels are used: TOP = topic, CONT = contrastive, NOM = nominative, ACC = accusative, DAT = dative, COMP = complementizer, COP = copula, and Q = question.

(180) Local scrambling:

a. [Ken-ga Naomi-ni ageta] mono-wa banana-da. (canonical)
Ken-NOM Naomi-DAT gave thing-TOP banana-COP
"The thing which Ken gave to Naomi was banana."

b. [Naomi-ni Ken-ga ageta] mono-wa banana-da. (scrambled)
Naomi-DAT Ken-NOM gave thing-TOP banana-COP
"The thing which Ken gave to Naomi was banana."

A relative clause is used to avoid the involvement of long-distance fronting.

Next, the following is an example of long-distance fronting.[22] Phonological prominence is placed to make the sentences more natural.

(181) Long-distance fronting:

a. **Naomi**-ga [Erika-ga banana-o tabeta] -to omotta. (canonical)
Naomi-NOM Erika-NOM banana-ACC ate -COMP thought
"Naomi thought Erika ate the banana."

b. Banana$_i$-wa **Naomi**-ga [**Erika**-ga t_i tabeta] -to omotta. (fronted)
banana-TOP Naomi-NOM Erika-NOM ate -COMP thought
"The banana, Naomi thought Erika ate."

Long-distance fronting is 'unbounded' in the sense that the fronting can originate in an arbitrarily deeply embedded clause (modulo processing limitation, as usual).

A few remarks on previous work are in order. Kiss [1981] argues that Japanese has a fixed information structure with the "*Topic – Focus – Background*" pattern. But we have seen that is not the only case. Miyagawa [1997] suggests that long-distance fronting is related to information structure but does not go beyond that point. Gunji [1987, Section 5.2, p. 219-220] distinguishes two type of topicalization (argument and non-argument cases) and emphatic fronting. But it is not clear whether the syntactic operation involved in topicalization (argument case) and fronting are really distinct.

Long-distance fronting is most commonly observed at the matrix level, and at this level, setting up information structure is a typical function. The following examples show such a case.

(182) *Q*: "Who thought who ate a/the banana?"

[22] Long-distance fronting is also called as long-distance scrambling. I will use (long-distance) fronting to easily distinguish from (local) scrambling.

A: Banana$_i$-wa **Naomi**-ga [**Erika**-ga t_i tabeta] -to omotta.
Banana-TOP Naomi-NOM Erika-NOM ate -COMP thought
"Naomi thought that Erika ate the banana."

Here, *banana*, the theme, is fronted from an embedded position to be contrasted with the two more informative *ga*-marked NPs.[23]

(183) *Q*: "What did Naomi thought Erika ate?"

A: **Banana**$_i$-o Naomi-wa [Erika-ga t_i tabeta] -to omotta.
Banana-ACC Naomi-TOP Erika-NOM ate -COMP thought
"Naomi thought that Erika ate the banana."

In this case, *banana* is the rheme and is again fronted to separate the rest of the utterance as the theme. In (183*A*), the *wa*-marking of *Naomi* is not clear whether we can say that it is a result of long-distance fronting (vacuous) or that it is in situ at the matrix clause.

But long-distance fronting is not limited to the matrix level.

(184) *a*. (in a situation where Naomi told multiple people that Erika ate either mango or banana)

b. **Banana**$_i$-o Naomi-ga [Erika-ga t_i tabeta] -to tutaeta hito
Banana-ACC Naomi-NOM Erika-NOM ate -COMP tole person
"the person whom Naomi told that Erika ate the banana"

Extraction from a relative clause is not impossible in Japanese,[24] but is strongly resisted. The above example shows that *banana* is the key element in the contrast among people and that long-distance fronting is not necessarily a matrix phenomenon. Thus, not every case of long-distance fronting licenses thematic *wa* either (but thematic *wa* cannot be found in a position where long-distance fronting is not applicable, e.g., embedded position). Since I have argued that direct information-structure marking must be a matrix phenomenon (Subsection 2.3.3), long-distance fronting cannot be so, much like the cleft construction in English.

In Japanese, discontiguous information structure of the pattern "*Theme – Rheme – Theme*" is fairly common. This reflects the tendency to front thematic materials and verb (even when it is a part of the theme) remains in situ due to strict verb-final property, as shown in (96) repeated below.

[23]In this case, the embedded and matrix verbs, which are also parts of the theme, are left in the original position. The consequence is a discontiguous information structure of "*Theme – Rheme – Theme*". We suspect that the strict verb-final property is the cause of this discontiguity.

[24]See Example (1) on p. 211.

(185) *Q*: Ken-wa nani-o tabeta-no?
Ken-TOP what-ACC ate-Q
"What did Ken eat?"

A: [Ken-wa]$_{Theme}$ [**banana**-o]$_{Rheme}$ [tabeta]$_{Theme}$.
Ken-TOP banana-ACC ate
"Ken ate a banana."[25]

This corresponds to the idea that pre-verbal position is a 'focus position' (a comparable idea in Hoffman [1995, Section 5.4.1]). But we cannot associate a pre-verbal position with a rheme, as we have already seen, e.g., (158*A*, 181*a*, 183*A*, 184*A*).

Long-distance fronting that is still bounded within an embedded clause actually has commonality with local scrambling. Although we leave it for future research, local scrambling and long-distance fronting may be more similar than previously thought. Information-structure-related function of long-distance fronting is in fact a combination of contextual link and semantic composition at the matrix level.

Long-distance fronting is a general-purpose constituent re-ordering device. At an embedded level, it does not separate information structure, but it can separate a contrastive element from the background elements. At the matrix level, it can still separate a contrastive element, but can also separate materials to set up information structure.

With respect to its functions, fronting in Japanese is similar to cleft in English (see Subsection 3.3.2). Both of these can appear at an embedded level, and re-order some elements for various pragmatic reasons. At the matrix level, fronting in Japanese functions in a way similar to the combination of topicalization and focus movement in English. They weakly mark information structure as re-ordering can affect the way semantic composition is done at the last stage of derivation.

5.4 Prediction of *wa* and *ga* from Information Structure

In this section, we combine the discussion up to this point and analyze the distinction between *wa* and case particles including *ga*. The complicated situation involving all these can now be seen in terms of the theory behind it. We then present a method to predict *wa* and *ga* from information structure and grammatical information.

[25] Depending on the situation, the definite article *the* may also be applicable.

Resulting Effects

The summary of the propositions we support are as follows:

(186) *a*. Phonological prominence is associated with 'contrast'.

b. The degree of contrast is distinct for the case with and without *wa*. We called the contrast involving *wa* 'strong'.

c. Long-distance fronting is a general constituent re-ordering mechanism possibly involving contrastiveness, contextual-link status, and information structure.

d. The thematic function of *wa* can appear without prominence only at the matrix level.

From these and some additional points discussed below, we can infer the resulting pattern of *wa* and case particles including *ga*.

In embedded environments, (186*a*, *b*) are sufficient to derive the results in Table 5.7. It is a three-way distinction with respect to contrastiveness between (i) case particle without prominence, (ii) case particle with prominence, and (iii) *wa* with prominence. An embedded clause cannot have an information-structure division within itself (except for constituents fronted into the matrix level). Thus, there is no information-structure marking. A conjecture is that local and long-distance fronting within an embedded clause marks contrastiveness.

Embedded case		Information structure	Contrastiveness
wa (TOP/CONT)	**Prominent**	Unspecified	Strong
	Non-prominent	Not available	
ga, o, ni (NOM, ACC, DAT)	**Prominent**	Unspecified	Weak
	Non-prominent	Unspecified	None

Table 5.7: *wa* and Case Particles in Embedded Environments

The situation is substantially more complicated at the matrix level. Now, let us compare *wa* with *ga*. First, matrix-level *ga*-marking with prominence is rhematic. It cannot be a theme, even a contrastive theme, as in the following example.

(187) *Q*: "What did Ken and Naomi eat?"

A: # **Ken**-ga **banana**-o tabeta.
Ken-NOM banana-ACC ate
"Ken ate a/the banana."

But *ga*-marking can appear without prominence at the matrix level. I take it that this type of *ga* corresponds to Kuno's [1973] neutral description assuming that his exhaustive listing requires prominence. Kuno [1973, p. 51] states that neutral description presents a "temporary state as a new event". More recent analyses found that this type of utterance is available with a 'stage-level' predicate (the definition later) [Shirai, 1986, p. 65; Heycock, 1994, p. 159] and that it is considered all-rheme [Choi, 1997, p. 546]. This situation contrasts with thematic *wa*, which can also be non-prominent. Therefore, regardless of prominence, *ga*-marked NP at the matrix-level is (a part of) the rheme. The contrast between *wa* and *ga* at the matrix level is summarized in Table 5.8. Note that non-prominent *ga* cannot be fronted from an embedded level. If fronting is for thematic purpose, it must be marked with a *wa*. Furthermore, we follow Heycock [1994, p. 161] and do not consider *ga* as a rheme marker. In embedded environments, *ga* may appear as a part of either theme or rheme. What we have seen above only shows that *ga* at the matrix level cannot be a theme.

Matrix case		Information structure	Contrastiveness
wa (TOP/CONT)	**Prominent**	Theme/Rheme	Strong
	Non-prominent	Theme	None
ga (NOM)	**Prominent**	Rheme	Weak
	Non-prominent	Rheme	None

Table 5.8: *wa* vs. *ga* at the Matrix Level

Second, let us consider other case particles, i.e., accusative case particle *o* and dative case particle *ni*. These case particles behave similarly to the case particle *ga*, but there is a difference. The difference seems to come from a grammatical constraint that multiple occurrences of thematic *wa* are not allowed [Kuno, 1973, p. 48]. Thus, if the subject is already marked with a thematic *wa*, other arguments stay with their case particles. The reason *o*/*ni* cannot compete with *ga* for a thematic *wa* is probably due to the fact that the subject tends to be the theme and thematic *wa* is statistically strongly associated with subject. Thus, non-prominent *o*/*ni*-marking may be either theme or rheme. The resulting situation is shown in Table 5.9. The above argument shows that a relatively small number of conditions (186) can account for the phenomenon at the matrix and an embedded levels.

Finally, let us briefly comment on the case of adverbials. As before, *wa*-marking on an adverbial with prominence is strongly contrastive. If a *wa*-marked adverbial is fronted and loses

Matrix case		Information structure	Contrastiveness
wa	**Prominent**	Theme/Rheme	Strong
(TOP/CONT)	Non-prominent	Theme	None
ga	**Prominent**	Rheme	Weak
(NOM)	Non-prominent	Rheme	None
o, *ni*	**Prominent**	Rheme	Weak
(ACC, DAT)	Non-prominent	Theme/Rheme	None

Table 5.9: *wa* and Case Particles at the Matrix Level

prominence, it is thematic. If *wa*-marking on an adverbial is the only *wa*-marking and the matrix subject is *ga*-marked, we expect that the adverbial is a part of the theme and the subject is a part of the rheme.

Particle Choice

Now, Table 5.9 can be used as our tool for choosing a particle at the matrix level. But, when we deal with written texts, prominence information is not available. Therefore, in theory, we cannot identify a theme in the way we have been discussing. But lack of various phonological properties can actually bring in other factors to compensate. In order to represent prominence in writing, one would use special construction, punctuation, etc. As a consequence, many instances of *wa*-marking at the matrix/fronted position are in fact thematic. The same Asahi Newspaper data (see p. 5.2.3) has 110 occurrences of matrix-level *wa*. Among them, 100 occurrences (91%) are thematic and 10 occurrences (9%) are contrastive *wa*. But none of the contrastive cases appears to be a rheme observing that the predicates for these cases are non-contextual links. Since Japanese allows dropping constituents freely, if the verb arguments are perfectly clear, they can be dropped. But, in written texts with a complex propositional structure, theme may not be that obvious. For this purpose, thematic *wa* can be effectively used.

Theoretically, we could still analyze texts with respect to contrastiveness and separate the instances of contrastive *wa*. But, computationally, general analysis of contrastiveness is still very difficult (see Prevost [1995] for a theory and implementation for a small domain). One way to tackle this situation is to analyze certain syntactic environments where contrastiveness is strongly associated, e.g., parallel contrastive structure and negative environment. We discuss these structures in the following.

One environment where contrastive *wa* is routinely used is in an explicit parallel clause [Noda, 1996, p. 200], as in the following example (see Subsection 5.2.2).

(188) Ken-wa Americazin de, Naomi-wa Canadazin da.
Ken-TOP/CONT American COP (and) Naomi-TOP/CONT Canadian COP
"Ken is an American and Naomi is a Canadian."

Note that the above case involves contrast between the two grammatical subjects. Other components of a clause can also be contrasted in an analogous way.

Another environment where contrastive *wa* is commonly used is the negative environment [Noda, 1996, p. 214].

(189) Ken-wa kuurudenai.
Ken-TOP/CONT cool.not
"Ken is not cool (but someone else is)."

Use of contrastive *wa* is usually justified by the assumption that the presupposition associated with strong contrast is available for negation due to the availability of positive knowledge [Russell, 1948, Chapter IX]. For the negative environment, the scope of negation and the use of contrastive *wa* must correspond. In the above example, the verb *kuuruda* (*cool*) predicates over the subject *Ken*. Thus, it is clear that the negation scopes over the *wa* attached to *Ken*. When the predicate has more than one argument, the situation is more complicated, as shown below.

(190) Ken-wa banana-o tabeai.
Ken-TOP/CONT banana-ACC eat.not
"Ken doesn't eat a/the banana."

Presupposition: (i) "Something other than banana is involved."
(negation scopes over banana)
(ii) "Ken eats something else." (negation scopes over Ken)

In general, we cannot tell whether the *wa*-marking on the subject is under the scope of the negation without phonological prominence.

There is a case where the use of *ga* can be predicted, which involves the distinction between stage and individual-level predicates. Informally, **stage-level predicates** are associated with temporary state while **individual-level predicates** are associated with permanent state [Carlson, 1980]. Two examples are shown below:

(191) *a*. Ken-wa kuuruda. (individual-level predicate)
Ken-TOP cool
"Ken is cool."

b. Ken-wa sinda. (stage-level predicate)
Ken-CONT died
"Ken died."

For (*a*), "*Theme – Rheme*" information structure is commonly observed. But, for (*b*), "*Theme – Rheme*" information structure is rare (all rheme with *ga*-marking is more common). A possible analysis for this situation is that the utterance (*b*) requires a specific 'situation' where the proposition must be interpreted. For the "*Theme – Rheme*" structure, this 'situation' and *Ken* must be jointly contextually-linked while *sinda* (*died*) is the rheme. But such a case seems to require elaborate set up not commonly observed in expository texts.

I suspect that the interaction between stage/individual-level predicates and information structure is not specific to Japanese. The conjecture is that the distribution of particles in Japanese and focus projection in English [Diesing, 1992, p. 46] can be explained by the same underlying theory based on the stage/individual-level distinction and information structure. This direction is left for future work.[26]

For our task of evaluating the identified information structures in English, we must be able to predict particle choice, which can be compared against human translation. Fig. 5.1 presents an example of applying the above analysis to a particle-choice procedure for grammatical subjects.

The procedure seems relatively straightforward for humans. But several steps, especially involving analysis of contrastiveness, are quite difficult for the computer. In Chapter 6, we implement only the case of *wa*/*ga* prediction based on theme/rheme distinction for the matrix subject.

Particle choice for non-subjects is slightly different. The situation for the embedded environment is identical to the case of subject. Strong contrastiveness invites *wa*, otherwise a case particle is used. At the matrix level, if the subject is not *wa*-marked, *wa*-marking of a non-subject is probably thematic, but, otherwise, it is likely to be contrastive. Since the subject tends to be a theme,

[26] My conjecture is that both particle distribution in Japanese and focus projection in English can be derived from the following two propositions:

(1) *a*. A stage-level predicate has an event argument while an individual-level predicate does not [Kratzer, 1995, p. 126].

b. Every utterance has a theme.

In this thesis, we have been assuming that all-rheme utterances are possible following [Vallduví, 1990] and [Choi, 1997].

Embedded case:	Predict:
• If strong contrastiveness is required,	*wa*
• Otherwise,	*ga*
Matrix case:	
• For a parallel clause (subject contrast),	*wa*
• For a negative construction (one-place predicate),	*wa*
• For other contrastive case,	*wa*
• For a one-place stage-level predicate,	*ga*
• Otherwise,	
• For a theme,	*wa*
• For a rheme,	*ga*

Figure 5.1: Particle Prediction in Japanese

the chance of a non-subject being marked with a *wa* is relatively low. This makes it more difficult in practice to use it as an evaluation tool for checking the information status on non-subjects.

5.5 Summary

We now have a reasonably precise and accurate idea about direct information-structure marking in Japanese, especially in relation to the use of *wa*, case particles, and long-distance fronting. With semantics and information structure, we can predict the use of wa and case particle. The results are used as a particle choice prediction procedure in the next chapter.

Chapter 6

Implementation of the Information-Structure Analyzer

In this chapter we demonstrate that the formalized theory can be implemented for practical applications and evaluation. In particular, we show that (1) the backbone of the system, CCG parser, is practical despite some previously-addressed concerns about spurious ambiguity and (2) the specifications of contextual link and information structure are implementable with some additional procedural aspects, which are modularly upgradable.

The chapter starts with an introduction of the overall architecture in the first section. The following two sections focus on the CCG parser and information-structure analyzer. In the latter section, we also discuss an implementation of particle prediction in Japanese based on the analysis in Chapter 5.

6.1 Introduction

The current system accepts text as input, analyzes its information structure, and predicts particle choice in Japanese as shown in Fig. 6.1. It has two main modules: the parser and the information-structure analyzer. Since our grammar, CCG, can recognize non-traditional constituency in accordance with divisions of information structure, analysis of information structure can proceed in

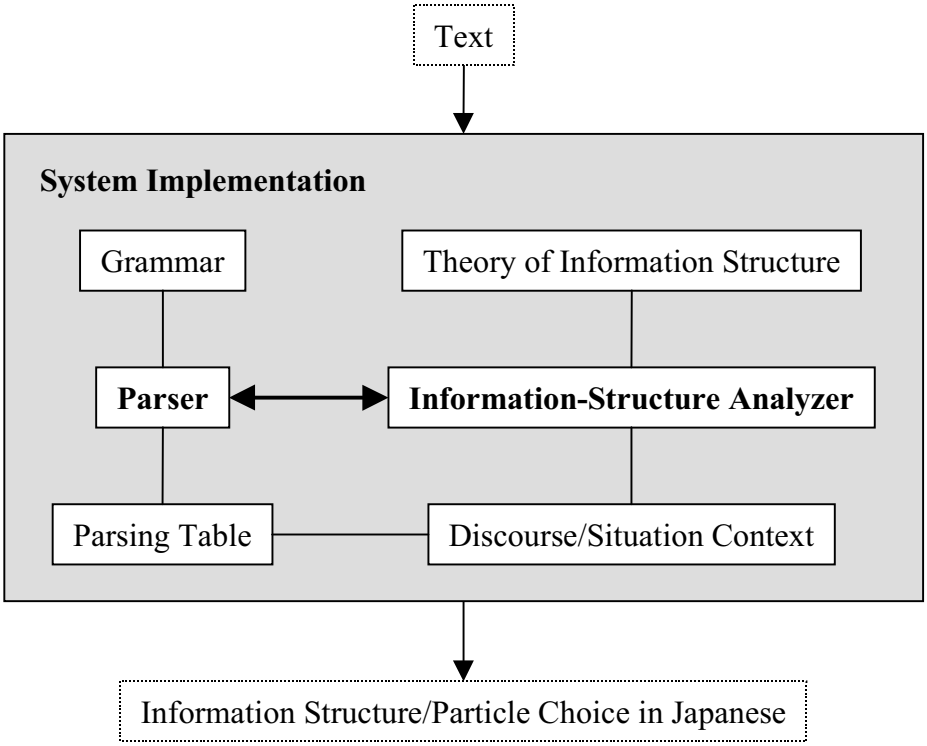


Figure 6.1: System Architecture

parallel to parsing.[1] This situation is represented by the bidirectional arrow '↔' between the parser and the information-structure analyzer in the figure. Also in the system, the parsing table is used to derive the results in an efficient way, avoiding redundancy. The information obtained through parsing is stored as a part of the context, and later used for identifying discourse status. The parser has evolved from an earlier implementation [Komagata, 1997a (for Japanese)]. Another previous implementation [Komagata, 1998a (for English)] included a module for identifying information structure with limited analysis of linguistic marking and no structured-meaning component.

The system is implemented on a Sun Ultra E4000 2×250MHz Ultrasparcs with 320MB memory running SunOS 5.5.1. The code is written entirely in Sicstus Prolog Ver. 3. The program source files are approximately 100KB in size and the data/grammar files are about 200KB (including both training and test data and also lexicons).[2]

[1]This also makes it possible to control parsing, e.g., disambiguation, by the result of discourse processing. This possibility is left for future work.

[2]The source code and data files are available through the author's thesis web page at

6.2 Practical CCG Parser

The practicality of our CCG parser depends primarily on the elimination of spurious ambiguity (i.e., multiple derivation of semantically-equivalent categories as introduced in Subsection 4.1.5) and some other engineering solutions such as preprocessing and the use of features.

We start this section with requirements for the parser. Then, we discuss the elimination of spurious ambiguity, processing of linguistic specifications, and the performance of the parser.

6.2.1 Requirements for the Parser

In order to process information structure as described in the previous chapters, we need a parser to derive semantic representations (to be precise, structured meanings) from input strings. In order to deal with this process, the parser needs to satisfy the following requirements:

1. Capable of processing referents (in our case semantic representations) across utterance boundaries for discourse-status analysis
2. Capable of parsing the complexity of real data, involving the following:
   (a) Spurious ambiguity
   (b) Genuine ambiguities (e.g., modification and coordination)
   (c) Factors beyond 'toy' grammars: including inflection, punctuation, and lexical specification
3. Scalable to larger data (no pre-set limitation associated with the initial data and scale)
4. Applicable to multiple languages (at least English and Japanese)
5. Efficient enough for interactive use (response in the order of *seconds*)

Some of these, but not all, have been addressed in previous work with respect to CCG and similar formalisms. The CCG parsers have been built for several languages: English [Wittenburg, 1986; Komagata, 1998a], Turkish [Hoffman, 1995], and Japanese [Whitelock, 1988 (focus on morphology); Komagata, 1997a]. Applicability to fairly large data has also been shown by Wittenburg

"`http://www.cis.upenn.edu/~komagata/thesis.html`".

[1986]. Application to long, complex sentences is shown to be practical [Komagata, 1997a] with a CKY-style parsing algorithm from [Aho and Ullman, 1972], cf. use of shift-reduce algorithm [Prevost, 1995; Hoffman, 1995]. Before proceeding, we need to distinguish **parsing** and **recognition**: the former derives semantic representation of a parse, and the latter only decides on grammaticality.

Since the problem with spurious ambiguity for practical parsing is only recently addressed, we include the discussion from Komagata [1997a] in the next subsection. The other issues are discussed in Subsection 6.2.3.

6.2.2 Elimination of Spurious Ambiguity

Let us first define several types of ambiguities involved in the parsing process:

(193) *a*. **Categorial ambiguity**: Availability of multiple categories (lexical/derivational), e.g., noun-verb ambiguity for *rose*

b. **Spurious ambiguity**: Multiple derivations of *semantically-equivalent* categories, e.g., "John visited Bill." has two derivations (left and right branching) in CCG with the identical semantic representation "$visited'\,(bill')\,(john')$"

c. **Genuine ambiguity**:

(i) Lexico-semantic ambiguity: Multiple semantic assignments to a single lexical category, e.g., financial *bank* vs. river *bank*

(ii) Attachment ambiguity: Multiple derivations of the same category with *distinct* semantics, e.g., PP attachment

Since spurious ambiguity is unnecessary and can result in an exponential explosion (see Section A.3), CCG parsers must implement some means of eliminating this type of ambiguity. We review three classes of approaches: (i) syntactic, (ii) semantic, and (iii) those which do not belong to the previous two.

First, syntactic approaches eliminate 'spurious derivations', which are not 'the normal form'. Each proposal defines its own 'normal form', but a simplistic example is to choose, e.g., left-branching as the normal form. Then, if there are multiple derivations, only the left-branching is chosen. This does not necessarily suffer from incompleteness because if the left-branching is

unavailable, the right branching can be chosen without conflict. The syntactic approach blends naturally with a theoretical polynomial parsing algorithm for CCG [Vijay-Shanker and Weir, 1990]. Vijay-Shanker and Weir also include a mechanism of eliminating spurious ambiguity during a stage after recognition. Among several proposals, Hendriks [1993] and König [1994] work on Lambek calculus. But Lambek calculus does not include functional composition as a primitive rule. Thus, their proposal does not immediately apply to CCG. Hepple and Morrill [1989] cover a subset of the current formalism but do not have crossing instances of function composition nor type raising. Eisner [1996] covers an even wider range of CCGs but the case including type raising remains to be shown correct. By definition, the syntactic approach does not take semantics into consideration. But our definition of spurious ambiguity refers to semantics. Therefore, normal form parsing does not necessarily match our definition of spurious ambiguity elimination. There is an approach called labeled deduction, which includes semantics within syntactic types [Morrill, 1994]. But the above-mentioned syntactic approaches are not automatically applicable to labeled deduction.

Karttunen [1986] proposes the following semantic method. A new derivation is discarded if its semantic representation is *equivalent to* (or mutually subsumes) that of some entry with the same category already derived and stored.[3] This directly enforces the definition of spurious ambiguity and does not depend on the syntax. Note that 'equivalence' depends on the form of semantic representation [for general discussion Thompson, 1991]. For the case where the semantics is represented in λ-calculus, equivalence is not generally computable [Paulson, 1991]. For the case of feature structure, equivalence is defined as alphabetic variants and characterized by the isomorphism between the structures [Carpenter, 1992]. Our case corresponds to the latter.

Pareschi and Steedman [1987] present a method that belongs to the third type. The approach integrates Karttunen's equivalence check in a CKY-style parsing algorithm, but invokes the mechanism for certain cases of category combination (i.e., a syntactic component). But the published algorithm is shown to be incomplete [Hepple, 1987]. Another approach by Wittenburg and Wall [1991] compiles the grammar so that only normal form derivation is possible. But this compilation replaces the original functional composition schemata with a 'predictive' version of composition schemata. As a consequence, certain non-traditional constituents such as subject-verb sequence

[3]For a detailed discussion of subsumption, see Shieber [1986].

that depend on the original functional composition are no longer available for coordination. Thus, a crucial property of CCG is compromised.

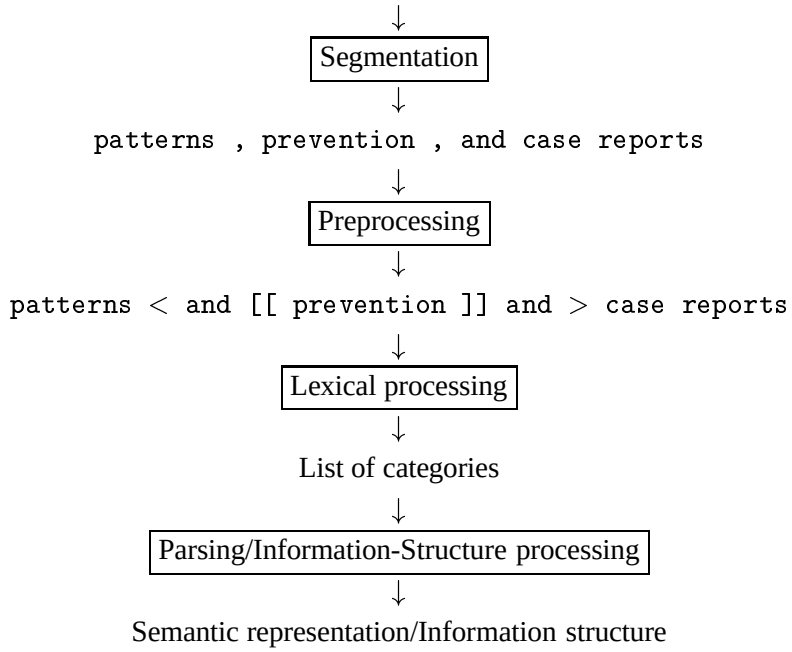
Among the methods discussed above, we adopt Karttunen's semantic equivalence check for its direct connection to the definition of spurious ambiguity and also for its conceptual simplicity. In support of this position, let us review some arguments against this approach. Eisner [1996] argues that a sequence of categories exemplified by "X/X ... X ... $X\backslash X$" can slow down a parser exponentially. Here we assume that X/X and $X\backslash X$ are 'modifiers' of X with *distinct* semantics, e.g., sentential adverbs. Then, this is an instance of genuine ambiguity because the result of "$[X/X+X]+X\backslash X$" and "$X/X+[X+X\backslash X]$" have distinct semantics. Syntactic approaches would consider them as spurious ambiguity. But, then, derivation of semantic representation would face incompleteness. Wittenburg [1987] objects to the cost of an equivalence check. But an equivalence check (for our semantics) is inherently easier than the general case of subsumption check. The latter requires the costly *occurs check* for soundness [Pereira and Shieber, 1987]. Hepple and Morrill [1989] raise another objection. While syntactic methods detect spurious ambiguities before deriving a result, an equivalence check needs to compare the derived result with every entry in the current table cell. However, the cost associated with the semantic method depends on how many genuinely ambiguous entries are in the cell but not on the number of spuriously-ambiguous entries (they are eliminated as soon as they are derived and do not accumulate). This does not introduce additional complexity that is specific to the spurious ambiguity check. Further, once semantics is involved, it is not possible to distinguish between spurious and genuine ambiguities unless we actually check the involved semantics.

The effect of spurious ambiguity for a practical parser is enormous. In the implementation of [Komagata, 1997a] (with a CKY-style parser), a sentence with more than 10 words mostly resulted in an out-of-space error. This result applies to both parsing and recognition because spurious ambiguity can derive syntactic types in an exponential manner. By eliminating spurious ambiguities with a mutual subsumption method, the performance of a CCG parser can be brought to a level comparable to other grammar-based parsers.

6.2.3 Linguistic Specification and Processing

In this subsection, we describe components of the parser in the order of processing as shown below.

(194) Input string: "*Pattern, prevention, and case reports*"



Segmentation

Segmentation is a simple finite-stage process that converts an input string of characters to a list of strings. Each string roughly corresponds to a word and many punctuation symbols. A sample specification (for English) to separate comma ',' from the attached word is shown below.

(195) `segmentation(e,",",[break_before=yes,break_after=yes,delete=no]).`

Preprocessing

The preprocessor is a finite-state string processor, and is an effective engineering solution to various problems. For example, frozen expressions, hyphens, numerals, and some punctuation are handled by the parser this way. The most significant effects can be seen in handling coordination. The preprocessor detects coordination patterns and replaces them in the following way (for 4-way coordination):

(196)
```
A, B, C, Coord D
        ↓
A < Coord [[ B Coord C ]] Coord > D
```

This allows the parser to apply the same coordinator for multiple-conjunct coordination. Replacement is done repeatedly for the preceding and succeeding parts, but not recursively (thus it is still finite-state). The processor tries to match patterns starting from 3-way coordination, up to the 5-way case. Currently, the preprocessor stops searching for alternative patterns once a solution is found. In this respect, the parser is not complete, but this has not been a problem in our case.

The double square brackets '`[[`' and '`]]`' force that the combination between them must be complete before combination with outside categories. This fixes the domain of, e.g., "`B Coord C`" and is found extremely effective, especially when `B` or `C` includes an embedded coordination (without comma) within a conjunct. The underlined phrase in the following example is fixed in this way.

(197) Laboratory work includes blood tests, liver and renal function studies, analysis of aspirated fluids, and sputum cultures.

Multiple instances of such a case are observed in the data.

There is a parasitic effect associated with comma replacement. Simply replacing a comma with a coordinator may destroy the original span of the conjuncts. The following example involves an instance of comma replacement (underlined as and) which may be analyzed incorrectly.

(198) Original: Treatment generally consists of daily doses of isoniazid, rifampin, and ethambutol.

a. * Treatment generally consists of {{daily doses of isoniazid and pyrazinamide} and ethambutol}.

b. Treatment generally consists of daily doses of {isoniazid and pyrazinamide and ethambutol}.

In order to avoid this problem, an additional pair of symbols '<' and '>' are used. They glue the entire span of the original coordination.

Preprocessing of frozen expressions is slightly different from the above case in that the specification is found in the lexical entries such as follows:

(199) `["x","-","ray"] := [lang=e,head="ray",class=n(c),infl=reg].`

For this reason, the preprocessor needs to maintain a set of frozen expressions and check the segmented list of strings against them. By adopting 'longest-match', the process prioritizes matching

frozen expressions over single-word entries.

Finally, the current process ignores discourse markers. Discourse markers indicate a relation between utterances and are not used in our analysis of information structure. This situation is currently handled by the preprocessor, which eliminates the following discourse markers: *and*, *but*, *so*, "*in addition*", *however*, and *therefore* in the environments shown below (with or without comma).

(200) *a*. However(,) ⟨*utterance*⟩

b. ⟨*subject*⟩(,) however(,) ⟨*rest of utterance*⟩

The second case is applied only to discourse markers without other functions, e.g., *however*, but not *and*, which is also a coordinator.

Lexical Processing

The lexical processing consists of identifying the matching lexical entry and assigning the corresponding categories. Some examples of lexical entries are shown below, but the details do not matter here.

(201) *a*.
```
"medicine" := [lang=e,class=n(u),infl=reg,pre_np=yes,arg=[pp(for)]].
```

b.
```
"his" := [lang=e,class=det(his),num_pers=[-,-,s3,-,-,p3],def=yes].
```

c.
```
"require" := [lang=e,class=v(reg),infl=reg-d,arg=[np,[np,pp(for)]]].
```

These are all English entries for the specified word classes. Information such as inflection, agreement, and subcategorization is also included. For our training data set (more detail on the data is in the next chapter) including 16 texts or 2300 words, there are about 900 such lexical entries including punctuation.

The following is an example of a singular noun inflection macro also including some other common properties. A macro is later used as a part of lexical assignment.

(202)
```
macro(e,noun_infl_sg,
      [if(class=n(_)),
      ifnot(pl_only=yes),
      lab=np(com),
      (if(human=H) -> [] ; [call(H=no)]),
      (if(sit=Sit) -> [] ; [call(Sit=no)]),
```

```
(if(class=n(c)) -> [call(UC=c),call(NP='-')] ;
[call(NP=np),(if(class=n(u)) -> [call(UC=u)] ; [call(UC=_)])]),
{if(gend=G)},
(if(implicit_arg=req) -> [call(Inf=yes)] ; []),
features=[agr=([-,-,s3,-,-,-],G,_),n_np=[n(UC),NP],human=H,def=_,
sit=Sit,inferrable=Inf],
locase_pf(Int),
int=Int,
cont=(cont,n:Int)]).
```

This macro only applies to the noun class `n`, specifies a syntactic type `np(com)`, sets features including agreement and human, and specifies the semantic representation as the lower case of the string. The above specification is written in a form of a simple procedural description language. There are assignment and conditional statements. These statements are interpreted by the system at the time of lexical look up.

The standard Montagovian analysis of NP is that there is a (common) noun category N (or CN) and determiner category NP/N. The current implementation deviates from this by assuming a single category NP for both of these. The distinction between N and NP is still maintained by the use of features. This approach has an advantage of reducing categorial ambiguity for, e.g., plural nouns. While N and NP are specified with features `n_np=[n(c),-]` and `n_np=[-,np]`, respectively, an ambiguous case is simply `n_np=[n(c),np]` (all for the countable case).

The above macro is used as a template for several lexical assignment for nouns including the following that subcategorizes a PP.

(203)
```
lex_assign(e,n(_),
[ifnot(num_req=yes),
incl(arg,pp(Prep)),
(macro(noun_infl_sg);macro(noun_infl_pl)),
lab=(Lab=>Lab/pp(Prep)),
(if(implicit_arg=req) -> [call(Src=arg)] ; [call(Src=self)]),
features=(F=>[npostmod=yes,composition=no]
/[composition=no,colon=no,context_link=proj(Src)]),
int=(Int=>PP^(Int-PP))]).
```

It calls the macro (both singular and plural cases), adds the PP argument, adds the features corresponding to the PP argument, and also adjusts the semantic representation to reflect this change.

For the training data set, there are about 200 lexical assignments (but not including macros) including different subcategorizations for verbs.

In the above, we have seen the specification of a noun class that takes a PP as an argument, rather than as a modifier. We take this position for most post-nominal PP's including situational ones. The motivation for the current position comes from difficulty with explosive ambiguity for considering PP's as post-nominal *modifiers*. Although this move might sound too restrictive, it actually corresponds to the difficulty in choosing the right preposition for a post-nominal PP modification, often experienced by non-native speakers. Thus, it seems justifiable that most noun-preposition relations must be specified.

The lexical processor reads the output of the preprocessor and assigns a set of categories to each string. For some string (e.g., *in*), over a dozen categories are assigned. In principle, looking up inflected forms takes a simple approach of generate and test. But to avoid the inefficiency of looking up unnecessary forms, the current implementation skips the cases where the stem is different from the target word. This technique has improved the performance of the current implementation over that of [Komagata, 1997a].

The system is capable of parsing both English and Japanese provided that the corresponding sets of lexical entries are prepared. But the current implementation only contains the English lexicon reflecting the scope of work. Although the implementation is also capable of dealing with generalized type-raised categories (GTRC) [Komagata, 1997a], which can simplify the grammar for Japanese, the capability is not activated because it is not necessary for English.

Use of Features

The focus of Komagata [1997a] was elimination of spurious ambiguity. The paper avoided the issue of genuine ambiguity by working mainly on recognition. In the previous implementation [Komagata, 1998a], I tackled a small, but noticeable part of genuine ambiguity as well. It is a subclass of genuine ambiguity that can be resolved by use of features. Let us call it 'absurd' ambiguity. This subclass must be distinguished from the main, and more difficult type of genuine ambiguity that requires domain-specific or more general world knowledge.

Absurd ambiguities are eliminated by using both syntactic and semantic features in the grammar. Some of these features are to (i) limit modification structures in and around NPs, (ii) restrict

coordination patterns, (iii) condition on the modification of adjectives by *more*/*most*, and (iv) apply the human/non-human distinction.

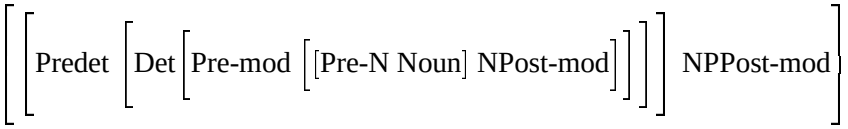
The following examples show an absurd modification with respect to syntax possibly allowed by a coarse grammar.

(204) *a*. * [minor skin] complication, cf. minor [skin complication]

b. *[tuberculosis in a young baseball] player, cf. tuberculosis [in a young baseball player]

A coarse grammar that allows noun-noun compounds in a reasonably general way may face absurd ambiguities like this. As a first technique to reduce these absurd ambiguities, the current grammar assumes and imposes the following structure in/around NP, mainly adopting the analysis in [Quirk et al., 1985].

(205)



a. Predet: predeterminers such as *such* and *half*

b. Det: determiners such as (in)definite article

c. Pre-mod: premodifier such as adjective

d. Pre-N: noun to form a noun-noun compound with the head noun

e. NPost-mod: post-nominal modifier such as PP, restrictive relative

f. NPPost-mod: post-NP modifier such as appositive, non-restrictive relative

This restriction is achieved by the use of features such as `premod=yes` or `npostmod=yes` for results of pre- and post-nominal modification. The head noun that allows noun-noun compounds has features `[premod=no,npostmod=no]` to avoid these 'heavy' words.

The distinction between nominal and NP modification is crucial. For example, "*acute injuries typical of the sport*" must be analyzed as "[*acute injuries*] *typical of the sport*", not as "*acute* [*injuries typical of the sport*]". The modifier "*typical of*" is assigned a feature `npostmod=no` and is prevented from modifying the noun *injuries*. If there is no adjective *acute*, the noun *injuries* can be successfully modified by a post-NP modifier because it is underspecified between a noun and an NP.

There is another case involving nouns that can form a noun-noun compound. For this case, coordination is the primary factor. A phrase "*exercise modifications or medications*" should be analyzed as "[*exercise modifications*] *or medications*" but not as "**exercise* [*modifications or medications*]". Now, both *modification* and *medication* are allowed to form a noun-noun compound, e.g., "*exercise modifications*" and "*antihypertensive medications*". Thus, a general form of coordination allows the unintended analysis. To avoid this, the current implementation adds a procedural constraint to exclude noun-noun compounds where the second noun is a result of coordination.

Absurd modification may also involve lexico-semantic aspects as can be seen in the following example.

(206) *a*. *[most lateral] ankle sprains

b. most [lateral ankle sprains]

c. cf. [most unusual] ankle sprains

The lexical specification of adjectives includes whether it can be modified by *more* and *most* (this information is shared as an inflectional feature whether the adjective can have suffix forms of comparative/superative).

The parser also uses the feature 'human' for various purposes including subject-verb agreement, modification, and coordination. Without this feature, the expression "refining rehabilitation" can be ambiguous between "an act of (someone's) refining rehabilitation" or "rehabilitation that refines something", as in the well-known "flying planes" example. In our case, the verb entry for *refine* specifies that the subject be 'human', eliminating the latter possibility.

Finally, the current grammar specifies agreement and subcategorization fairly accurately. For example, in addition to subject-verb agreement, the grammar specifies agreement for relative clauses including the possessive form and various coordination patterns. This helps the disambiguation process substantially.

Parsing

The list of sets of categories obtained through the lexical processing is now fed to the CKY-style parser based on Aho and Ullman [1972]. Informally, a CKY-style algorithm parses "*Felix praised Donald*" using a chart, as shown in Fig. 6.2.

	Column *a*	Column *b*	Column *c*
Row 1	Felix	praised	Donald
Row 2	[Felix praised]$_{1a+1b}$	[praised Donald]$_{1b+1c}$	
Row 3	[Felix praised Donald]$_{2a+1c}$ [Felix praised Donald]$_{1a+2b}$		

Figure 6.2: CKY-Style Parsing Table

Starting from the lexical categories for the entries in row 1, the parser proceeds to a lower row by combining the component categories specified in the subscript. In Row 3, multiple entries with exactly the same results are obtained. This is an example of spurious ambiguity. As we have discussed, we adopt a mutual subsumption check to eliminate spurious ambiguity. For the above case, since the two entries in 3*a* have equivalent semantics, they are reduced to a single entry. This process takes place whenever a new entry is entered into a cell. The situation gets more complicated once structured meaning is introduced. We will come back to this topic in the next section.

The linguistic specification file contains the following CCG rules to combine categories.

(207)
```
ccg_rule(e,[x/y, y] =>x,[]).                (>)
ccg_rule(e,[y, x\y] =>x,[]).                (<)
ccg_rule(e,[x/y, y/z] =>x/z,[]).            (>B)
ccg_rule(e,[x/y, y/z/u] =>x/z/u,[]).        (>B²)
ccg_rule(e,[y?z, x\y] =>x?z,[]).            (<B(×))
ccg_rule(e,[x, &, x] =>x,[]).               (<&>)
```

The question mark '?' is used for underspecifying the slash directionality (but the two instances of '?' must agree). This rule specification is interpreted by the program for the corresponding operation. In the present implementation, type raising is considered a unary rule, and is activated dynamically when categories *NP* or *PP* are inserted into the CKY table. *NP* is type raised to $S/(S\backslash NP)$ and $S\backslash(S/NP)$, and *PP* is type raised to $S\backslash(S/PP)$.

6.2.4 Performance

In this subsection, we discuss the performance of the parser for the training data (i.e., before extension to the test data), and show that it provides a reasonable backbone for analyzing information structure.

The system parses 16 introduction sections of medical case reports including 131 utterances. The average word length for an utterance (after preprocessing and including punctuation symbols) is 20, and the maximum, 42. There are four utterances beyond this level. Since they slow down the process so much, they are divided into two segments. Unfortunately, the utterances of 40 or more words seem to be beyond the capacity of the system. After this preparation, the measured CPU time is on average 16 seconds per utterance.[4] The average number of parses per utterance is 16. While there are a number of utterances that take too long for an interactive response, many utterances can be parsed in the order of *seconds*.

The above performance does not appear as good as the previous version [Komagata, 1998a] implemented for the abstracts of the same journal. The average parse time was about 2 seconds per utterance. But there are several factors involved in this difference. The average utterance length increased from 17 to 20. If we assume cubic parsing complexity in practice, this translates to 60% increase in parse time. The total size of the lexicon has increased about 50%. This proportionally slows down the lexical look up time. The parser now processes structured meaning. As we discuss at end of Subsection 6.3.3, structured meaning can introduce additional complexity. This seems to be reflected in the increase in average number of parse from 2 to 16. Considering all these, the performance of the present parser seems to scale reasonably from the previous implementation.

Since the goal of this implementation is to provide an adequate platform for analyzing information structure, no comprehensive comparison with other parsing systems is made. Informal side comments are that those long sentences are very difficult for a large-scale grammar-based parsers. For example, the XTAG parser [Doran et al., 1994] would have difficulty parsing many of the long sentences in our texts. Since the XTAG system has hundreds of thousands of lexical entries and up to dozens of trees for each lexical entry, this is only a confirmation that parsing real data is still challenging.

We thus conclude that the parser can be a reasonable platform for analyzing information structure. Two major factors are the use of CKY-style parsing algorithm and the elimination of spurious ambiguity, in comparison to earlier experimental parsers [Prevost, 1995; Hoffman, 1995]. The

[4]Time measurement is done by Sicstus built-in predicate `statistics`. The time measurement includes most of the stages: segmentation, preprocessing, lexical processing, and CKY parsing with derivation of semantic representations (structured meanings). There are a few off-line processes such as asserting (i) relations between a word-form and the canonical form and (ii) a set of frozen expressions. These can be done in a negligible time.

parser also demonstrates improvements over the version in Komagata [1997a] due to preprocessing, more efficient lexical processing, and use of features.

6.3 Processing Information Structure

This section presents the key element of the system, the information-structure analyzer. This module is a straightforward implementation of the theory developed in Chapter 3 and formalized in Chapter 4. It includes a small number of procedural aspects, but they are modularly specified and can be upgraded when necessary.

Each step of processing information structure is associated with a step of parsing. Parsing steps consist of lexical processing and combination of two (non-coordination) or three categories (coordination). Thus, this section only describes *local* processes applicable to either lexical or combinatory process.

In this section, we discuss the three properties for contextual links, composition of structured meaning, identification of information structure, and prediction of particle choices in Japanese.

6.3.1 Discourse Status and Domain-Specific Knowledge

Discourse Status

As a consequence of adopting a CKY-style algorithm for parsing CCG, semantic representations corresponding to information structure units are available in CKY table cells. In order to analyze discourse status, we modify the CKY table so that table cells only contain *pointers* to categories, not categories themselves. Categories are stored in the discourse context by the `assert` predicate of Prolog. Then, we can easily decide whether the category should remain in the context. Basically, we keep all the categories that are used in a successful parse.[5] Then, we can define the notion of discourse-oldness as presence of an equivalent category in the discourse context. The process of identifying discourse-old referents utilizes Prolog's unification mechanism. In order to correctly identify the existence of an equivalent semantic representation, we use mutual subsumption, not

[5]The actual process of asserting categories is slightly more complicated. Categories are initially assigned a temporary status until the parsing process completes. After completion, a top-down process traces down the successful parses and changes the temporary status to a permanent one. The unsuccessful categories are then eliminated.

simple unification.[6]

When there are multiple occurrences of identical semantic representations in a single utterance, only one instance is asserted and pointed to from multiple CKY-table cells. At this point, analysis of discourse-oldness is applied only across utterances. Thus, intra-utterance reference cannot be made. This is not a problem for the analysis of information structure, as will be seen in the next chapter.

Situationally-Available Referents and Domain-Specific Knowledge

The above discourse-status processing can be applied to the analysis of situationally-EVOKED referents as well. For example, pronouns such as *we* and *they* are asserted at the beginning of an analysis under the assumption that these are situationally available.

The present proposal also assumes domain-specific knowledge that referents such as *physician*, *clinician*, and *patient* are available in the domain. This assumption can be implemented exactly the same way as for the above case of pronouns. That is, common nouns *physician*, *clinician*, and *patient* are asserted at the beginning of an analysis. Thus, this case too can be handled by the same mechanism as that for discourse status.

Use of Morphological Forms

There is one procedural aspect added to the lexical process. In identifying discourse status, we also use morphological forms as a cue [see Dahl et al., 1987 and Palmer et al., 1993 for an analysis of derivational forms]. For example, the use of a verb *damage* is assumed to imply that there is a damaging event. Then, a NP "*the damage*" may be considered to refer to that event. This is in a sense a combination of linguistic marking of contextual linking and discourse status because we identify the contextual-link status of a word only if a morphologically-related referent is discourse-old.

Currently, the system deals with the following cases:

(208) *a*. Nouns: between singular and plural forms

b. Adjectives: between base, comparative, and superative forms

[6] Sicstus Prolog has a built-in predicate called `variant` which does exactly the mutual subsumption, i.e., identity except for variable names.

c. Verbs: between inflected forms, e.g. *damages* and *damaged*

d. Derivation: between a noun and a verb with a shared sense

The system realizes the above condition by keeping content information (usually a dictionary form) as in the underlined portion below.

(209)
```
macro(e,noun_infl_sg,
     [if(class=n(_)),
          .
          .
     locase_pf(Int),
     int=Int,
     cont=(cont,n:Int)]).
```

Here, the attribute `cont` (for content information) has a pair of values. The first component `cont` indicates that the entry is a content word and not a function word. The second component `n:Int` indicates that the entry is a noun with a key value shared among different word classes.[7] For example, a noun *damage* contains a feature `cont=(cont,n:damage)` and the verb *damage* contains `cont=(cont,v:damage)`. The system checks the noun-verb relation by comparing the specification but ignoring the difference between the word classes `n` and `v`. For inflection, the entire content specifications are compared. We need to use this feature rather than semantic representations because the latter naturally differ between the cases mentioned above.

The above process for morphologically-related forms is only available at the lexical level. But its effect may project to a more complex structure exactly like other contextual links.

6.3.2 Linguistic Marking of Contextual Links

This subsection describes how the system processes linguistic marking of contextual links. The discussion covers the following topics: lexical assignment of categories, composition of two categories, a special case involving utterance-initial modifiers, and coordination.

Lexical Processing

There are a few cases where linguistic marking of contextual links needs to be processed at the time of lexical processing. First, function words are assigned contextual-link status. This

[7] In this example, `Int` is unified with the phonological form of the entry.

class includes: auxiliary verbs, modals, prepositions, and subordinators. They have the feature `cont=(func,`*FuncWordType*`)` where *FuncWordType* specifies a type of function word. Since function words are available in the grammar, we can associate them with zero-inference. Thus, it seems natural to assume a contextual-link status for them.

Another case is two-place nouns such as *page* (see discussion on p. 68 in 3.3.1). This type of nouns can be considered to have an implicit argument without a PP argument, and thus is assigned a contextual-link status. The process needs to check if the category is *NP* without arguments and the feature `implicit_arg=req` is specified.

Finally, numerals with the category `num` are assigned a *non*-contextual-link status.

Composition of Two Categories

In Chapter 4, we presented a specification of linguistic marking (of contextual links) in terms of feature unification associated with categories. The system still uses features for this purpose, but implements them in a slightly different way. To avoid cluttering the feature area and to consolidate specifications shared in different categories, the system includes a special module to deal with assignment and projection of contextual-link statuses.

For example, a contextual-link projection from the argument is specified as a feature "$cl = proj\,(arg)$" (cl for contextual link) on the argument of the functor category, and the special module processes structured meaning according to the specification shown below.

(210)

	Determiner	Noun
Example:	*many*	*researchers*
Syntactic type:	$NP/\underset{cl=proj(arg)}{N}$	N
	$\langle C_1, - \rangle$	$\langle -, N_2 \rangle$
Syntactic type:	NP	
	$\langle -, N' \rangle$	

There are three more features corresponding to the specification of contextual-link assignment/projection: "$cl = set$", "$cl = reset$", and "$cl = proj\,(self\,f)$".

We now move to specific cases. First, let us discuss some special cases: composition with dummy categories (e.g., punctuation) and function words as an argument (e.g., particles), and composition of two function words. In these cases, function words are handled transparently. That

is, the result is a projection of the contextual-link status of the other component. Composition of two function words is treated as a new function word.

Next, the process sets the contextual-link status to the following: definite determiner, indefinite generic, and utterance-initial modifier. The type of compositions involving the definite determiner can be represented as "$X_{[def=yes]}/Y+Y$". That is, only the feature `def=yes` on the result category specifies the process. This specification is more general than explicitly specifying "$NP_{[def=yes]}/NP+NP$". Thus, the notion of 'definiteness' (for the purpose of setting a contextual link) can be extended to other categories as well. For an indefinite generic, the composition can be represented as: "$X_{[def=no]}/Y+Y$" with the additional condition that Y is a contextual link. The contextual-link status is set only if the right category, Y, is a contextual link. Utterance-initial modifiers receive a contextual-link status if the result of composition is S/S. Inverted phrases also assign a contextual-link status.

The case that assigns a non-contextual-link status is analogous. For an indefinite article, "$X_{[def=no]}/Y+Y$" is specified. For numerals of the modifier type, the following pattern is detected and processed "$X/Y_{[cl=reset]}+Y$".

The system may also project the contextual-link status from an argument to the result. This takes place for the pattern "$X.../Y_{[cl=proj(arg)]}+Y$", its directional variant "$Y+X...\backslash Y_{[cl=proj(arg)]}$", and for a complex argument "$X.../\left(Y_{[cl=proj(arg)]}/Z\right)+(Y/Z)$".

Projection of the contextual-link status from itself is similar. The same set of patterns are currently implemented: "$X.../Y_{[cl=proj(self)]}+Y$", "$Y+X...\backslash Y_{[cl=proj(self)]}$", and "$X.../\left(Y_{[cl=proj(self)]}/Z\right)+(Y/Z)$".

Composition of an Utterance-Initial Modifier and the Subject

There is a case where linguistic marking functions slightly differently from the previous cases. It involves an utterance-initial modifier, analyzed as $\langle C_1,-\rangle$, composing with the main clause with a structured meaning $\langle \underset{\text{contextual link}}{C_2}, \underset{\text{non-link}}{N_2} \rangle$ where C_2 is the subject. If the combination of C_1+C_2 is a contextual link, the resulting structured meaning is $\langle C',N_1\rangle$ where $C'=C_1+C_2$. This is a kind of discontiguous information structure and possible even though the combination of the utterance-initial modifier and the subject cannot form a constituent in English.[8] If C' is not a contextual link

[8]For example, such a phrase cannot form a conjunct in English.

on its own, the resulting structured meaning would be $\langle C_1, C_2 + N_2 \rangle$, i.e., the entire main clause becomes a rheme. But an observation of the experiment data suggests that in many cases, the informational partition in the main clause seems as strong as the case without an utterance-initial modifier. To accommodate this situation, we assume the following hypothesis:

(211) (Operational hypothesis) The utterance-initial modifier is not only a contextual-link marker of the modifier phrase itself but also a marker of the discontiguous theme including the subject where the subject is a contextual link.

Such a discontiguous theme can satisfy the condition of a discontiguous structured-meaning component: the semantic representation of the utterance-initial modifier and the subject can compose to derive a sound semantic representation. For example, an adverb *yesterday* with "$\lambda X.S//yesterday'$" and the subject *John* with $\lambda P.P(john')$ can derive "$\lambda P.[P(john')]//yesterday'$". Note that the notation $X//Y$ is used for a modification (or adjunct) structure, which is distinguished from the functor-argument structure. In terms of information structure, there is no reason such a semantic representation cannot be a (discontiguous) theme. In fact, Japanese allows coordination of a phrase corresponding to "*yesterday–John*" with another phrase, say, "*today–Mary*". In this case, the subject must be compatible with the type-raised form $S/(S\backslash NP)$. Then, a modifier-subject composition can be recognized as "$S/S + S/(S\backslash NP) \Longrightarrow S/(S\backslash NP)$" with the intended semantics.

In terms of assignment/projection of contextual links, we can consider that utterance-initial modifiers either (i) project the contextual-link status of the subject or (ii) project the status of itself. In the system, the same function is performed by the above-mentioned module that deals with assignment/projection of contextual links.

Summary

The process of identifying contextual links is summarized in Fig. 6.3 on page 180.

6.3.3 Composition of Structured Meaning

Perhaps, the most innovative feature of the current system is implementation of structured meaning in a fairly general sense. This subsection describes the implementation of the ideas formalized in

Section 4.3. At the end, we also describe the way we deal with spurious ambiguity in relation to structured meaning.

Data Types for Structured Meaning

Let us represent a structured meaning in the following form:
$\langle \underset{\text{contextual link}}{C}, \underset{\text{non-link}}{N} \rangle_{LeftBoundary-RightBoundary}$. Although we allow arbitrary discontiguous construction of C and N, we distinguish instances of structured meaning only by the boundary categories. Then, we have the following six possible types of structured meanings for inputs and results: $\langle C,N\rangle_{C-C}$, $\langle C,N\rangle_{C-N}$, $\langle C,N\rangle_{N-C}$,$\langle C,N\rangle_{N-N}$, $\langle -,N\rangle$, $\langle C,-\rangle$.[9] As a consequence, the number of composition rules is bounded. The recursive process of dealing with structured meanings is defined for the lexical and the derivation steps (Subsection 4.3.1). The existence of the bound on the derivational process thus guarantees a closed operation.

To be complete, we have to discuss all the possible combinations, i.e., 216 (see p. 115). But, since it is tedious and not all the cases are equally common, the system only implements about 20 possibilities. In the following, we look at a few common cases among those discussed in Subsection 4.3.1. Note that we use the notations $Type_C$ and $Type_N$, representing the syntactic type corresponding to C and N, respectively.

Composition Type: $\langle C_1,N_1\rangle_{N-C} + \langle -,N_2\rangle$

Let us first recall the case where this type of composition is needed. In Subsection 4.3.1, we observed the non-traditional derivation of "[**Fred** praised] [**Donald**]", e.g., as a response to "Who praised who?". The component "**Fred** praised" is analyzed as $\langle praised', fred' \rangle_{fred'-praised'}$ where the contextual-link and non-link components of the structured meaning are $praised'$ and $fred'$, respectively, and the left and the right boundaries are $fred'$ and $praised'$, respectively.

Now, the composition in question, $\langle C_1,N_1\rangle_{N-C} + \langle -,N_2\rangle$, would result in another structured meaning, $\langle C_1,N'\rangle_{N-N}$ where, N' is a semantic composition of N_1 and N_2 and the boundary $N-N$ indicates that C_1 is not at the boundaries. But, as we have discussed in Subsection 4.3.1, this N' must satisfy certain conditions so that the composition of C_1 and N' can result in the correct

[9] Note that $\langle C,-\rangle$ and $\langle -,N\rangle$ are the cases where the entire phrase is a contextual link and a non-contextual link, respectively.

semantic representation corresponding to the entire phrase. For the current example, it must be $\lambda P.P(donald')(fred')$. We say that this semantic representation is 'correct' reflecting that it can combine with the verb $\lambda X.\lambda Y.praised'(X)(Y)$ with the correct result. We also require that this be guided by an appropriate syntactic process, i.e., functional composition of two type-raised categories $S/(S\backslash NP)$ and $(S\backslash NP)\backslash(S\backslash NP/NP)$ with the result, $S\backslash(S\backslash NP/NP)$.[10]

The conditions described above can be stated in the following way (the notation $Type_{fred'}$ denote the syntactic type corresponding to $fred'$, i.e., $S/(S\backslash NP)$ in the above example):

(212) *a*. There is some syntactic type $Type_{N'}$ such that $Type_{N_1} + Type_{N_2} = Type_{N'}$
There is some semantic representation N' such that $N_1 + N_2 = N'$

b. Either of the following holds:

(i) "$Type_{C_1} + Type_{N'}$" results in the correct syntactic type of the entire phrase *and* $C_1 + N'$ results in the correct semantic representation of the entire phrase

(ii) "$Type_{N'} + Type_{C_1}$" results in the correct syntactic type of the entire phrase *and* $N' + C_1$ results in the correct semantic representation of the entire phrase

The condition (*b*) allows either direction because the position of C_1 relative to the composition of N_1 and N_2 no longer corresponds to the surface order, and becomes 'virtual'.

If the above conditions are not satisfied, this derivation is not available. Another possibility for the above example is the traditional derivation, "[**Fred**] [praised **Donald**]". The conditions for this case is analogous, but the current implementation has a fail-safe case, which allows the result of the form $\langle -, N'' \rangle$ where N'' is the semantic representation of the entire phrase. That is, no contextual link survives the composition.

Composition Type: $\langle C_1, N_1 \rangle_{C-N} + \langle -, N_2 \rangle$

This case corresponds to the example "[Felix **praised**] [**Donald**]", e.g., in response to "What about Felix?". The non-traditional constituent in this case is $\langle felix', praised' \rangle_{felix'-praised'}$ where $felix'$ and $praised'$ are contextual-link and non-contextual link, respectively, in that order at the surface. The condition is similar to the previous case except that in this case, the surface order between C_1 and N' $(N_1 + N_2)$ is fixed.

[10] The detail of this composition is described in Subsection 4.3.3.

The conditions are thus specified as follows:

(213) *a*. There is some syntactic type $Type_{N'}$ such that $Type_{N_1} + Type_{N_2} = Type_{N'}$

There is some semantic representation N' such that $N_1 + N_2 = N'$

b. "$Type_{C_1} + Type_{N'}$" results in the correct syntactic type of the entire phrase *and*

"$C_1 + N'$" results in the correct semantic representation of the entire phrase

For this case, the traditional derivation "[Felix] [**praised Donald**]" is also possible. And, it is probably more natural in general. Thus, the analysis may end up with a spurious ambiguity. Elimination of spurious ambiguity involving structured meanings will be discussed at the end of this subsection.

Composition Type: $\langle C_1, - \rangle + \langle C_2, - \rangle$

The last case examined here is a composition of two contextual links. For example, consider an example $\langle praised', - \rangle + \langle donald', - \rangle$. This would result in $\langle \lambda Y.praised'(donald')(Y), - \rangle$ if $\lambda Y.praised'(donald')(Y)$ is indeed a contextual link. We enforce this requirement (146*b*) as follows:

(214) *a*. "$Type_{C_1} + Type_{C_2}$" results in the correct syntactic type

"$C_1 + C_2$" results in the correct semantic representation C'

b. C' is a contextual link.

Then, the resulting structured meaning is $\langle C', - \rangle$. Otherwise, the current implementation assumes $\langle C_1, C_2 \rangle_{C-N}$, but not $\langle C_2, C_1 \rangle_{N-C}$. This is a disambiguation heuristic and a weak form of 'theme-first' principle. In practice, when both the subject and the predicate are contextual links (and thus either can be a theme), this heuristics appears as choosing the subject as a theme.

Other cases discussed in Section 4.3.1 are analogous.

Structured Meaning and Spurious Ambiguity

Integration of structured meaning with our CCG parser complicates the situation involving spurious ambiguity. The elimination method based on mutual subsumption needs to be redefined because the comparison between the result semantic representation does not reflect potential difference in structured meaning. The adopted solution is to apply mutual subsumption check to each

component of structured meaning. For example, to compare $\langle C_1, N_1 \rangle$ and $\langle C_2, N_2 \rangle$, mutual the subsumption of C_1 and C_2 *and* that of N_1 and N_2 are checked.

When both of the involved structured meanings are the type of $\langle C, - \rangle$ or $\langle -, N \rangle$, the case reduces to the original mutual subsumption check. Although components C and N in $\langle C, N \rangle$ may consist of discontiguous elements, the proposed method is along the same line with the original mutual subsumption check, which ignores the syntax.

Since we have estimated that the practical maximum of distinct structured meanings for a category is 16 (p. 114), we also expect that each category may correspond to up to 16 structured meanings. But integration of structured meaning even with the presence of spurious ambiguity does not in practice introduce an exponential explosion.

6.3.4 Identification of Information Structure

As discussed in Subsection 4.3.2, once we adopt the structured meaning approach, its identification of information structure is almost trivial. At the last semantic composition, simply retrieve C and N from $\langle \underset{\text{contextual link}}{C}, \underset{\text{non-link}}{N} \rangle$. But there is one procedural aspect we should consider here.

In Subsection 2.3.4 (p. 42), we have mentioned the possibility of accommodated theme. The following is the example used there.

(215) Q: Who did Felix praise?

A_1: [Felix praised]$_{Theme}$ [Donald]$_{Rheme}$. (direct response)

A_2: [Felix]$_{Theme}$ [praised Donald]$_{Rheme}$.

A_3: [Felix praised Donald]$_{Rheme}$.

The current implementation adopts a heuristic to picks up only the possibility (A_1), corresponding to the maximal theme. This is achieved by checking the dominance relation between categories. This way, only the themes that are not dominated by another survive. This process does not guarantee a unique theme, though. There may be incomparable maximal themes, e.g., the subject and the object where the verb is a non-contextual link.

In order to choose the most likely particle based on the information structure, we adopt an additional heuristic. The system prioritizes the patterns of information structure according to the conditions below. They are arranged from the highest priority to the lowest.

(216) *a*. The subject + verb is the theme. Thus, the subject is a part of the theme. (Code 12)

b. The predicate is the rheme. Thus, the subject is the theme. (Code 49)

c. The predicate is the rheme and is one-place and negative. Thus, the subject is a contrastive theme. (Code 50)

d. The predicate is the theme. Thus, the subject is the rheme. (Code 51)

e. The adjectival predicate is the rheme. Thus, the subject is the theme. (Code 82)

f. The subject + verb is the rheme. Thus, the subject is a part of the rheme. (Code 88)

g. The verb is the rheme. Thus, the subject is a part of the theme. (Code 91)

h. The verb is the theme. Thus, the subject is a part of the rheme. (Code 99)

The general idea is to choose a larger theme. So far, no obvious errors have been observed due to the above prioritization.

6.3.5 Prediction of Particle Choice in Japanese

The last step of the automatic process is prediction of particles in Japanese. Following the analysis in Chapter 5, the procedure simply predicts *wa* for the matrix-level subject that is a part of the theme, and *ga* otherwise.

The prediction procedure in Chapter 5 includes special cases such as parallel clauses, one-place negative predicates, and one-place stage-level predicates. Among these, only the case of one-place negative predicate has been implemented.[11] The other two cases may result in incorrect predictions. In particular, our lexicon does not yet reflect stage and individual-level distinction. There is one case where stage-level predicates appear in a coordination.

For the case of purely semantically-conditioned contrastive *wa*, there is no way of mechanically identifying them. But we have seen that very few of them are rhematic through the mini-corpus analysis in Subsection 5.2.3 and that there is none in our experimental data. Thus, we do not consider the case where a contrastive *wa* must be chosen within a rheme in place of *ga* marking.

To identify the matrix-level grammatical subject, we adopt a definition of subject as the least oblique argument of a predicate [Steedman, 1996, p. 21]. The system first checks for all-theme

[11] Although this case is implemented for the process of identifying information structure, it is not used in the evaluation process.

$\langle C, - \rangle$ and all-rheme $\langle -, N \rangle$ cases. Once these possibilities are excluded, the system identifies the presence of a modifier-clause relation between C and N in $\langle \underset{\text{contextual link}}{C}, \underset{\text{non-link}}{N} \rangle$. After excluding the clausal modifier, if any, the process checks the matrix-level predicate-argument structure and detects whether the subject, the least oblique argument, is in C or N. Depending on whether the subject is in C or N, the system identifies its theme/rheme status and thus predicts *wa* or *ga*. The semantic representation of adjectival and passive predicates are treated similarly. When a *by*-phrase is present in a passive construction, it is placed as an argument more oblique than the subject.[12]

The actual output of the program looks like the following. The listing is to demonstrate the state of implementation, the detail does not concern us.

```
>>>>> Utterance Number: 12-1 <<<<<

Seg: Osteoporosis in Active Women : Prevention , Diagnosis , and treatment (11 wo
rds)

Preprocessed: Osteoporosis in Active Women : Prevention < and [[ Diagnosis ]] and
> Treatment (14 words)

Result: cat(bas(np(com),[colon=yes]),osteoporosis-(pat_agt-(woman:pl//active-woma
n:pl))//colon(and:[prevention,diagnosis,treatment]),[nil,181,id])

Result: cat(bas(s(fin),_28902)-(/,bas(s(fin),_28902)-( bas(np(com),[colon=yes]))
),((osteoporosis-(pat_agt-(woman:pl//active-woman:pl))//colon(and:[prevention,diag
nosis,treatment]))^_28873)^_28873,[nil,182,id])

Number of parses: 2

 CPU time: 1330 ms  Elapsed: 1600 ms

>>>>> Utterance Number: 12-2 <<<<<

Seg: Osteoporosis has been defined as ‘ ‘ a disease characterized by low bone mas
s and microarchitectural deterioration of bone tissue , leading to enhanced bone f
ragility and a consequent increase in fracture risk . ’ ’ (36 words)

Preprocessed: Osteoporosis has been defined as ‘‘ a disease characterized by low
bone mass and microarchitectural deterioration of bone tissue , leading to enhance
```

[12]One evidence for this position is the binding phenomenon, e.g., “Felix is praised by himself” vs. “*Himself is praised by Felix”. This suggests that the subject of a passive is in a ‘commanding’ position in whatever the structure assumed for binding process, e.g., predicate-argument structure in our case.

d bone fragility and a consequent increase in fracture risk . '' (34 words)

Result: cat(bas(s(fin),[be_verb=no]),aux(perf)-(define-risk-(as-and:[indef-(disea
se//characterize-(by-and:[mass-(? -bone)//low-(mass-(? -bone)),deterioration-(of-(
tissue-(? -bone)//lead-(to-(fragility-(? -bone)//enhance-(fragility-(? -bone))-_23
71))-(tissue-(? -bone))))//microarchitectural-(deterioration-(of-(tissue-(? -bone)
//lead-(to-(fragility-(? -bone)//enhance-(fragility-(? -bone))-_2371))-(tissue-(?
-bone)))))])-disease),indef-(increase-(in-fracture)//consequent-(increase-(in-frac
ture)))])-osteoporosis),[40,2951,cn])

(54 other parses omitted)

Number of parses: 55

CPU time: 218370 ms Elapsed: 331630 ms

*** IS analysis:
- Theme(osteoporosis/40):Rheme(_3268^(aux(perf)-(define-risk-(as-and:[indef-(dise
ase//characterize-(by-and:[mass-(? -bone)//low-(mass-(? -bone)),deterioration-(of-
(tissue-(? -bone)//lead-(to-(fragility-(? -bone)//enhance-(fragility-(? -bone))-_3
174))-(tissue-(? -bone))))//microarchitectural-(deterioration-(of-(tissue-(? -bone
)//lead-(to-(fragility-(? -bone)//enhance-(fragility-(? -bone))-_3174))-(tissue-(?
-bone)))))])-disease),indef-(increase-(in-fracture)//consequent-(increase-(in-fra
cture)))])-_3268))/2951)

(1 another information-structure analyses omitted)

=> Particle prediction (matrix subject): >>wa<< (case 49)

>>>>> Utterance Number: 12-3 <<<<<

Seg: Although anyone can develop osteoporosis , postmenopausal women and young fe
males with menstrual irregularities are most commonly affected . (19 words)

Result: cat(bas(s(fin),[]),affect-and:[woman:pl//postmenopausal-woman:pl,female:p
l-(prop-(irregularity:pl//menstrual-irregularity:pl))//young-(female:pl-(prop-(irr
egularity:pl//menstrual-irregularity:pl)))]-_61406//most//commonly//although-(aux(
can)-(develop-osteoporosis-pron(anyone))),[3288,3440,cn])

(2 other parses omitted)

Number of parses: 3

```
 CPU time: 4430 ms  Elapsed: 7320 ms

*** IS analysis:
- Theme(_49161^(_49161//although-(aux(can)-(develop-osteoporosis-pron(anyone))))/
3288):Rheme((affect-and:[woman:pl//postmenopausal-woman:pl,female:pl-(prop-(irregu
larity:pl//menstrual-irregularity:pl))//young-(female:pl-(prop-(irregularity:pl//m
enstrual-irregularity:pl)))]-_49116//most//commonly)/3440)

=> Particle prediction (matrix subject):  >>ga<<  (case 89)
```

6.3.6 Potential Applications to Generation

Before concluding this chapter, let us briefly discuss the possibility of applying the identification process (of information structure) to natural language (NL) generation. As implemented in Prevost [1995] and Hoffman [1995], the basic idea is that certain linguistic forms are associated with either the theme or the theme. Once we identify the information structure of an utterance, we can eliminate linguistic forms incompatible with the identified information structure. Here, we consider two examples. Text-to-speech generation in English and English-Turkish machine translation.

For the case of text-to-speech generation, we can identify the information structure of the utterances in the text. As we have discussed earlier, certain pitch accents are associated with the theme and the rheme [Steedman, 1991a], e.g., L+H* for a theme and H* for a rheme. Depending on whether a contrast falls within a theme or a rheme, we can predict the appropriate intonation. This process of predicting intonation applies to arbitrary word class, cf. particle choice for the matrix subject in Japanese.

In English-Turkish machine translation, we may adopt the function of word order in relation to information structure. For example, following Hoffman [1995], we may identify the utterance-initial element as a theme and the pre-verbal element as a rheme. Once we identify the information structure of the utterances in the texts in English, we can choose an instance of word order in Turkish that is consistent with the identified information structure. Thus, for the type of NL generation where the input is a text, the current approach can provide useful information for generating contextually appropriate linguistic forms with respect to information structure.

6.4 Summary

In this chapter, we demonstrate that our CCG parser performs reasonably well for the purpose of information-structure analysis. The most critical element in the implementation is elimination of spurious ambiguity. We show that the semantic equivalence check can be extended to the case where structured meanings are also involved.

The module that analyzes information structure is realized as a straightforward implementation of the specification of contextual link and integration of structured meaning. In addition, several procedural aspects are addressed and integrated in a modular fashion. This allows us to upgrade the system with new specification for these procedural aspects.

Note: CL for a contextual-link status and NL for a non-contextual-link status

- **Initial context** (once at the beginning of a text): Set CL for the following:
  - Pronouns and situation words: *he*, *it*, *such*, *these*, *they*, *this*, *those*, *today*, *we*
  - Nouns available as domain-specific knowledge: *physician*, *clinician*, *patient*
- **Lexical processing** (for each lexical instance):

  Set CL for the following:
  - Function words: modals, prepositions, etc. (specified as `cont=(func,`*FuncType*`)`)
  - Two-place nouns (specified by the feature `implicit_arg=req`) with no argument
  - Entries sharing the content information (specified as `cont=(cont,`*Class*`:`*Content*`)`) with a contextually-linked category (i.e., morphological variation)

  Set NL for numerals (specified by the category `num`)

  Designate the "project from itself" status for a denominal adjective with `denom=yes`
- **Composition** (two categories):
  - Special case: ignore dummy categories (e.g., punctuation) and function words as arguments; set CL for composition of two function words
  - Set CL status to the result of the following:
    - Definite determiner: for $X_{[def=yes]}/Y+Y$
    - Indefinite generic: for $X_{[def=no]}/Y+\underset{CL}{Y}$
    - Utterance-initial modifier: if the result is S/S
    - Inverted phrase: $S_{[inv=yes]}/X+X$
  - Set NL status to the result of the following:
    - Indefinite article: for $X_{[def=no]}/Y+Y$
    - Numeral: for $X/Y_{[cl=reset]}+Y$
  - Project the contextual-link status from an argument to the result:
    $\underset{Status}{X}\ldots/Y_{[cl=proj(arg)]}+\underset{Status}{Y}$ (also directional variations)
  - Project the contextual-link status from itself:
    $X\ldots/Y_{[cl=proj(self)]}+Y$ (also directional variations)
  - Project the contextual-link status either from an argument or from itself: utterance-initial modifier with a CL subject, i.e., $S/S+S$ where both S/S and the subject of S are CL
- **Coordination** (three categories): Project CL if both conjuncts are CL, set NL otherwise

Figure 6.3: A Summary of the Procedure to Identify Contextual-Link Status

Chapter 7

Evaluation of the Theory Using Parallel Texts

To overcome the problem with the previous implementations, this chapter develops an evaluation process that allows us to demonstrate that the proposed theory performs better than some alternative hypotheses underlying previous implementations. For practical reasons, the process shown here is an evaluation of the procedure corresponding to the proposed theory, not a direct evaluation of the theory. Nevertheless, we may call the process "evaluation of the theory" considering the fairly transparent nature of the implementation, as discussed in the previous chapter.

In this chapter, we first describe the data used for the experiment, and then develop an evaluation method that compares system's particle prediction with human translation. In the final section, we apply the evaluation method to reserved test data and present the results.

7.1 The Data

Our experimental data are taken from a journal, "The Physician and Sportsmedicine", downloaded from the journal web site "`http://www.physsportsmed.com/index.html`". We prepared two sets of texts: the **training data set** used for the development of the theory, system, and evaluation method and the **test data set** reserved for the evaluation of the theory. Some basic properties are shown in Table 7.1. We have already seen Text 12 as an example in earlier chapters.

Once the data sets are downloaded, they are manually processed in the following way. First,

	Training Data Set	Test Data Set
Source	Vol 25 - No. 9 - September 1997 to No. 12 - December 1997	Vol 26 - No. 12 - December 1998 to Vol 27 - No. 2 - February 1999
Number of texts	16 (Text 1 to 16)	8 (Text 17 to 24)
Number of utterances	131	66
Number of words	2314	1203

Table 7.1: Training and Test Data Set

utterances are segmented after each title and at each sentence boundary. Compound sentences are broken down into multiple utterances. In this case, the coordinator such as *and* and *but* are treated as discourse markers of the latter utterance(s). After this stage, utterances are identified as (*T*-*U*) where T/U correspond to the text/utterance IDs (the utterance ID starts at 1 for the title).

There are several places where additional adjustments have been made.

(217) *a*. In the coordinate structure of the form "*A, B, C*", an *and* is added before *C* to make it "*A, B, and C*".

"*Cheerleading Injuries: Patterns, Prevention,* □ *Case Reports*" (3-1)

b. A period after a non-sentence in a parenthetical is removed.

"*(See "The Years Surrounding Menopause: Practical Terms for a Complex Time," below* □.*)*" (17-3)

c. Several utterances are separated into two to avoid extremely long processing.

The main concerns in evaluating acute extremity injuries are to (1) determine the type and severity of injury (severe sprains, which may be difficult to differentiate from fractures, receive similar initial treatment),

↑ separated

(2) assess the distal neurologic and vascular status, (3) determine the need for radiographic imaging and specialty treatment, and (4) select appropriate splinting for immediate protection. (6-3) [also (4-4), (7-5), (12-5), and (19-3)]

d. A comma is replaced with an *or* to avoid excessive complication due to the ambiguity associated with the comma category.

"*but whether the activity is recreational or professional* □, *organized or spontaneous, the level of play makes little difference in the type or severity of foot injury.*" (19-6)

The number of utterances in Table 7.1 is the figure after these adjustments.[1]

[1]The data, instruction to the translators, translation, and an Excel file for analysis are all available through the

Next, we describe the case where some utterances are excluded from evaluation. There are three major classes of conditions for exclusion: (i) properties of text (language independent), (ii) properties of English, and (iii) properties associated with English-Japanese translation. In this section, we list the following exclusion cases corresponding to (i) (those corresponding to (ii) and (iii) are discussed in the next section).

(218) *a*. Title

b. Discourse marker

c. Citation

d. Direct quote

Titles are parsed, and semantic representations are derived and stored in the discourse context. For a title that has the NP type, the system does not analyze the information structure. For a title that has the sentence type, the system outputs an information structure, but we exclude it from evaluation. Discourse markers are automatically removed by the preprocessor as described in the previous chapter. Citations are manually removed from the data. One utterance entirely consisting of a direct quote (15-10) is also manually excluded because the situation within a direct quote is distinct from that of the text.

7.2 Development of an Evaluation Method Using the Training Data

The next step is the development of an evaluation method. The path for automatic particle prediction and that for human translation are separated, and the results are compared manually (see Fig. 1.3 on p. 11). This stage uses the training data, and the test data had been withheld from analysis.

The proposed theory is designed to identify the information structure of the entire utterance. But our current evaluation method concentrates on the theme/rheme status of the matrix-level grammatical subjects for the following reasons. First, in Japanese, the choice of particle for grammatical subjects is most crucial and most discussed, as we have seen in Chapter 5. Second, evaluation involving other components is possible but requires a project of much larger scale. At this point, it is more immediate to establish a methodology and obtain some results for a prominent case.

---

author's thesis web page at "`http://www.cis.upenn.edu/~komagata/thesis.html`".

This section starts with a review of particle prediction by the system, describes the process of collecting translations, and presents an evaluation method. We also discuss some difficult cases and possibility of extending the evaluation using components other than grammatical subjects.

7.2.1 Mechanical Prediction of Particle Choices in Japanese

The system's sample particle predictions for Text 12 are shown below. Here, grammatical subjects are in *italics* and materials excluded from analysis are enclosed in ⟨...⟩. We make a few remarks at the end of the data.

(219) *i.* (Title) ⟨Osteoporosis in Active Women: Prevention, Diagnosis, and Treatment⟩

ii. [*Osteoporosis* *wa*]$_{Theme}$ [has been defined as "a disease characterized by low bone mass and microarchitectural deterioration of bone tissue, leading to enhanced bone fragility and a consequent increase in fracture risk."]$_{Rheme}$

iii. [Although anyone can develop osteoporosis,]$_{Theme}$ [*postmenopausal women and young females with menstrual irregularities* *ga* are most commonly affected.]$_{Rheme}$

iv. [*An estimated 20% of women more than 50 years old* *ga* have]$_{Rheme}$ [osteoporosis.]$_{Theme}$ (see the note below)

v. [Although most studies have focused on women of this age-group, *osteoporosis* *wa*]$_{Theme}$ [is potentially more deleterious in younger women because they haven't yet attained peak bone mass, and early bone loss therefore can affect the rest of their lives.]$_{Rheme}$

vi. [Whether patients are younger or older, *the social costs of osteoporosis* *wa*]$_{Theme}$ [are enormous.]$_{Rheme}$

vii. [*The yearly estimated healthcare bill for osteoporotic fractures* *wa*]$_{Theme}$ [is between $2 billion and $6 billion.]$_{Rheme}$

viii. [*About 200,000 osteoporosis-related hip fractures* *ga* occur each year]$_{Rheme}$ [in the United States,]$_{Theme}$

ix. ⟨and⟩ [*the mortality rate 1 year after fracture* *wa*]$_{Theme}$ [is estimated to be as high as 20%.]$_{Rheme}$

The first remark is that in utterances (*v, vi*), the span of the theme includes the utterance-initial modifier and the subject of the main clause. These themes are identified due to the operational hypothesis (211) on p. 170, and are actually discontiguous. The process of derivation and information-structure analysis are shown below.

(220) *a*. [Whether patients are younger or older,]$_{CL_1}$ [*the social costs of osteoporosis*]$_{CL_2}$ [are enormous.]$_{NL}$

b. [Whether patients are younger or older,]$_{CL_1}$ [*the social costs of osteoporosis* are enormous.]$_{\langle CL_2, NL\rangle}$

c. [Whether patients are younger or older, *the social costs of osteoporosis* are enormous.]$_{\langle \underset{\downarrow}{CL_1}+\underset{\textit{Theme}}{CL_2},\ \underset{\textit{Rheme}}{NL}\ \rangle}$

Second, the information-structure analysis for (*iv*) appears incorrect. I.e., the verb *have* should belong to the theme because it cannot receive a pitch accent at the end of the rheme. The system includes *have* within the rheme for the following reason. This instance of *have* is analyzed as a main verb, not the auxiliary counterpart.[2] All main verbs are currently treated as content words. Thus, its contextual-link status depends on the discourse status. Since no occurrence of *have* (main verb) appears prior to this one, it is judged as a non-contextual-link. Although we leave the problem as is for now, this can be fixed by assigning the main verb *have* a status distinct from other main verbs. In this chapter, we focus on the information-structure status of grammatical subjects.

Although the system analyzes the information structure of every utterance (except for titles with the NP type), there are cases excluded from evaluation for reasons specific to English. The system is not designed to analyze the following type of constructions.

(221) *a*. Expletive: e.g., "*it's important to detect PCL injuries*" (10-3)

b. Correlative between clauses: e.g., "*Not only is it responsible for 200,000 deaths yearly, but in men over 40 it ranks second only to coronary heart disease as a cause of disability.*" (11-4)

c. Adverbial modification scoping over a clausal coordination: e.g., "*Among athletes, ankle sprains are the most common injury, and inversion injuries are frequent.*" (16-4)

[2]The auxiliary verb *have* and the *be* verb are analyzed as a function word.

We have not included an analysis of expletive, and thus the system cannot distinguish the expletive *it* from the pronoun *it*. The correlative in (*b*) combines two clauses but cannot be separated as a compound. The last case also involves clause coordination, which cannot be separated into two utterances.

While we could deal with these cases within the current framework, we leave them for future work because there are only a few instances of this kind.

7.2.2 Human Translation

Collecting Translations

To identify an appropriate data collection methodology, a preliminary experiment was conducted. It included the following three tasks.[3]

(222) *a*. English-to-Japanese translation of one text (translation of medical terms was provided)

b. After reading a text in English, the subject is asked to answer one question about the text (to make sure that the original text in English is read), and then asked to fill-in appropriate particles in the prepared translation in Japanese

c. Evaluation of instances of *wa* and *ga* in their own translation: indicate whether their choice could be replaced with the other particles

My initial expectation was to use a fill-in survey of the type (*b*) to obtain human judgment on particle choice because it is relatively easy and cost-effective. Unfortunately, it appears that the subjects are heavily influenced by the sentence constructions given in the translations, including word order. The third task, (*c*), of evaluating their own translation shows uncertainty of the subjects about 'judgment'. When they are asked to evaluate and consider the alternative, they tend to show a great tolerance to whichever choice. It seems unrealistic to expect translators to provide their intuition corresponding to what we expect for 'contextual appropriateness'. The conclusion is that the only remaining possibility is full translation, (*a*).

Four subjects are found through local and public newsgroups to translate the training data.[4] They are all native speakers of Japanese (two male and two female). Three of them have some

[3]The texts used in this preliminary experiment are taken from the same journal but not included in the training nor the test data.

[4]The newsgroups are: "`upenn.general`", "`upenn.nihon-club`", "`upenn.asian-student-union`", "`sci.lang.japan`", and "`fj.sci.lang`".

experience in translation, none of them is full-time professional translators. The following is the instruction given to them.

(223) *a*. The translation should contain all of the information in the original text in English.

b. The translation should correctly reflect the idea in the original.

c. The translation should be sentence-by-sentence as segmented for each text.

d. The translation should sound natural. After the translation is done, please read all the texts aloud and make necessary adjustments so that the translation sounds natural to the listener.

e. No artistic or rhetoric consideration should be made.

f. The translator can choose the level of politeness.

Recording Particle Choices

Translators' particle choices are recorded manually. First, all utterances are aligned with the output of the system. Then, for each utterance, we identify the phrase in Japanese that corresponds to the grammatical subject in the source utterance in English.

There are several cases where translation from English to Japanese introduces additional complications. At this point, the following cases (identified for each translator) are marked 'not available' for the evaluation.

(224) *a*. The subject in English corresponds to discontiguous parts in Japanese.

b. The subject in English corresponds to a phrase in Japanese that is not marked with either *wa* or *ga*.

c. The subject in English corresponds to an embedded phrase in Japanese.

d. The matrix-level predicate of the target subject in Japanese is negated.

e. The matrix-level predicate of the target subject in Japanese is a one-place, stage-level predicate.

The case (*a*) can be observed for a complex NP subject in English. For example, the modifying PP can be separated and preposed in the translation. There are a few possibilities for the case (*b*). The translators occasionally choose a construction distinct from the original argument structure

in English. For example, the subject in English may appear as the object (usually *o*-marked) or adjunct in Japanese. In some translations, the particle *mo* (*also* or *too*) is used for the target subject. In Section 5.4, we have discussed several special cases for *wa*/*ga* choice. The case (*c*) corresponds to one of them. But, if a phrase is extracted from the embedded clause, typically from a complement clause, it must be considered at the matrix level and the case (*c*) does not apply. The case (*d*) is another special case discussed in Section 5.4. Note that a positive construction in English, e.g., one involving *few*, may be translated into a negative one in Japanese. Finally, the case (*e*) is yet another special case.

For the remaining cases, we record the particle choices between *wa* and *ga*. As long as *wa*-marking is used, even if it appears as non-subject or after other case particle such as *ni* (dative), we count it as *wa*-marking (see Section 5.4). In addition, if the entire phrase corresponding to the English subject is *dropped*, it can be analyzed as a part of the theme and can be classified as *wa*-marking, because no rheme can be dropped.

This process of recording translators' particle choice is singly done by the author. Although there is a possibility of errors and variability, we assume that this process is reasonably accurate. In a sense, it is comparable to a task, in English, to identify a phrase in an utterance, corresponding to a particular semantics (e.g., given a phrase in French), and to check its definiteness from the determiner. It is difficult to automate this process because finding corresponding phrases in English and Japanese from semantic representations requires much more than simple unification.

A summary of translators' choice for Text 12 is shown in Table 7.2. The result appears consistent although there are cases where translators opt for constructions without *wa*/*ga* marking.

Utterance	Translator				*wa*/*ga* choice		
	N	A	F	I	*wa*	*ga*	n/a
(*ii*)	*wa*	*wa*	*wa*	*wa*	4	0	0
(*iii*)	*ga*	n/a	n/a	n/a	0	1	3
(*iv*)	*ga*	*ga*	*ga*	n/a	0	3	1
(*v*)	*wa*	wa_{drop}	*wa*	wa_{drop}	4	0	0
(*vi*)	*wa*	n/a	*wa*	n/a	2	0	2
(*vii*)	*wa*	*wa*	*wa*	*wa*	4	0	0
(*viii*)	*ga*	*ga*	*ga*	*ga*	0	4	0
(*ix*)	*wa*	*wa*	*wa*	*wa*	4	0	0

Table 7.2: Particle Choices by Human Translators (Text 12)

The distribution of *wa* and *ga* for all the texts in the training data is shown in Table 7.3. At first glance, this table may not appear very coherent. But we should note the following. The translators have a great degree of freedom. A choice between *wa* and *ga* surfaces as only one of the factors involved in the process. Thus, the case of 'n/a' must be considered as non-commitment to *wa*/*ga* choice, and the difference among translators about the degree of commitment for choosing either *wa* or *ga* is not a concern here.

Translator	*wa*	*ga*	n/a
N	89	15	5
A	79	10	20
F	85	4	20
I	57	14	38
	Total = 109		

Table 7.3: Particle Choices by Translators (Training Data)

The uneven distribution of *wa* and *ga* in the data (80 to 90% are *wa*) might lead one to think that *wa* is the default particle for the matrix-level subject and *ga* is a special case. We have already assumed the opposite position in Chapter 5. The predominance of *wa* in the matrix environment is a consequence of the tendency that matrix-level subjects are a part of the theme. Most of embedded subjects are marked with *ga*. The overall distribution including both matrix-level and embedded subjects is much more even, as shown in Chapter 5.

Agreement among Translators

In order to analyze the agreement among translators in a standard way, we use the κ statistic, following the procedure described in Siegel and Castellan [1988].[5] The κ statistic is developed for nominally-scaled data where no ranking or interval is observed among data categories. The process utilizes an agreement table like Table 7.2 as input and computes the level of agreement as a number between 0 (no agreement; corresponding to a chance distribution) and 1 (perfect agreement). It has also been found that for a large sample, the κ statistic distributes approximately normally. Therefore, it is possible to estimate the significance of a κ statistic in terms of, in our case, a *z* score. Since the κ statistic simply scales from chance to perfect agreement, comparing κ statistics

[5]The standard reference for the κ statistic is Cohen [1960], and the extension for multiple raters is due to Fleiss [1971].

for different cases without reference to variance is meaningless.

We compute a κ statistic for the binary choice between *wa* and *ga*, excluding 'n/a' cases. This is because the agreement among translators about not to use these particles is not our concern. But, then, we can only use the data where all translators choose either *wa* or *ga*.[6] For example, in Table 7.2, Utterances (*iii*), (*iv*), and (*vi*) are no longer available for the four-rater comparison.

First, the κ statistics and the *z* scores for the case of two-translator agreement is shown in Table 7.4. We observe that the agreement for the pair in boldface is significant $(p < .05)$,[7] but not for two other cases. Both of the two cases involve the translator F. Thus, it seems that F is not in agreement with the rest of the group. For this reason, the evaluation process requires multiple translators to obtain a representative sample of the population of native Japanese speakers.

Translator	N	A	F	I
N	–	**$\overset{\kappa}{0.59}$/$\overset{z}{2.69}$**	0.42/1.56	**0.46/2.31**
A	–	–	**0.46/1.65**	**0.39/1.71**
F	–	–	–	0.19/0.77
I	–	–	–	–

Table 7.4: Agreement between Two Translators (Training Data)

The κ statistics and the corresponding *z* scores for the agreement among all four translators on binary choice between *wa* and *ga* is 0.38 with $z = 1.98$. Thus, we can conclude that the agreement is significant $(p < .05)$. We now justify to use the set of translations as a reasonably coherent group for evaluation. Although choices between *wa* and *ga* by multiple subjects has been analyzed in narrative context [e.g., Clancy and Downing, 1987; Maynard, 1987], there have been few reports on particle choice agreement among translators. Thus, the present project also provides interesting data for further study.

7.2.3 Evaluation Methodology

We are now in a position to evaluate the machine-generated predictions in comparison to the human translations. For the evaluation purpose, we construct a set of **target** particle choices for a hypothetical translator from the translators' data in the following way:

[6] We still include the dropping case, though.

[7] For $\alpha = .05$, the cutoff point of the region of rejection is $z = 1.64$. For $\alpha = .01$, it is $z = 2.32$.

(225) *a*. Choose *wa* as the target if the number of translators who choose *wa* is (i) more than one and (ii) greater than those who choose *ga*

b. Choose *ga* as the target if the number of translators who choose *ga* is (i) more than one and is (ii) greater than those who choose *wa*

c. Otherwise, exclude the utterance from evaluation

This scheme is applicable to arbitrary number of translators. It excludes cases where only one translator chooses *wa*/*ga* and those where the choice is a tie. After this process, we have 82 instances (90%) of *wa* and 9 instances (10%) of *ga* as the target data.

For evaluation, we use the measure of recall/precision commonly used in information retrieval and other areas of computational linguistics. In our case, it is a measure of agreement between the target particle choices (hypothetical translator) and the predictions of the system (or other hypotheses). The definition is given as follows:

(226) *a*. **Recall** $= \dfrac{\text{number of } \textit{correctly-predicted} \text{ target data}}{\text{number of } \textit{total} \text{ target data}}$

b. **Precision** $= \dfrac{\text{number of } \textit{correctly}\text{-predicted data}}{\text{number of } \textit{total} \text{ predicted data}}$

Recall/precision is calculated for several alternative hypotheses, as shown in Table 7.5.

	wa (Target = 82)				*ga* (Target = 9)			
	Predicted		Recall	Precision	Predicted		Recall	Precision
Hypothesis	Correct	Total	(%)	(%)	Correct	Total	(%)	(%)
All *wa*	82	91	82/82= 100	82/91= 90	0	0	0/9 0	0/0= *n/a*
Chance (random)	74	82	90	90	1	9	11	11
Discourse status only	26	26	32	100	9	65	100	14
Definiteness only	40	40	49	100	9	51	100	18
Proposed	**73**	**73**	**89**	**100**	**9**	**18**	**100**	**50**

Table 7.5: Comparison of Hypotheses on the Training Data

The trivial hypothesis 'all *wa*' happens to exhibit a high recall and precision on *wa* due to the uneven distribution of *wa* and *ga*. It has nothing to say about the choice of *ga*. Even though the absolute number of errors is only 9 and the lowest among the hypotheses, there is no information about the distribution of *ga* and there is no room for improvement.

The chance case is calculated as follows. Since the probability of a *wa* occurrence is 90% for the training data, the number of *wa* predictions is 90% of the target number of *wa*. Thus, we expect 74 instances of correct predictions. The number of *ga* predictions is 10% of the target number of *ga*. Thus, only 1 instance of correct prediction is expected, which gives a very poor result.

For the hypothesis 'discourse status only', we assume that a process can predict particles for the matrix-level subject. The procedure would consider the discourse status of the subject. But we extend this slightly and assign particles for certain pronouns (e.g., *we* and *they*) and domain-specific nouns (e.g., *physician* and *patient*) because these can be asserted in the initial context and analyzed as discourse-old (as we do in our implementation). But we exclude any structural analysis from this hypothesis. This hypothesis misses too many instances of *wa*.

For the hypothesis 'definiteness only', the particle choice is applied only to the matrix-level subject based on its information-structure status. This hypothesis only utilizes structural information including definiteness on the subject. But pronouns and domain-specific nouns are also included because they can be lexically identified. The hypothesis fails to identify many instances of *wa* much like the previous one.

Although the proposed algorithm is far from perfect, it performs better than the other hypotheses. This is the only hypothesis that can predict both *wa* and *ga*-marking in a balanced way. The remaining problem for our hypothesis is that there still are a substantial number of incorrect predictions of *ga* instances. We will discuss this problem shortly.

For the reasons of coverage and specification, we cannot directly compare the above results with the previous computational approaches. For example, Hahn [1995] uses a partial parser, and has limitations in recognizing different types of themes. Hajičová et al. [1995] and Hoffman [1996] cannot deal with realistic texts like ours. While Hoffman [1996] mentions the possibility of processing INFERRABLE, no specification is provided. Therefore, we only point out that the 'discourse status only' and the 'definiteness only' hypotheses are underlying mechanisms for Hahn's [1995] and Hajičová et al.'s [1995], respectively. Hahn's algorithm may perform better than the 'discourse status only' hypothesis because it has a limited inference mechanism. Hoffman's [1996] algorithm combines properties underlying both of these hypothesis, and would be the closest to ours only if it is applicable to realistic data.

Let us examine one more property of the proposed theory. The κ statistic for the group of all four translators *and* the system's prediction is 0.33 with $z = 2.09$. This is a significant agreement $(p < .05)$, and inclusion of the predicted data even increases the z score (from $z = 1.98$ for 4 translators). Thus, from a statistical point of view too, we may say that the prediction is on the right track.

7.2.4 Analysis of Errors

The 'errors' found in the result of the training evaluation (9 of them) are all incorrect predictions of *ga* for the translators' choice of *wa*. They can be classified into the following two types:

(227) *a*. Indefinite inferrable in (2-3), (3-5), (5-4), (5-10), (9-6), (14-3) (6 instances)

b. Discourse-initial accommodation in (6-2), (9-2), (16-2) (3 instances)

Each type is discussed in the following.

Indefinite Inferrable

This is by far the predominant type of errors. The following example taken from (3-5) illustrates the case. The problematic subject is underline in the last utterance.

(228) *i.* Cheerleading Injuries: Patterns, Prevention, Case Reports

ii. Cheerleading began at the turn of the century when a University of Minnesota football fan stood in his seat and led the crowd in a verse in support of their team.

iii. From that humble beginning has blossomed a competitive athletic activity that includes nearly a million participants at the elementary, high school, college, and professional levels.

iv. Cheerleading competitions are held at regional and national levels,

v. and <u>training</u> is a year-round activity.

In the last utterance, the system predicts *ga*-marking because the grammatical subject is not discourse-old, not specified in the domain-specific knowledge, and without linguistic marking for contextual linking. But three translators choose *wa*-marking and only one chooses *ga*-marking. For human, it is most likely to infer the relation such as "*cheerleading* ***requires*** *training*". Thus, this can be considered an instance of indefinite inferrable. On the other hand, *training* inferred

from cheerleading is not as specific as the relation between "*the door*" and "*a house*" as seen in (51).

Other instances of grammatical subjects involving indefinite inferrable are listed below.

(229) *a*. "*A fiberglass cast with a waterproof liner that "breathes" "* inferrable from "*A Waterproof Cast Liner*" in the title [translators' choices between *wa*:*ga*:'n/a' is 3:0:1] (2-3)

b. "*Musculoskeletal weakness, stiffness, and pain*" inferrable from "*unwelcome changes*" in the preceding utterance [translators' choices 4:0:0] (5-4)

c. "*reduced capacity for exercise*" inferrable from "*decreased mobility*" in an earlier utterance [translators' choices 2:1:1] (5-10)

d. "*Many researchers*" inferrable from "*sports medicine*" in an earlier utterance [translators' choices 2:1:1] (9-6)

e. "*Exercise-related symptoms in the upper GI tract*" inferrable from "*Gastrointestinal Disorders*" in the title [translators' choices 4:0:0] (14-3)

These inferrables are all specific to the domain of discussion. Thus, we could capture the above inferrable cases within the domain-specific knowledge. But the use of domain-specific knowledge in our theory is to *bound* general inference. As soon as we include this type of inference within domain-specific knowledge, there is a danger of re-introducing general inference in our theory. Thus, at this point, we accept errors of this kind and leave the problem with inference as a whole for future work.

Discourse-Initial Accommodation

The second type of errors can be seen in the following example from (6-2):

(230) *i.* (title) Field Splinting of Suspected Fractures: Preparation, Assessment, and Application

ii. Initial on-site management of serious musculoskeletal injuries can pose a number of diagnostic and treatment challenges for the team physician.

No properties of our theory can be used to analyze the underlined subject as a part of the theme and thus *ga*-marking is predicted. The agreement among the translators is perfect (all 4 translators chose *wa*) for all three discourse-initial subjects that are predicted for *ga*. An obvious possibility is that even with the presence of the title, a discourse-initial matrix subject can be accommodated. In

addition, discourse-initial accommodation has a simple mechanical solution because its position can be identified with an extremely simple kind of discourse structure. But, since we exclude the discussion of discourse structure in general, we leave these errors as they appear.

7.2.5 Possibility of Extending the Evaluation

Let us next discuss the possibility of evaluating information-structure status of elements other than matrix-level subjects.

First, it is more difficult to use *wa*-marking for evaluation of the information-structure status on arguments other than subject. As we have discussed in Section 5.4, a thematic object may receive *wa*-marking only when the subject is not *wa*-marked and the object is 'fronted' (possibly including the vacuous case at the matrix level) or the object becomes a subject by passivization or use of an unaccusative verb. Considering the fact that 80-90% of subjects are *wa*-marked, there is little room for other elements to be fronted and get a *wa*. But, there is one example involving this case (7-4).

(231) *a*. (Translators A and I)

The original utterance in **English**: Predisposing factors can put [many active patients]$_{wa}$ at risk.

Their translation in **Japanese** (literally translated back into English): *Many active patients have risk due to predisposing factors.*

b. (System) [Predisposing factors can put many active patients]$_{Rheme}$ [at risk.]$_{Theme}$ (incorrect)

The system correctly analyzes that the original subject is a part of the rheme. But the analysis for the rest of the utterance is incorrect. The reason "*at risk*" is incorrectly analyzed as a theme is as follows. The noun *risk* is currently assigned as a two-place noun, i.e., as "risk of something" (see Section 3.3). Without an argument PP, it is assigned a contextual-link status. This status is projected through the preposition. At the same time, the system correctly identifies the contextual-link status of "*many active patients*" by projecting the domain-specific knowledge through adjective and non-definite determiner. There is a stage where the following three components are identified (*CL* and *NL* stand for contextual link and non-contextual link).

(232) [Predisposing factors can put]$_{NL}$ [many active patients]$_{CL}$ [at risk.]$_{CL}$

Due to the incorrect status on "*at risk*", the system fails to project the middle *CL* to the final structured meaning. If the last two CL's could combine into a single CL, "*many active patients at risk*", this case would result in a "*Rheme – Theme*" pattern where the combined CL is the theme. But, since "*at risk*" is only available as an argument of the verb, this possibility is rejected. The only remaining possibility is that the rightmost CL gives rise to the sole CL of the matrix clause.

There is another possibility: similar patterns of object-to-subject conversion may end up with *ga*-marking. The following example (3-6) demonstrates such a case. Note that the Japanese translation is literally translated back into English in all of the following examples.

(233) *a*. (Translators F and N)

English: Cheerleading routines can include [gymnastic elements, tumbling runs, partner stunts, pyramid formations, and dance routines.]$_{ga}$

Japanese: *Among cheerleading routines, there are gymnastic elements, tumbling runs, partner stunts, pyramid formations, and dance routines.*

b. (System) [Cheerleading routines]$_{Theme}$ [can include gymnastic elements, tumbling runs, partner stunts, pyramid formations, and dance routines.]$_{Rheme}$

The system's analysis is consistent with the *ga*-marking on the subject in Japanese (the original subject is *wa*-marked after postposition *ni* as an adverbial). There are several more examples of this kind. In addition, *ga*-marking on adjectival complements and that-complement are also observed and predicted as a part of the rheme.

An interesting case of *wa*-marking is found in the following example (10-7):

(234) *a*. (Translators N, A, and I)

English: With that in mind, the focus of [this paper]$_{wa}$ is on injury assessment and detection.

Japanese: *With that in mind, this paper places the focus on injury assessment and detection*.

b. (System) [With that in mind, the focus of this paper]$_{Theme}$ [is on injury assessment and detection.]$_{Rheme}$

In this case, only the complement of a preposition within the subject is extracted and *wa*-marked in Japanese. This is not inconsistent with the system's prediction, but excluded from the evaluation

because the subject NP in English does not appear as a constituent in Japanese. There are a few more examples of this type.

There is another case where even a verb in English is nominalized and *ga*-marked (10-4).

(235) *a*. (Translator F)

English: Though athletes can often function at a high level after an undiagnosed PCL injury, untreated injuries may [result]$_{ga}$ in disability years later.

Japanese: *Though ..., without treating injuries, the result of being disabled may occur years later.*

b. (System) [Though athletes can often function at a high level after an undiagnosed PCL injury, untreated injuries]$_{Theme}$ [may result in disability years later.]$_{Rheme}$

The system's analysis is again consistent with the *ga*-marking.

Although adverbials cannot be *ga*-marked, they can be *wa*-marked, as in the following example (7-5).

(236) *a*. (Translators A and I) [Especially in 18- to 40-year-olds,]$_{wa}$ these include close contact with a number of people (as in team travel or dormitory living), time of year, possible overtraining, and being debilitated from hectic schedules that leave little time for sleep.

b. (System) [Especially in 18- to 40-year-olds, these]$_{Theme}$ [include close contact with a number of people (as in team travel or dormitory living), time of year, possible overtraining, and being debilitated from hectic schedules that leave little time for sleep.]$_{Rheme}$

Several similar cases are observed. There is an example of *wa*-marking on an utterance-initial *if*-clause. These are consistent with our hypothesis that utterance-initial modifiers are a part of the theme.

The occurrence of these cases are limited and we could not collect a sufficient number in a small-scale evaluation like ours. But the above examples demonstrate that the proposed theory of information structure is not limited to grammatical subjects and the result could be evaluated with more data.

7.3 Evaluation of the Theory Using the Test Data

We now face the test data. Naturally, our expectation is that the properties observed for the training data generalize to the test data. This section describes the preparation, and then presents and discusses the results.

7.3.1 Extension of the System for the Test Data

First of all, we must be clear that our case of the evaluation on test data cannot be directly compared to tests commonly practiced by corpus-based approaches. In their case, systems are trained on millions of words and tested on another set of large data. Once a system is trained, it is used for testing without any modification. In our case, the system is designed for only 16 texts, and is being tested against another 8 texts. Since the lexical and grammatical coverage for 16 texts is no way general enough to cover another 8 texts, it is inevitable that the lexicon and grammatical features will need to be extended for the test data. Since information-structure-related specifications are also encoded in the lexicon, the way we extend the system *affects* the result of the evaluation. At this stage of developing and conducting an evaluation for an information-structure analyzer, this situation seems unavoidable. Nevertheless, we expect to demonstrate that the core of the theory and implementation with respect to information structure generalizes to a new data set.

Due to the complexity of contextual-link and structured-meaning analysis, the implementation for the training data is still underspecified in many respects. During the course of the extension, instantiation of such specifications becomes necessary. This demonstrates the system's capability to accommodate a new data set within the design criteria.

Extension of the system is mostly confined to a single file to delineate what is being *added*. The following is a summary of the extension.

(237) *a*. The test data contains 1203 words, an approximately 52% increase of the training data set with 2314 words.

b. The number of lexical entries (i.e., 'word' entries) increased by 291 (33%) from 883. Among the original, 56 are modified.

c. The number of lexical category assignments increased by 28 (15%) from 190. Among the original, 23 are modified.

d. The following are added to the initial context: *we*, *others*, *many* (as a pronoun)

e. The following is added to the composition of structured meaning:
"$\langle CL_1, - \rangle + \langle CL_2, NL_2 \rangle_{NL-CL} \Rightarrow \langle CL_1, NL' \rangle_{CL-NL}$" for the case where the following stronger condition fails "$\langle CL_1, - \rangle + \langle CL_2, NL_2 \rangle_{NL-CL} \Rightarrow \langle CL', NL_2 \rangle_{CL-NL}$"[8]

Since the data size increased by 52%, a change of 52% means no generalization while 0% change means perfect generalization. Naturally, a lexicon of this small size could not generalize to an additional data set. Many new words need to be added. Many of the changes to the existing lexical entries are due to additional subcategorizations that were not initially specified. There are cases where information-structure related features such as `implicit_arg=req` for two-place nouns and `denom=yes` for denominal adjectives (see Fig. 6.3 on page 180) are adjusted when these features were not initially specified.

Lexical category assignment shows some generalization (15%). Most of them are additional verb subcategorizations and modification frameworks for adverbs. The changes made to the existing lexical assignments are correction for syntactic/semantic reasons or specifications of contextual-link projection that was originally not given.

The basic grammatical framework stays. Most of the components related to the information-structure and contextual-link processing stay as in the original.

7.3.2 Results

For the test data, we gained two translators and have a total of six. The distribution of particle choice is shown in Table 7.6. The balance between *wa* and *ga* is slightly more even for this data set.

The κ statistics and the corresponding z scores for two-translator agreement is shown in Table 7.7. We observe that the agreement for the pairs in boldface is significant $(p < .05)$, but not for the other cases. In this case, translator I seems in least agreement with the rest of the group. Note that for the training data, F (not I) was in least agreement with the group. Thus, this situation again warns us about individual variation and requires us to use the data collectively.

Let us now turn to the level of agreement as a group. The κ statistic for all six translators on binary choices between *wa* and *ga* is 0.44 with $z = 2.25$ ($z = 1.98$ for the training data). Thus, we

[8]Here, NL' is a composition of CL_2 and NL_2, and CL' is a composition of CL_1 and CL_2.

Translator	*wa*	*ga*	n/a
N	45	8	4
A	35	10	12
F	39	5	13
I	24	11	22
K	38	9	10
U	37	9	11
	Total = 57		

Table 7.6: Particle Choices by Translators (Test Data)

Translator	N	A	F	I	K	U
N	–	**0.60$^{\kappa}$/2.50^{z}**	**0.48/1.67**	0.28/1.14	**0.55/2.36**	**0.44/1.83**
A	–	–	0.25/0.89	0.26/1.07	0.54/2.09	**0.48/1.93**
F	–	–	–	0.16/0.60	**0.47/1.66**	0.36/1.15
I	–	–	–	–	0.27/1.12	0.31/1.23
K	–	–	–	–	–	0.36/1.40
U	–	–	–	–	–	–

Table 7.7: Agreement between Two Translators (Test Data)

conclude that the agreement is significant $(p < .05)$, which justifies the use of the set of translations for evaluation as a group.

We adopt the same criterion (225) to set up the target particle choice. The result of the comparison among alternative hypotheses (same criteria) is shown in Table 7.8.

	wa (Target = 44)				*ga* (Target = 7)			
	Predicted		Recall	Precision	Predicted		Recall	Precision
Hypothesis	Correct	Total	(%)	(%)	Correct	Total	(%)	(%)
All *wa*	44	51	100	86	0	0	0	n/a
Chance	38	44	86	86	1	6	14	14
Discourse status only	14	14	32	100	7	37	100	19
Definiteness only	23	23	52	100	7	28	100	25
Proposed	**36**	**37**	**82**	**97**	**6**	**14**	**86**	**43**
Proposed (training)	–	–	89	100	–	–	100	50

Table 7.8: Comparison of Hypotheses on the Test Data

This resulting pattern in Table 7.8 parallels that in Table 7.5. The first two hypotheses cannot predict the occurrence of *ga*-marking. The hypotheses "discourse-status only" and "definiteness only" cannot collect a sufficient number of *wa*-markings. The proposed theory is again far from

perfect and the recall/precision figures are slightly worse than those for the training data. But they are substantially better than the other hypotheses compared in the table. The κ statistic for the group of all six translators *and* the machine prediction is 0.31 with $z = 1.84$. Thus, we conclude that the agreement still results in a significant level $(p < .05)$. From this, we can conclude that the proposed theory generalizes to a new data set reasonably well.

7.3.3 Discussion

Analysis of Errors

In the result, there is 1 error of incorrect prediction of *wa* and 8 errors of incorrect predictions of *ga*. The latter includes 4 cases of indefinite inferrables and 1 case of discourse-initial accommodation, and 2 more cases that may be classified both indefinite inferrable and discourse-initial accommodation. These cases are basically the same as we have discussed for the training data. In the following, we discuss two new types of errors (1 incorrect *wa* and 1 incorrect *ga* prediction) in detail. This is to explore even further development of the proposed theory, which has basically met our expectations.

The first (18-6) is the case of incorrect *wa* prediction. The problematic subject is underlined in the last utterance.

(238) *i.* Stress Urinary Incontinence in Women: Removing the Barriers to Exercise

ii. A growing number of women are exercising and thereby gaining benefits ranging from an improved sense of well-being to increased cardiovascular endurance, musculoskeletal strength, and mobility.

iii. But as more women have formed the exercise habit, more attention has been focused on complaints of stress urinary incontinence (SUI) during physical activity.

iv. The prevalence of SUI was suggested by a recent survey in which 28% of a group of nulliparous elite athletes reported experiencing the problem during exercise.

v. For women who are troubled by incontinence while working out, effective treatment may be essential to enable them to continue their regimen.

vi. Thus an understanding of SUI and the wide range of available treatments is important for fitness-oriented physicians.

All translators have chosen *ga*-marking. Let us first trace the system's analysis. It first detects the discourse-old status of *SUI* and the definiteness of "*the wide range of available treatments*". The coordination of these conjuncts thus results in a contextual link. This status is projected through the preposition *of*, to the N+PP combination. A composition of an indefinite article and a contextual link is, at this point, analyzed as a generic and set as a contextual link. This puts the subject as a part of the theme, and predicts *wa*. Since all the translators chose *ga*-marking for the *wa*-prediction of the system, we must suspect the system's prediction, i.e., our conjecture about indefinite generic (p. 68) in particular. This shows a benefit of a mechanical procedure for objective evaluation.

On the other hand, we may also investigate other possibilities. The problematic subject is a fairly complex NP. In this regard, it is different from the simple case of an indefinite generic discussed on page 68. We need finer conditions for analyzing indefinite generics.

Interestingly, we have a very similar use of indefinite in the following example (20-8).

(239) *i.* Overuse Injuries in Children and Adolescents

ii. The benefits of regular exercise are not limited to adults.

iii. Youth athletic programs provide opportunities to improve self-esteem, acquire leadership skills and self-discipline, and develop general fitness and motor skills.

iv. Peer socialization is another important, though sometimes overlooked, benefit.

v. Participation, however, is not without injury risk.

vi. While acute trauma and rare catastrophic injuries draw much attention, overuse injuries are increasingly common.

vii. Diagnostic and treatment efforts should focus on how the injury developed and consider issues that are unique to growing athletes.

viii. An understanding of these concepts provides the basis for making specific injury-prevention recommendations.

Naturally, the system does basically the same thing and predicts a *wa*. In this case, three translators have chosen *wa*, two *ga*, and one chose a different construction. According to our criterion (225), the target for this case is set as *wa*, and thus this case is evaluated as correct. One possible analysis is that the property of the predicate affects the information structure. For example, "*is important*" might set the subject as a rheme.

The other case of an error is the following (20-4).

(240) *i.* Overuse Injuries in Children and Adolescents

ii. The benefits of regular exercise are not limited to adults.

iii. Youth athletic programs provide opportunities to improve self-esteem, acquire leadership skills and self-discipline, and develop general fitness and motor skills.

iv. Peer socialization is another important, though sometimes overlooked, benefit.

The system predicts *ga*-marking. Three translators have chosen *wa*-marking and the other three used constructions where no *wa*/*ga* choice is available. Thus, the target is chosen as *wa*. Two translators have chosen *mo*-marking (*also* or *too*), which is natural considering the presence of *another* in the predicate.

Although I did not classify the subject "*peer socialization*" as an indefinite INFERRABLE, one may do so. In fact, the three translators who chose *wa*-marking are likely to have considered it that way. Our theory does not have a specification for the phrase "*another X*", but this phrase seems special in the following way. When we say "*another X*", it is likely that there is some *X* already in the context. In this regard, "*another X*" may well be an INFERRABLE. If "*peer socialization*" is BRAND-NEW and "*another X*" is INFERRABLE, the theory predicts "*Rheme – Theme*". If both components are INFERRABLEs, the prediction is ambiguous between "*Theme – Rheme*" and "*Rheme – Theme*". Thus, like other clearly inferrable cases, the present analysis faces the difficulty associated with INFERRABLES.

Applicability to a New Domain

The evaluation process shows that the lexicon and, to some extent, the grammar needs to be adjusted for a new data set in the same domain. The possibility of applying the present theory/system to information-structure analysis to a new domain is a natural question we need to address. But let us still limit ourselves to expository texts because most applications for expository texts today, e.g., reference resolution algorithms, are not automatically applicable to, say, spoken discourse.

The present theory of information structure specifically includes domain-specific knowledge as a component. Thus, this component must be adjusted for a new domain. For example, for the domain of financial news, the assumption for medical case reports is no longer applicable.

That is, physicians and patients are not in general situationally available. But it is likely that the other components, i.e., discourse status and linguistic marking of contextual links, remain as we analyzed. The evaluation method presented in this chapter is of course available for testing such a hypothesis.

7.4 Summary

We develop an evaluation method for the training data set and apply its extension to a test data set. The results demonstrate that the proposed theory performs better than other alternative hypothesis underlying previous implementations of information-structure analyzers, and that the results extend to a new data set. We thus conclude that the theory of information structure and its implementation exhibit a reasonable level of generality.

Chapter 8

Conclusion

Summary

In computational applications such as machine translation, speech generation, and writing assistance, the effect of information structure is critical for contextually appropriate processing of natural language. This thesis focuses on the problem of identifying information structure in expository texts.

But, as we review in Chapter 2, the existing analyses of information structure cannot directly be applied to the Identification Problem. They basically do not address the problem, and are not sufficiently explicit for the purposes of formalization and implementation either. The computational proposals directly responding to the problem are mostly not applied to realistic texts and do not provide an evaluation method.

Our response to this situation is to propose an explicit theory of information structure, formalize and implement it, and evaluate the result with respect to an independent observation. In Chapter 3, we develop a theory of information structure with the Identification Problem in mind. The main hypothesis is that information structure is a semantic composition between a theme and a rheme and the theme is necessarily contextually-linked. Following the Montagovian tradition, we analyze instances of semantic composition along the syntactic derivation. This way, the analysis of contextual links in an utterance can be used to identify a information structure of the utterance. The present approach captures the properties of contextual linking in terms of logic-external properties: discourse status, primitive domain-specific knowledge, and linguistic marking. Each of

these properties is precisely described.

For two potential problems with binomial partition of information structure, i.e., non-traditional constituency and discontiguous information structure, we adopt a flexible notion of constituency recognized by Combinatory Categorial Grammar (CCG) and an additional degree of freedom gained by structured meanings compositionally built for CCG constituents as semantic representations.

To establish the connection between the proposed theory and a practical implementation, we formally describe the theory, including the specification of contextual links and structured meanings, within an extended form of the CCG framework (Chapter 4). We also show that variants of CCGs are comparable to the related formalisms with respect to generative capacity and theoretical parsing efficiency.

For the evaluation purpose, we take advantage of the particle choice problem in English-Japanese translation. Chapter 5 provides the basis for this approach by investigating the Japanese particle *wa* and other case markers, and the function of long-distance fronting in detail. After identifying several exceptional cases, we analyze that *wa* and *ga* at the matrix level mark (a part of) theme and rheme, respectively.

The next step is to provide a procedure to identify information structure. In Chapter 6, we first show the practicality of our CCG parser, and then implement the specification of contextual linking and information structure. There are certain procedural aspects associated with our information-structure analysis. These are introduced in a modular fashion, and can be considered reasonable through the examination of the experiment (training) data. As the last step of the mechanical procedure involved in the current project, we apply the analysis of Japanese and predict particle choices for matrix subjects based on the identified information structure.

Finally, the crucial element of this thesis is the evaluation of the theory (Chapter 7). The methodology is to compare the particle predictions made by the system and human translations. We first develop our evaluation method using the training data, and then show that the theory generalizes to previously-withheld data. This demonstrates that the proposed theory is an improvement over the alternative hypotheses underlying the existing computational approaches, and that the proposed theory generalizes to new data.

Contributions

The main contribution of this thesis is a demonstration, including an evaluation on test materials withheld from the development set, that information structure can be correctly interpreted and used in practical applications such as machine translation for limited domains. This development advances the state where the notion of information structure has rarely escaped the intuition of some researchers.

The first crucial step in this demonstration is to squarely face the Identification Problem. Like other computational approaches to the Identification Problem, but, unlike most theoretical work in linguistics, the current proposal can directly connect the result of the project to practical applications.

The present work is distinguished from other computational approaches in that the results are evaluated based on an independently-observable phenomenon. As a consequence, the readers can judge for themselves whether or not the main point of the thesis (10) holds. The same does not apply to the previous computational approaches simply because they do not provide an evaluation procedure. Their results often appear arbitrary, and cannot really be judged for this reason. The presented evaluation method is limited to matrix subject positions, and the accuracy is still not very high. But it can be extended to a wider range of utterance components as shown in Chapter 7, and other languages can be used for the same purpose. Thus, we can increase the coverage and the accuracy of the evaluation beyond what is presented here.

The thesis also covers a wider range of linguistic constructions, including various real-text properties, than previous work. Although the lexicon and the grammar still need to be extended, the information-structure analysis can be applied to a new set of realistic texts for further evaluation with little adjustment in terms of the theory of information structure. Thus, we have overcome Levinson's [1983] skepticism about the applicability of information-structure analysis for an arbitrarily complex linguistic structure.

There is one other factor associated with the main contribution. That is, the theory is made sufficiently explicit so that it is readily formalized and implemented as a procedure. This development contrasts with the situation where most theoretical works in linguistics are at a level that does not easily allow formalization and implementation. It also contrasts with most computational approaches, which lack the connection between their procedure and linguistic theories.

In addition to the above, the thesis contributes several points to the field of computational linguistics. By adopting a grammar-based parser, albeit one that is rather flexible in terms of dealing with constituency, the implementation of the theory retains the ability of precisely capturing various syntactic and semantic properties, and can integrate pragmatic factors in a straightforward manner. This provides a precise connection between utterance-level linguistic description and certain discourse-level concepts.

As a backbone of the system, we developed a practical parser for the CCG framework, overcoming the potential problem of spurious ambiguity. This point should remove the skepticism surrounding parsing CCG.

The thesis develops a comprehensive formalization and implementation of structured meanings. This not only captures the informational contrast present at every step of derivation, but also provides a platform for other properties including 'contrast' in a more general way than existing applications of structured meaning.

Finally, we provide an analysis of Japanese from the view point of a modern information-structure analysis. The functions of Japanese particles and long-distance fronting have been under discussion for a long time. Unfortunately, even the current literature does not fully reflect the recent advancement in studies of information structure and referential status. The current work updates this situation and provides materials useful for language-specific and cross-linguistic analyses. In addition, through the discussion on both English and Japanese in terms of information structure and contextual linking, we are able to relate certain underlying mechanisms of various pragmatic functions.

Future Directions

One natural continuation of the present work is to integrate the information-structure analyzer with the applications discussed in the Introduction. For example, in most machine translation projects, a parser is already built in. While not all types of parsers can recognize constituents as flexibly as CCG parsers can, we may still use the derived linguistic structure and identify information structure based on the present approach. Then, the results can be used for prediction of particles in Japanese and word order in, e.g., Turkish.

Another application that I have a great interest is a Computer-Assisted Writing system, which can analyze text readability with respect to information structure. During the development of the present thesis, we seriously considered this project as an application domain and proposed a prototype (Section 2.4). A preliminary result on analyzing journal abstracts gives an impression that this application would make a noble, useful tool for writers. But the idea was not pursued for the present thesis because of the difficulty with evaluation. But I still consider this as an interesting long-term project.

The evaluation method proposed in the present work concentrates on the information-structure status of grammatical subjects. We may extend this to components other than subjects as briefly touched on in Chapter 7. We may also use other languages that marks information structure differently from the way it is done in Japanese. Since the linguistic marking of information structure in a single language by no means covers all the constructions, a multi-lingual analysis seems to be required for a more complete coverage.

Another direction is to use larger-scale parallel corpora available on the Internet. We have seen that the current accuracy of the prediction is at a level comparable to the individual variation (for unconstrained translation). Thus, using a larger number of texts written by different individuals may yield similar results without obtaining multiple translations.

As we mentioned in Section 2.1, there is a related problem of identifying definiteness in English encountered in an application such as Japanese-English machine translation. Our position is that the definiteness-identification problem is distinct from the Identification Problem for information structure. But there is a great deal of overlap. Both problems contain basically the same components: definiteness marking, contextual linking, and information structure. It is interesting to see how much the present theory can tell about the relation between the two, both shared and distinct elements.

The present thesis separates important areas of reference, inference, and discourse structure. Further exploration about the connection between information structure and these areas is a challenging but exciting future work.

Finally, the analysis of the present work may also apply to second-language education both in English and Japanese. A student of Japanese may learn certain concrete information about the use of particles, which is often perceived difficult or vague. A student of English may learn

the functions of various constructions in terms of a fairly small number of properties including contextual linking.

Appendix A

Generative Power and Parsing Efficiency of CCG-GTRC

In Section 4.1.4, we briefly touched on generative power and parsing efficiency for CCG involving Generalized Type-Raised Categories (CCG-GTRC). This Section explores these properties in detail based on two technical reports [Komagata, 1997d; Komagata, 1997b] (with minor revision on the notation). The main points are that a restricted version of CCG-GTRC is equivalent to the standard CCG, and that CCG-GTRC is polynomially parsable theoretically and practically. The results have also been presented as Komagata [1997c] and Komagata [1997a].

This section is organized as follows. Subsection A.1 motivates and introduces the formal framework of CCG-GTRC. Subsection A.2 proves the equivalence of CCG-GTRC and standard CCG under specific conditions. Subsections A.3 and A.4 discuss theoretical and practical results, respectively, on polynomial parsing for CCG-GTRC.

A.1 CCG with Generalized Type-Raised Categories

Motivation: Unbounded NP Sequence

In languages including Japanese, a NP sequence can form a constituent for coordination and extraction as seen in Section 4.1.4. A similar type of constituent can also be formed of NPs extracted from different levels of embedding, as in the following example:

(1) Japanese: Rinyouzai-wa natoriumu-ni, β syadanzai-wa koukan sinkei kei-ni,
kankei-no aru kouketuatu-no hito-ni kikimasu.
Gloss: {Diuretic-TOP sodium-DAT} & {β blocker-TOP sympathetic nervous system-DAT}
relevance-GEN exist hypertension-GEN person-DAT effective.
Translation: "Diuretic is effective for the person with hypertension related to sodium, and β blocker
[is for the person with hypertension related] to sympathetic nervous system."

The underlined part is another instance of non-traditional constituent, which includes an extraction from the relative clause. Its structure is schematically shown as follows:[1]

(2) [t_1 hypertension$_2$-GEN person-DAT effective.]
[t_2 t_3 relevance-GEN exist]

As we have seen in (130) on page 102 (Subsection 4.1.4), NP sequence in Japanese can form a category of the form $S/((S\backslash NP)\backslash NP)$. Assuming that the competence grammar does not place a bound on the levels of embedding [Miller and Chomsky, 1963], we may have unboundedly-many extractions [Becker et al., 1991; Rambow and Joshi, 1994; Rambow, 1994]. Since no systematic constraint has been identified for the bound on the composition of such extracted constituents, we also assume that these constituents can compose without a limit, potentially resulting in an unboundedly-long NP sequence. As in the case of embedding, the degraded acceptability of long sequences can be attributed to performance issues. These assumptions calls for an infinite set of type-raised categories such as $(S\backslash X_n...\backslash X_1)/((S\backslash X_n...\backslash X_1)\backslash NP)$ associated with NP. We capture this polymorphic situation by using variables as in $T/(T\backslash NP)$.

The formal properties of the standard CCGs not involving variable (CCG-Std) are relatively well-studied (see Section 4.1.5). But the use of variables can destroy these properties. For example, Hoffman [1993] showed that a grammar involving categories of the form $(T\backslash x)/(T\backslash y)$ can generate a language $\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n\mathtt{d}^n\mathtt{e}^n$, which no mildly context-sensitive grammar can generate. The use of variables in the coordination schema "x^+ **conj** $x \Rightarrow x$" is also believed to generate a language $(\mathtt{wc})^n$ beyond LIG's power [Weir, 1988]. At a level higher in the scale, Becker et al. [1991], Rambow and Joshi [1994], and Hoffman [1995] propose formalisms that are more powerful than the standard CCG to account for 'doubly'-unbounded scrambling. 'Doubly'-unbounded scrambling has the following properties: (i) there is no bound on the distance of scrambling and (ii) there is no

[1]The use of trace t_i is for illustration purposes only. The current approach does not assume the notion of gap or movement as the theories which employ trace.

bound on the number of unbounded dependencies in one sentence. As we know that full context-sensitive capacity is too powerful to be a formal model of natural language syntax [e.g., Savitch, 1987], it is essential to identify the generative power of the formalism that interests us.

Component: Generalized Type-Raised Categories

CCG-GTRC involves the class of constant categories (Const) and the class of Generalized Type-Raised Categories (GTRC).

A constant (derivable) category c can always be represented as $F|a_n...|a_1$ where F is an atomic **target** category and a_i's with their directionality are **arguments**. We use '$A,...,Z$' for atomic, constant categories, '$a,...,z$' for possibly complex, constant categories, and '$|$' as a meta-variable for directional slashes $\{/,\backslash\}$. Categories are in the 'result-leftmost' representation and associate left. Thus, we usually write $F|a_n...|a_1$ for $(...(F|a_n)...|a_1)$. We call '$|a_i...|a_j$' a **sequence** (of arguments). The length of a sequence is defined as $\left||a_i...|a_1\right| = i$ while the null sequence is defined to have the length 0. Thus, an atomic constant category is considered a category with the target category with the null sequence. We may also use the term 'sequence' to represent an ordered set of categories such as '$c_1,...,c_2$' but these two uses can be distinguished by the context. The standard CCGs (CCG-Std) solely utilize the class of Const.

GTRC is a generalization of **Lexical Type-Raised Category** (LTRC). A LTRC has the form $\overset{T}{\mathsf{T}}\left\langle \left(\mathsf{T}\right\rangle a\right)|b_i...|b_1$ associated with a lexical category $a|b_i...|b_1$ where $\overset{T}{\mathsf{T}}$ is a variable over categories with the atomic target category T. The target indication may be dropped when it is not crucial or all the atomic categories are allowed for the target. We assume the order-preserving form of LTRC using the following notation. '$\left\langle$' and '$\left\rangle$' indicate that either set of slashes in the upper or the lower tier can be chosen but a mixture such as '$/$' and '$/$' is prohibited [see Steedman, 1991b for a related discussion]. **GTRC** is defined as having the form of $\mathsf{T}\left\langle(\mathsf{T}|\underbrace{a_m...|a_2}\right\rangle a_1)\quad\underbrace{|b_n...|b_1}$

inner sequence outer sequence

resulting from compositions of LTRCs where $m \geq 1$, $n \geq 0$, and the directional constraint is carried over from the involved LTRCs. When the directionality is not critical, we may simply write a GTRC as $\mathsf{T}|(\mathsf{T}|a_m...|a_2|a_1)|b_n...|b_1$. For $gtrc = \mathsf{T}|\left(\mathsf{T}|a_m...|a_1\right)|b_n...|b_1$, we define $|gtrc| = n+1$, ignoring the underspecified valency of the variable. Note that the introduction of LTRCs in the lexicon is non-recursive and thus does not suffer from the problem of the overgeneration discussed

by Carpenter [1991].

These categories can be combined by combinatory rule schemata. Rules of (forward) "generalized functional composition" have the following form:[2]

(3) $x/\underline{y} \qquad \underline{y}|z_k...|z_1 \qquad \Longrightarrow \qquad x|z_k...|z_1 \qquad (>B^k)$

functor category input category result category

The integer 'k' in this schema is bounded by k_{max} specific to the grammar, as in CCG-Std.[3] Rules of functional application, "$x/y \quad y \Rightarrow x$", can be considered a special case of (3) where the sequence z_i's is null. We say "the combination of '$x/\underline{y}$' and '$\underline{y}|z_k...|z_1$' *derives* $x|z_k...|z_1$", and "$x|z_k...|z_1$ *generates* the string of nonterminals '$x/y, y|z_k...|z_1$' or the string of terminals '`ab`"' where the terminals `a` and `b` are associated with x/y and $y|z_k...|z_1$, respectively. The case with backward rules is analogous.

The use of variable for polymorphic type drew attention of researchers working on Lambek calculus [Moortgat, 1988; Emms, 1993]. In particular, Emms showed decidability for an extension called Polymorphic Lambek Calculus. The use of variables in the current formulation is limited to type raising. This reflects the intuition about the choice of rules based on 'combinators' [Steedman, 1988]. But, otherwise, we do not assume that categories are wildly polymorphic.

One way to represent this situation is to use two distinct subclasses of the type 'category' constructed as follows:

(4)

	Type construction	Example	
a.	*const* (Target, Arguments)	$F\backslash a_n...\backslash a_1$	$\mapsto const\,(F, \backslash a_n...\backslash a_1)$
b.	*gtrc* (Target, IDir, ISeq, OSeq)	$\overset{T}{\mathsf{T}}/(\mathsf{T}\backslash a_m...\backslash a_1)\backslash b_n...\backslash b_1$	$\mapsto gtrc\,(T, /, \backslash a_m...\backslash a_1, \backslash b_n...\backslash b_1)$

Such type construction can be defined in ML as follows:

(5)
```
datatype target A | B | C ...                (* atomic categories *)
datatype dir left | right
datatype complex_cat = Complex of target * arg
and arg = Arg of dir * complex_cat           (* mutually recursive *)
datatype seq = Seq of arg list
datatype cat = Const of target * seq
```

[2] Vijay-Shanker and Weir [1994] call the functor and input categories as 'primary' and 'secondary' components, respectively.

[3] Weir [1988] comments that the categorial grammars defined by Friedman and Venkatesan [1986] is more powerful than CCGs due to no bound on k.

```
| GTRC of target * dir * seq * seq
```

Then, we can define the combinatory rules on instantiated categories. Theoretically, no unification of variable is required although our implementation based on the proposed formalism uses variable unification for convenience. Although dealing with a greater number of cases is tedious, the technique is straightforward. This leads to a favorable result that CCG-GTRC is not only decidable but also polynomially recognizable.

Composition Involving GTRCs

Inclusion of GTRCs calls for a thorough examination of each combinatory case depending on the involved category classes. All the possible combination of category classes are described below. Some cases are subdivided furthermore. Although the traditional categorial representation is used below, the complete description for the constructor format can be defined. A summary of the cases is given in Table A.1. In the following, a combination of two constant categories is written as Const+Const. Note that all of the following cases are written for '$>B^k$' and the other direction is analogous.

(6) Const+Const: $a/\underline{b} \quad \underline{c}|d_k...|d_1 \implies a|d_k...|d_1$

(7) GTRC+Const

a. Functor GTRC has an outer sequence:

$\mathsf{T}|(\mathsf{T}|a_m...|a_1)|b_n...|b_2/\underline{b_1} \quad \underline{c}|d_k...|d_1 \implies \mathsf{T}|(\mathsf{T}|a_m...|a_1)|b_n...|b_2|d_k...|d_1$

Example: $\mathsf{T}\backslash(\mathsf{T}/PP)/\underline{NP} \quad \underline{NP} \implies \mathsf{T}\backslash(\mathsf{T}/PP)$

b. Functor GTRC has no outer sequence:

$\mathsf{T}/\underline{(\mathsf{T}|a_m...|a_2\backslash a_1)} \quad \overset{\overset{c_0|c_m...|c_1}{\underline{\quad\quad}}}{c} \quad |d_k...|d_1 \implies c_0|d_k...|d_1$

Example: $\mathsf{T}/\underline{(\mathsf{T}\backslash NP\backslash NP)} \quad \underline{S\backslash NP\backslash NP} \implies S$

(8) Const+GTRC

a. $k < |\text{input}|$:

$a/\underline{b} \quad \underline{\mathsf{T}|(\mathsf{T}|c_m...|c_1)|d_n...|d_{k+1}}|d_k...|d_1 \implies a|d_k...|d_1$

Example: $(S/(S\backslash NP\backslash NP))\backslash(S/(S\backslash NP\backslash NP))/\underline{(S/(S\backslash NP\backslash NP))} \quad \underline{\mathsf{T}/(\mathsf{T}\backslash NP\backslash NP)}$

$\rightarrow (S/(S\backslash NP\backslash NP))\backslash(S/(S\backslash NP\backslash NP))$

b. $k = |\text{input}|$ (and $k \geq 1$):

$a/\underline{b} \quad \underline{\mathsf{T}}|(\mathsf{T}|c_m...|c_1)|d_{k-1}...|d_1 \implies a|(b|\overbrace{c_m...|c_1}^{\text{unbounded}})|d_{k-1}...|d_1$

Example: $S/\underline{S} \quad \underline{\mathsf{T}}/(\mathsf{T}\backslash NP\backslash NP) \implies S/(S\backslash NP\backslash NP)$

c. $k > |\text{input}|$ (and $k \geq 2$):

$a/\underline{b} \quad \underline{\mathsf{T}_0}|\mathsf{T}_1|(\mathsf{T}_0|\mathsf{T}_1|c_m...|c_1)|d_{k-2}...|d_1 \implies a|\underset{\text{residual}}{\mathsf{T}_1}|(b|\mathsf{T}_1|\overbrace{c_m...|c_1}^{\text{unbounded}})|d_{k-2}...|d_1$[4]

(9) GTRC+GTRC

a. Functor GTRC has an outer sequence *and* $|\text{input}| > k$:

$\mathsf{T}|(\mathsf{T}|a_m...|a_1)|b_n...|b_2/\underline{b_1} \quad \underline{\mathsf{U}|(\mathsf{U}|c_p...|c_1)|d_n...|d_{k+1}}|d_k...|d_1$

$\implies \mathsf{T}|(\mathsf{T}|a_m...|a_1)|b_n...|b_2|d_k...|d_1$

b. Functor GTRC has an outer sequence *and* $|\text{input}| = k$ (and $k \geq 1$):

$\mathsf{T}|(\mathsf{T}|a_m...|a_1)|b_n...|b_2/\underline{b_1} \quad \underline{\mathsf{U}}|(\mathsf{U}|c_p...|c_1)|d_{k-1}...|d_1$

$\implies \mathsf{T}|(\mathsf{T}|a_m...|a_1)|b_n...|b_2|(b_1|\overbrace{c_p...|c_1}^{\text{unbounded}})|d_{k-1}...|d_1$

c. Functor GTRC has an outer sequence *and* $|\text{input}| < k$ (and $k \geq 2$):

$\mathsf{T}|(\mathsf{T}|a_m...|a_1)|b_n...|b_2/\underline{b_1} \quad \underline{\mathsf{U}_0}|\mathsf{U}_1|(\mathsf{U}_0|\mathsf{U}_1|c_p...|c_1)|d_{k-2}...|d_1$

$\implies \mathsf{T}|(\mathsf{T}|a_m...|a_1)|b_n...|b_2|\underset{\text{residual}}{\mathsf{U}_1}|(b_1|\mathsf{U}_1|\overbrace{c_p...|c_1}^{\text{unbounded}})|d_{k-2}...|d_1$

d. The functor GTRC has no outer sequence *and* $|\text{input}| > k$:

(i) T spans greater than U ($\mathsf{T} = \mathsf{U}|(\mathsf{U}|c_p...|c_1)|d_n...|d_{k+m+1}$):[5]

$\mathsf{T}/\underline{(\mathsf{T}|a_m...|a_2\backslash a_1)} \quad \underbrace{\mathsf{U}|(\mathsf{U}|c_p...|c_1)|d_n...|d_{k+m+1}|\overbrace{d_{k+m}...|d_{k+1}}^{|a_m...\backslash a_1}}_{\mathsf{T}}|d_k...|d_1$

$\implies \mathsf{U}|\overbrace{(\mathsf{U}|c_p...|c_1)}^{\text{inner seq of GTRC}}|d_n...|d_{k+m+1}|d_k...|d_1$

Example: $\mathsf{T}/\underline{(\mathsf{T}\backslash NP)} \quad \underline{\mathsf{U}/(\mathsf{U}\backslash PP)\backslash NP} \implies \mathsf{U}/(\mathsf{U}\backslash PP)$

[4] T could also be decomposed into $\mathsf{T}_0|\mathsf{T}_k...|\mathsf{T}_1$ for a larger k but all of them share the same characteristics with the above scheme.

[5] Here, the most general unifier is considered.

Case		Functor cat		Input cat		Result cat		
		Class	Outer seq	Class	$\|\text{input}\| \lessgtr k$	Class	Residual variable	Unbounded const argument
	6	Const	-	Const	$\geq$	Const	no	no
	7*a*	GTRC	yes	Const	$\geq$	GTRC	no	no
	7*b*	GTRC	no	Const	$\geq$	Const	no	no
	8*a*	Const	-	GTRC	$>$	Const	no	no
	8*b*	Const	-	GTRC	$=$	Const	no	possible
*	8*c*	Const	-	GTRC	$<$	neither	yes	possible
	9*a*	GTRC	yes	GTRC	$>$	GTRC	no	no
	9*b*	GTRC	yes	GTRC	$=$	GTRC	no	possible
*	9*c*	GTRC	yes	GTRC	$<$	neither	yes	possible
	9*d*i	GTRC	no	GTRC	$>$	GTRC	no	no
	9*d*ii	GTRC	no	GTRC	$>$	Const	no	no
	9*e*	GTRC	no	GTRC	$=$	GTRC	no	no
*	9*f*	GTRC	no	GTRC	$<$	neither	yes	possible

Table A.1: Combinatory Cases for CCG-GTRC

(ii) T spans no greater than U ($\mathsf{T}|a_m...|a_{m-j+1} = \mathsf{U}$):

$$\underbrace{\mathsf{T}/(\mathsf{T}|a_m...|\overbrace{a_{m-j}...}^{a_{m-j,0}|a_{m-j,p}...|a_{m-j,1}}|a_2\backslash a_1)}\quad \underbrace{\underbrace{\mathsf{U}_0}_{\mathsf{T}}|\underbrace{\mathsf{U}_j...|\mathsf{U}_1|(\mathsf{U}_0|\mathsf{U}_j...|\mathsf{U}_1|c_p...|c_1)}_{|a_m...\quad a_{m-j}=F|a_{(m-j,q)}...|a_{(m-j,1)}}|\underbrace{d_n...|d_{k+1}}_{...\backslash a_1}}|d_k...|d_1$$

$$\Longrightarrow\ F|a_{(m-j,q)}...|a_{(m-j,q-j-p)}|d_k...|d_1 \qquad \text{where } q \geq j+p$$

Example: $\mathsf{T}/\underline{(\mathsf{T}\backslash(S/NP))}\quad \underline{\mathsf{U}\backslash(\mathsf{U}/NP)} \Longrightarrow S$

e. The functor GTRC has no outer sequence *and* $|\text{input}| = k$ (and $k \geq 1$):

$$\mathsf{T}/\underline{(\mathsf{T}|a_m...|a_2\backslash a_1)}\quad \underline{\mathsf{U}}|(\mathsf{U}|c_p...|c_1)|d_{k-1}...|d_1$$

$$\Longrightarrow\ \mathsf{T}|(\mathsf{T}|\overbrace{a_m...|a_2\backslash a_1|c_p...|c_1}^{\text{inner seq of GTRC}})|d_{k-1}...|d_1$$

Example: $\mathsf{T}/\underline{(\mathsf{T}\backslash NP)}\quad \underline{\mathsf{U}}/(\mathsf{U}\backslash NP) \Longrightarrow \mathsf{T}/(\mathsf{T}\backslash NP\backslash NP)$

f. The functor GTRC has no outer sequence *and* $|\text{input}| < k$ (and $k \geq 2$):

$$\mathsf{T}/\underline{(\mathsf{T}|a_m...|a_2\backslash a_1)}\quad \underline{\mathsf{U}_0}|\mathsf{U}_1|(\mathsf{U}_0|\mathsf{U}_1|c_p...|c_1)|d_{k-2}...|d_1$$

$$\Longrightarrow\ \mathsf{T}|\underbrace{\mathsf{U}_1}_{\text{residual}}|(\mathsf{T}|a_m...|a_2\backslash a_1|\mathsf{U}_1|\overbrace{c_p...|c_1}^{\text{unbounded}})|d_{k-2}...|d_1$$

The three cases indicated by '$*$' in Table A.1 introduce categories that are neither Const nor GTRC due to the residual variables. This is an unintended, accidental use of functional composition. The closure of the system must be maintained by excluding these cases by the following condition:

(10) **Closure Condition**: The rule "$x/\underline{y} \quad \underline{y}|z_k...|z_1 \implies x/|z_k...|z_1$" (and analogously for the other direction) must satisfy $\left|y|z_k...|z_1\right| \geq k$.

Note that the distinction between constant categories and GTRCs must be made. This condition is particularly important for implementation since the residual variables can behave beyond our imagination and the parser must be able to compute the length of a category distinctively for constant categories and GTRCs.

Framework: CCG-GTRC

We define the most general form of CCG-GTRC$_0$ as follows:

Definition 1 A CCG-GTRC$_0$ is a five tuple (V_N, V_T, S, f, R) where

- V_N is a finite set of nonterminals (atomic categories)
- V_T is a finite set of terminals (lexical items, written as $\mathtt{a}, ..., \mathtt{z}$)
- S is a distinguished member of V_N
- $\mathcal{T}$ is a countable set of variables[6]
- f is a function that maps elements of V_T to finite subsets of "Const $\cup$ LTRC"[7]
- R is a finite set of rule instances of Generalized Functional Composition observing Closure Condition (i.e., those summarized in Table A.1 except for those with '$*$').

CCG-GTRC$_0$ differs from CCG-Std in some crucial respects.

(11) *a*. The set of arguments is not bounded. Not only the inner sequence of GTRC is unbounded, but also an argument of a constant category can be unboundedly long.

[6]Each instance of GTRC must be assigned a new variable when the GTRC is instantiated at a particular string position in order to avoid unintended variable binding.

[7]Our definition does not include the empty string in the domain of f as in [Vijay-Shanker and Weir, 1993] but unlike [Vijay-Shanker and Weir, 1994].

b. Combinatory rules cannot be specified in a 'finite' manner as described in [Vijay-Shanker and Weir, 1994].[8] The reason is that both functor and input categories can be unboundedly long unlike CCG-Std.

From both complexity and parsing points of view, this situation seems to require more 'power' to deal with. The conjecture is that this grammar is not equivalent to CCG-Std nor polynomially parsable. What I will do in the following is to find a subclass of CCG-GTRC$_0$ that still satisfies the original motivation and can be proved weakly-equivalent to CCG-Std. We discuss the following three problems in turn: (i) the bound of the arguments of constant categories, (ii) mixed directionality in GTRC inner sequence, and (iii) the behavior of GTRC outer sequences.

First, we want to apply the same techniques of CCG-Std to the "Const+Const" case. For this purpose, the set of arguments must be bounded [Vijay-Shanker and Weir, 1994; Vijay-Shanker and Weir, 1990]. Thus, we place a bound on the length of an argument.

(12) **Bounded Argument Condition**: Every argument except for the inner sequence of GTRC must be bounded by the grammar.

Then, the rules indicated as 'unbounded argument' in Table A.1 must be restricted to those satisfying the condition while the inner sequence of GTRCs can grow without limit. We now have the following property:

(13) The set of arguments of a constant category and the set of arguments of the inner and outer sequences of a GTRC are all finite. We denote the set of all these arguments as Args.

An alternative to the Bounded Argument Condition is to place a bound on the length of GTRC inner sequence. But, then, we need to re-evaluate our assumption about the unbounded NP sequence and the system degenerates to CCG-Std since every instance of GTRC can be represented as a constant.

The new subclass of CCG-GTRC$_0$ is defined as follows:

Definition 2 CCG-GTRC$_{bound_arg}$ is a subclass of CCG-GTRC where the Bounded Argument Condition is observed.

The second problem is with the mixed directionality of the GTRC. For example, consider a GTRC $\mathsf{T}/(\mathsf{T}/a_m.../a_2\backslash a_1)$ derived from "$\mathsf{T}/(\mathsf{T}\backslash a_1) \quad \Big(\mathsf{T}\backslash(\mathsf{T}/a_2) \quad ... \quad \mathsf{T}\backslash(\mathsf{T}/a_m)\Big)$". This

[8] This 'finiteness' corresponds to the instantiation of the input categories. The functor category of a combinatory rule still needs a meta-variable since categories can grow without limit.

may proceed with the following derivation: "$\mathsf{T}/\underline{(\mathsf{T}/a_m.../a_2\backslash a_1)}$ $\underline{c/a_m.../a_2\backslash a_1}|d_k...|d_1$". Although the input category, $c/a_m.../a_2\backslash a_1|d_k...|d_1$, seeks the arguments $a_2,...,a_m$ to its right, the arguments are actually found on the left of the category. In addition, although the GTRC $\mathsf{T}\backslash(\mathsf{T}/a_m)$ stands adjacent to $c/a_m.../a_2\backslash a_1|d_k...|d_1$, a_m is unboundedly-deep in the category $c/a_m.../a_2\backslash a_1|d_k...|d_1$. In a sense, this difficulty corresponds to the mixture of non-order preserving type raising and the unbounded version of generalized functional composition so that $\mathsf{T}/(\mathsf{T}/a_m)$ can combine with $s|d_k...|d_1/a_m.../a_2\backslash a_1$ (i.e., no limit on k_{max}). The current position is to stipulate the following condition:

(14) **Unidirectional GTRC Condition**: The inner sequence of a GTRC must have the uniform directionality as in: $\mathsf{T}\Big\langle\Big(\mathsf{T}\Big\rangle a_m...\Big\rangle a_1\Big)|b_n...|b_1$.

This condition is closely related to the linguistic aspect of long-distance 'movement' across the functor. Our motivation does not depends on these phenomena. For example, the gapping conjuncts of two underlined NPs in the English sentence, "John helped Mary, Bill, Rose." might involve
$S/(S\backslash NP/NP)$ from "$S/(S\backslash NP)$ $(S\backslash NP)\backslash((S\backslash NP)/NP)$". But I believe that such a case is inherently bounded and does not require a GTRC involving variables. We define the following subclass:

Definition 3 CCG-GTRC$_{uni}$ is a subclass of CCG-GTRC$_{bound_arg}$ where the Unidirectional GTRC Condition is observed. The third problem is related to 'quasi-island' condition exemplified as follows:

(15) *a*. CCG-GTRC: $S/A \quad \Big(\mathsf{T}\backslash(\mathsf{T}/A)/B \quad B\Big) \Longrightarrow S$

b. CCG-GTRC: $B \quad S/A \quad \overset{S}{\mathsf{T}}\backslash(\mathsf{T}/A)\backslash B \Longrightarrow *$

c. CCG-Std: $S/A \quad \Big(A/B \quad B\Big) \Longrightarrow S$

d. CCG-Std: $B \quad (S/A \quad A\backslash B) \Longrightarrow S$

With respect to the interaction with input categories of constant class, GTRCs behave like an island. But we do not have a general way in CCG-Std proper to *exactly* capture the effect. Our next step is to exclude outer sequence from the GTRCs altogether.

Definition 4 CCG-GTRC$_{no_outer}$ is a subclass of CCG-GTRC$_{uni}$ where no GTRC has outer sequence.

This limits the instances of GTRCs to a finite set since the inner argument of a GTRC is 'frozen'. It can only act as its own.[9] Although the expressiveness is greatly limited, it can still represent the example we started with in addition to the coverage of CCG-Std.

These conditions may appear unnatural. But note that they are applied when the grammar is constructed and do not change the way the grammar is used to recognize a string in CCG-GTRC. Thus, they are legitimate way to 'define' subclasses of grammar. In the rest of this paper, we focus on CCG-GTRC$_{no_outer}$ and prove its weak equivalence to CCG-Std. The only relevant cases are now (6), (7b), (8a, b), and (9$d$ii, e) where no outer sequence of GTRC is present.

A.2 Weak Equivalence of CCG-GTRC and CCG-Std

This section presents the proof of the equivalence of CCG-Std and CCG-GTRC$_{no_outer}$ (CCG-GTRC hereafter). Let $\mathcal{G}_{std}$ and $\mathcal{G}_{gtrc}$ be the classes of CCG-Std and CCG-GTRC, respectively. A grammar is represented by G_{index} where the subscript is optionally used to distinguish grammars. The proposition to prove is the following:

Proposition 1 $\mathcal{G}_{gtrc}$ is weakly equivalent to $\mathcal{G}_{std}$.

Since any $G \in \mathcal{G}_{std}$ is also $G \in \mathcal{G}_{gtrc}$ by definition, we only need to show that for each $G_{gtrc} \in \mathcal{G}_{gtrc}$, there is a $G_{std} \in \mathcal{G}_{std}$ such that G_{gtrc} and G_{std} generate the same language, i.e., $L(G_{gtrc}) = L(G_{std})$. The proof is by the following lemma with the start category set to S.

Lemma 1 (Main Lemma) For any $G_{gtrc} \in \mathcal{G}_{gtrc}$, there is a $G_{std-sim} \in \mathcal{G}_{std}$ such that a terminal string `w` is generated by a constant category c in G_{gtrc} iff `w` is generated by c' in $G_{std-sim}$ where c' is the category in $G_{std-sim}$ corresponding to c in G_{gtrc}.

We construct $G_{std-sim}$ from G_{gtrc} so that $G_{std-sim}$ *simulates* G_{gtrc}.[10] The process starts by translating G_{gtrc} to the base CCG-Std, $G_{std-base}$ as follows:

[9] The case (9$d$ii) may result in decomposition of the inner argument in a restricted way. This will be treated as Bounded GTRC in a later section.

[10] The word 'simulation' is also used to describe operations involved in the process.

(16) *a*. Copy all the constant categories in G_{gtrc} to $G_{std-base}$ assigned to the same terminal symbol.

b. For each LTRC $\mathsf{T}\left\langle \left(\mathsf{T}\right\rangle a\right)$ in G_{gtrc}, add an atomic category $\langle a \rangle$ to the lexicon of $G_{std-base}$ assigned to the same terminal symbol.[11] Note that the use of an atomic category is to avoid decomposition of the inner argument. This is possible since the inner arguments of GTRC never unifies with a target category and are never decomposed in the current formulation.[12]

Then, $G_{std-base}$ is extended to $G_{std-sim}$ to simulate G_{gtrc}. This situation is shown schematically as follows:

(17)

	translation		
G_{gtrc}	$\longrightarrow$	$G_{std-base}$	
	$\nwarrow$	$\downarrow$	extension
simulation		$G_{std-sim}$	

Since CCG-GTRC extends the way CCG captures phenomena including unbounded, but restricted 'permutation', it is crucial to identify the properties of GTRCs and provide appropriate methods for simulation. Once we have the right simulation, the equivalence can be shown by the set inclusion for both directions by invoking the simulation as needed. Two simulation techniques, 'wrapping' and 'bounded GTRC', and the proof of both directions will be described in the following.

Wrapping

CCG-GTRC allows permutation, as observed in the following example:

[11] The directionality of LTRC can be captured by features such as '$-left$' or '$-right$'. We will ignore this aspect for simplicity.

[12] This depends on the 'no-outer sequence' condition.

(18) *a.*

a	b	c
$\mathsf{T}/(\mathsf{T}\backslash A)$	$\underline{\mathsf{T}/(\mathsf{T}\backslash B)}$	$\underline{S\backslash A\backslash B}$
	$S\backslash A$	
	S	

b.

b	a	c	(permutation)
$\mathsf{T}/(\mathsf{T}\backslash B)$	$\underline{\mathsf{T}/(\mathsf{T}\backslash A)}$	$\underline{S\backslash A\backslash B}$	
	$S\backslash B$		
	S		

First, we attempt to simulate such a permutation by *wrapping* the arguments of a lexical category [the idea has been around for a while, e.g., Bach, 1979; Dowty, 1979]. For example, '$\backslash A$' in $S\backslash A\backslash B$ can wrap across '$\backslash B$' with the result of $S\backslash B\backslash\langle A\rangle$. We use '$\langle\rangle$' to represent the wrapped argument as an atomic category that will unify with the GTRC-translated category also represented in the same way. This corresponds to the permutation of (18*b*) as follows:

(19) *a.*

b	a	c	(CCG-GTRC)
$\mathsf{T}/(\mathsf{T}\backslash B)$	$\underline{\mathsf{T}/(\mathsf{T}\backslash A)}$	$\underline{S\backslash A\backslash B}$	
	$S\backslash B$		
	S		

b.

b	a	c	(CCG-Std)
B	$\langle A\rangle$	$S\backslash B\backslash\langle A\rangle$	
	$S\backslash B$		
	S		

The above-mentioned technique of wrapping arguments only applies to local permutation within a lexical category. But CCG-GTRC allows permutations across lexical categories, as seen below.

(20)

$\mathsf{T}/(\mathsf{T}\backslash A)$	$\underline{T\backslash B}$	$\underline{S\backslash A\backslash T}$
	$S\backslash A\backslash B$	
	$S\backslash B$	

Since we assume that GTRCs can compose without limit, there is no bound on the composition of the input to GTRCs.

(21) $$\dfrac{\dfrac{\mathsf{T}/\underline{(\mathsf{T}\backslash A_n...\backslash A_2)} \quad \underline{T}\backslash A_1 \quad ... \quad \underline{\dfrac{T\backslash A_{n-1}\backslash T \quad S\backslash A_n\backslash \underline{T}}{S\backslash A_n\backslash A_{n-1}\backslash \underline{T}}}}{\underline{S\backslash A_n...\backslash A_2}\backslash A_1}}{S\backslash A_1}$$

Then, we want to obtain a wrapped category like $S\backslash A_1\backslash\langle A_n\rangle...\backslash\langle A_2\rangle$. This situation can be captured by using the technique of *argument passing* as follows:[13]

(22) *a.* $S\backslash B\backslash\langle A\rangle \Longleftarrow \underline{T^{\{\backslash B\}}} \quad S\backslash B\backslash\langle A\rangle\backslash\underline{T^{\{\backslash B\}}}$

Note: Since subscripts are frequently used for indexing the categories in this paper, the features are placed as superscript.

b. $S\backslash A_1\backslash\langle A_n\rangle...\backslash\langle A_2\rangle \Longleftarrow$
$$\left(T^{\{\backslash A_1\}} \quad \left(T^{\{\backslash A_1\}}\backslash\langle A_2\rangle\backslash T^{\{\backslash A_1\}} \quad +s \quad \left(\underline{T^{\{\backslash A_1\}}}\backslash\langle A_{n-1}\rangle\backslash T^{\{\backslash A_1\}} \quad S\backslash A_1\backslash\langle A_n\rangle\backslash\underline{T^{\{\backslash A_1\}}}\right)\right)\right)$$

The arguments that are crossed by wrapping are placed as a feature on the target category and on the first argument. They are then passed on to the category corresponding to a deeper position of the composed category. As in the case of '$\langle\rangle$', we consider the category with passed arguments as an atomic category. This also applies for the case where the canceled category is complex such as: $(S/A)^{\{/C\}}/B$.

This simulation depends on the fact that the list of passed arguments is bounded. First, observe (a) below. A particular argument can be crossed by any number of arguments by wrapping, which is the source of unbounded permutation. On the other hand, an argument can cross only a finite number of other arguments by wrapping, as seen in (b). This latter case is bounded by k_{max} of functional composition.

(23) *a.*
$$\begin{array}{ccccccccc} B & \mathsf{T}/(\mathsf{T}\backslash A_n) & +s & \mathsf{T}/(\mathsf{T}\backslash A_1) & S & \backslash A_n...\backslash A_1 & |B & \\ & & & & & & | & \\ B & \langle A_n\rangle & +s & \langle A_1\rangle & S & & |B & \backslash\langle A_n\rangle...\backslash\langle A_1\rangle \end{array}$$

[13] Argument passing is conceptually similar to the techniques found in grammar formalism and logic including: SLASH feature of GPSG/HPSG [Gazdar et al., 1985; Pollard and Sag, 1994] and assume/discharge of natural deduction [Hepple, 1990]. But it is finite and limited in its power.

$b.\quad \mathsf{T}/(\mathsf{T}\backslash A) \qquad S \quad \backslash A \quad |B_k...|B_1$

$\qquad\qquad\qquad\qquad\qquad\qquad\quad |$

$\qquad\quad \langle A\rangle \qquad\quad S \qquad\quad |B_k...|B_1 \quad \backslash\langle A\rangle$

Recall that the set of arguments Args is bounded. Thus, at any juncture of rule application, there are only finitely many possibility of argument passing. We add all these cases to the lexicon.

To describe wrapping concisely, we introduce the following notation: Depending on how we divide a category into the 'function' and the 'arguments', a category $c = F|a_m...|a_1$ can be viewed with different valencies, i.e., $c = \underset{F}{f_m}|a_m...|a_1$,..., $c = \underset{F|a_m...|a_{i+1}}{f_i}|a_i...|a_1, c = \underset{F|a_m...|a_2}{f_1}|a_1, c = \underset{F|a_m...|a_1}{f_0}$. Let us refer to f_i as the *functional forms* of c. The functional forms with every valency can then be represented as follows: $c = F|a_m...|a_1 = f_i\mathbb{A}_i$ where $0 \leq i \leq m$ and $\mathbb{A}_i = |a_i...|a_1$.

The process of wrapping is now presented as follows:

(24) **Wrapping**: Consider functional forms of a lexical category $c = F|a_m...|a_1 = f_i[\mathbb{A}_i|a_1]$ where $\mathbb{A}_i = |a_i...|a_2$ and '[ ]' indicates the optionality. In case a_1 is not null, consider all the possible sequence of arguments $\mathbb{I} = |d_k...|d_1$ (as passed arguments) where $|\mathbb{I}| \leq k_{max}$. For a concatenation of $\mathbb{A}_i\mathbb{I}$ (including $|\mathbb{I}| = 0$), apply all the possible wrapping. Note the use of $\langle a_i\rangle$ to represent the wrapped argument a_i. Optionally, designate the last $j \leq k_{max}$ arguments as $\mathbb{O}$, and place them as the feature on f_i. The process can be abbreviated as follows: $f_i\mathbb{A}_i|a_1 \twoheadrightarrow f_i^{\{\mathbb{O}\}}\mathbb{A}_i'|a_1^{\{\mathbb{I}\}}$ where A_i' is obtained by wrapping as described above and the both categories are assigned to the same terminal. Categories with passed arguments are considered atomic categories.

Categories including a wrapped argument and/or a passed argument, do not interact with constant category until these features are canceled. For example, the following unintended cases all fail.

(25) *a.* $\underline{D} \quad \underline{S\backslash B\backslash\langle D\rangle} \implies *$

b. $C/\underline{(S\backslash B\backslash A)} \quad \underline{S\backslash C\backslash\langle A\rangle} \implies *$

c. $\underline{S}\backslash B \quad \underline{S\backslash B\backslash A\backslash S^{\{\backslash B\}}} \implies *$

The use of '⟨⟩' avoids overgeneration of the following kind as well:

(26) *a.* $S/C/\underline{A} \quad \underline{\left(A/B \quad B\right)} \implies S/C$ (potential overgeneration)

b. $S/C/\underline{\langle A\rangle} \quad \underline{\left(\langle A/B\rangle \quad B\right)} \implies *$ (implemented)

Bounded GTRC

When GTRCs appear as input category, their instances are bounded, as shown in (13). Thus, we can replace the variables with constants. For example, suppose that coordination is lexical, defined for each instance of conjunct category, and the set of conjuncts is bounded. Coordination of non-traditional constituents might need the conjunctive category like $\left(S/\left(S\backslash NP\backslash NP\right)\right)/\left(S/\left(S\backslash NP\backslash NP\right)\right)\backslash\left(S/\left(S\backslash NP\backslash NP\right)\right)$. Then, we can derive $S/\left(S\backslash NP\backslash NP\right)$ as "$S/\left(S\backslash NP\right)\quad\left(S\backslash NP\right)/\left(S\backslash NP\backslash NP\right)$". Both of the instances must be added to the lexicon since $G_{std-sim}$ has no other way to represent this non-traditional constituency. Since we are motivated to deal with unboundedly-long inner sequence of GTRC, we cannot apply this technique to (9*e*). Wrapping has been introduced for this purpose. The procedure of adding GTRC instances is described as follows:

(27) **Bounded GTRC**:

(8*a*): Suppose that the whole GTRC $\mathsf{T}\langle\backslash\left(\mathsf{T}\rangle a_m\ldots\rangle a_1\right)$ unifies with some argument of a category, i.e., a member of the set of arguments Arg in G_{gtrc}. The GTRC must be derived uniquely from a sequence of LTRCs, $\mathsf{T}_m/\left(\mathsf{T}_m\backslash a_m\right),\ldots,\mathsf{T}_1/\left(\mathsf{T}_1\backslash a_1\right)$ or $\mathsf{T}_1\backslash\left(\mathsf{T}_1/a_1\right),\ldots,\mathsf{T}_m\backslash\left(\mathsf{T}_m/a_m\right)$, depending on the directionality (cf. **Lemma 3**).[14] Add the ground instances of the LTRCs to the lexicon of $G_{std-sim}$.

(8*b*): Since we have set a bound on the instances on $\left(b\rangle c_m\ldots\rangle c_1\right)$, add the LTRCs that derives $b\langle\left(b\rangle c_m\ldots\rangle c_1\right)$.

(9*d*ii): The only possibility is the following:

"$\mathsf{T}/\underline{\left(\mathsf{T}\backslash a\right)}\quad\underbrace{\underbrace{\mathsf{U}}_{\mathsf{T}}|\underbrace{\left(\mathsf{U}|c_p\ldots|c_1\right)}_{a=F|a_m\ldots|a_1}}\rightarrow F|a_m\ldots|a_{m-p}|d_k\ldots|d_1$". The functor category must be an LTRC and the instances of a is bounded. We add those instances in the lexicon.

Proof: $L\left(G_{gtrc}\right)\subseteq L\left(G_{std-sim}\right)$

Now the simulation is established for the given CCG-GTRC. The proof of the direction from G_{gtrc} to $G_{std-sim}$ is by induction on the height h of a derivation in G_{gtrc}. The primary recursion (for both directions) deals only with constant categories (of CCG-GTRC) since we are concerned with

[14] This can be proved by induction on the length of the inner sequence.

derivations of a constant category, S in particular. The current direction also involves GTRCs as the source derivation and these are handled by **Lemma 3** and wrapping handled by **Lemma 4** introduced below. The latter lemma sets a mutually-recursive situation with this direction of the main lemma (**Lemma 2**).

Lemma 2 The direction $L(G_{gtrc}) \subseteq L(G_{std-sim})$ of the Main Lemma.

Base case ($h = 0$): c is a lexical category. Then, c is also in $G_{std-sim}$ assigned to the same terminal symbol.

Induction hypothesis (IH2): The lemma holds for all $h' \leq h-1$.

Induction step ($h \geq 1$): We only consider the following relevant cases where the result category is Const.

(28) *a*. (Const+Const, 6) The same derivation is available in $G_{std-sim}$. For the left and right categories, which are constant categories of smaller height, apply the induction hypothesis (IH2). The pair of strings obtained by the application of the induction hypothesis in the same order can be concatenated to provide the desired string in $G_{std-sim}$.

b. (GTRC+Const, 7*b*) $\mathsf{T}/\underline{(\mathsf{T}\backslash a_m ... \backslash a_1)} \quad \underbrace{c}_{\underline{c_0|c_m...|c_1}} \quad |d_k...|d_1 \implies c_0|d_k...|d_1$

This case requires the simulation. Note that c is unbounded. **Lemma 4** provides us the wrapped form $c_0|d_k...|d_1\backslash a_m...\backslash a_1$ from $c_0\backslash a_m...\backslash a_1|d_k...|d_1$. **Lemma 3** shows that there is a sequence of categories with the corresponding string that can combine with $c_0|d_k...|d_1\backslash a_m...\backslash a_1$ in the same order with the same string. Thus, after applying each category of the sequence to the wrapped category, we have the desired result $c_0|d_k...|d_1$ with the same string.

c. (Const+GTRC, 8*a*) $a/\underline{b} \quad \underline{\mathsf{T}|(\mathsf{T}|c_m...|c_1)} \implies a$

Since the GTRC is bounded, we have the corresponding category in $G_{std-sim}$ by (27). The rest is similar to the previous case.

d. (Const+GTRC, 8*b*) $a/\underline{b} \quad \underline{\mathsf{T}}|(\mathsf{T}|c_m...|c_1) \implies a|(b|c_m...|c_1)$

Since the GTRC is bounded by the stipulated Bounded Argument Condition, we have the corresponding category in $G_{std-sim}$ by (27). The rest is similar to (a).

e. (GTRC+GTRC, 9*d*ii) $\mathsf{T}/(\mathsf{T}\backslash \underbrace{a}_{\underbrace{a_0|a_p...|a_1}}) \quad \underline{\mathsf{U}|(\mathsf{U}|c_p...|c_1)} \implies a_0$

Since a is bounded, the process is similar to the previous case.

■

Lemma 3 If the derivation of $\mathsf{T}/(\mathsf{T}\backslash a_m...\backslash a_1)$ from the string $\mathtt{w}$ can combine with $x\backslash a_m...\backslash a_1$ in G_{gtrc}, $\langle a_m\rangle,...,\langle a_1\rangle$ that is associated with the same terminal string can combines with $x\backslash y_m...\backslash y_1$ for some x in $G_{std-sim}$ where y_i may be a_i or $\langle a_i\rangle$.

Proof: By induction on the height h of derivation.[15]

Base case ($h=0$): The category must be an LTRC, $\mathsf{T}/(\mathsf{T}\backslash a)$. Thus, there is $\langle a\rangle$ and a assigned to the same terminal in $G_{std-sim}$ by the simulation. Then, either $\langle a\rangle$ or a can combine with $x\backslash a$ or $x\backslash\langle a\rangle$ as desired.

Induction hypothesis: The lemma holds for $h' \leq h-1$.

Induction step ($h \geq 1$): The GTRC $\mathsf{T}/(\mathsf{T}\backslash a_m...\backslash a_1)$ must be derived as "$\mathsf{T}/\underline{(\mathsf{T}\backslash a_m...\backslash a_{i+1})} \quad \underline{\mathsf{U}}/(\mathsf{U}\backslash a_i...\backslash a_1) \rightarrow \mathsf{T}/(\mathsf{T}\backslash a_m...\backslash a_1)$" for some i (9*e*). Apply the induction hypothesis to the input category. Then, we have a sequence of $\langle a_i\rangle,...,\langle a_1\rangle$, which generates the same string as $\mathsf{U}/(\mathsf{U}\backslash a_i...\backslash a_1)$. Since each of $\langle a_i\rangle$ has the corresponding a_i, the sequence can apply to $x'\backslash y_i...\backslash y_1$ in series to derive some $x' = x\backslash y_m...\backslash y_{i+1}$ in $G_{std-sim}$. Next, apply the induction hypothesis to the functor category and $x\backslash y_m...\backslash y_{i+1}$ to obtain x in $G_{std-sim}$ from the same string as desired.

■

Lemma 4 Consider a category $c|a_m...|a_1|d_k...|d_1$ derivable in G_{gtrc} where $k \leq k_{max}$ and $m \geq 0$. If this category combines with a GTRC $\mathsf{T}/(\mathsf{T}\backslash a_m...\backslash a_1)$ to reduce to "$\mathsf{T}/(\mathsf{T}\backslash a_m...\backslash a_1) \quad c\backslash a_m...\backslash a_1|d_k...|d_1 \implies c|d_k...|d_1$", then there is a category $c|d_k...|d_1\backslash y_m...\backslash y_1$ in $G_{std-sim}$ where y_i is either a_i or $\langle a_i\rangle$, which generates the same terminal string as $c\backslash a_m...\backslash a_1|d_k...|d_1$ in G_{gtrc}.

The proof is by the following lemma that is a more general version.

[15] Induction on the length of the inner sequence also works for this case.

Lemma 5 Consider a category $x := c\backslash a_m ... \backslash a_1 | d_{k-1} ... | d_1 | e$ in G_{gtrc} where $k \geq 1$, $k \leq k_{max}$, $m \geq 0$, '$|e$' may be nil. c and e may be associated with passed arguments as a feature. If "$\mathsf{T}/(\mathsf{T}\backslash a_m ... \backslash a_1)$ $\quad c\backslash a_m ... \backslash a_1 | d_{k-1} ... | d_1 | e \implies c | d_{k-1} ... | d_1 | e$", there is a category $y := c^{\{\mathbb{O}\}} | b_j ... | b_1 \backslash \langle a_m \rangle ... \backslash \langle a_1 \rangle | e^{\{\mathbb{I}\}}$ in $G_{std-sim}$ where $j \leq k_{max}$, $m \geq 0$, $\mathbb{O}$ and $\mathbb{I}$ are sequences of arguments shorter than k_{max} (possibly nil) such that y derives the same terminal string as x.

Proof: By induction on the height h of derivation.

Base case ($h = 0$): $c\backslash a_m ... \backslash a_1 | d_{k-1} ... | d_1 | e$ is a lexical category. By (24), there is $c | d_{k-1} ... | d_1 \backslash \langle a_m \rangle ... \backslash \langle a_1 \rangle | e$ in $G_{std-sim}$ which is associated with the same terminal.

Induction hypothesis: The lemma holds for $h' \leq h - 1$.

Induction step ($h \geq 1$): Consider the following cases:

(29) *a*. Reduction: $c\backslash a_m ... \backslash a_1 | d_{k-1} ... | d_1 | e \Longleftarrow c\backslash a_m ... \backslash a_1 | d_{k-1} ... | d_{p+1} / f \quad f | d_p ... | d_1 | e$

By induction hypothesis, $c\backslash a_m ... \backslash a_1 | d_{k-1} ... | d_{p+1} / f$ has the corresponding $c | d_{k-1} ... | d_{p+1} | d_p ... | d_1 \backslash \langle a_m \rangle ... \backslash \langle a_1 \rangle / f^{\{|d_p ... |d_1\}}$ which generates the same string, and $f | d_p ... | d_1 / e$ has the corresponding $f^{\{|d_p ... |d_1\}} | e$ which generates the same string. This case may involve a GTRC as the input category ($p = 0$). But such a case is limited to a bounded form. We can thus consider the bounded instances as if they are constants.

b. Reduction: $c\backslash a_m ... \backslash a_1 | d_{k-1} ... | d_1 | e \Longleftarrow c\backslash a_m ... \backslash a_{i+1} / f \quad f \backslash a_i ... \backslash a_1 | d_{k-1} ... | d_1 | e$

By induction hypothesis, $c\backslash a_m ... \backslash a_{i+1} / f$ has $c | d_{k-1} ... | d_1 \backslash \langle a_m \rangle ... \backslash \langle a_{i+1} \rangle / f^{\{|d_{k-1} ... |d_1\}}$, and $f \backslash \langle a_i \rangle ... \backslash \langle a_1 \rangle | d_{k-1} ... | d_1 / e$ has $f^{\{|d_{k-1} ... |d_1\}} \backslash a_i ... \backslash a_1 | e$.

c. Reduction: $c\backslash a_m ... \backslash a_1 | d_{k-1} ... | d_1 | e \Longleftarrow c / f \quad f \backslash a_m ... \backslash a_1 | d_{k-1} ... | d_1 | e$

By induction hypothesis, $f \backslash a_m ... \backslash a_1 | d_{k-1} ... | d_1 / e$ has $f | d_{k-1} ... | d_1 \backslash \langle a_m \rangle ... \backslash \langle a_1 \rangle | e$. By the induction hypothesis of the main lemma (IH2) there is a constant category that generates the same string as c / f.

■

Proof: $L(G_{gtrc}) \supseteq L(G_{std-sim})$

We will use the following classification for the categories in $G_{std-sim}$.

(30) *a*. Const_2: Categories translated from Const of G_{gtrc}. Exclusive of the following.

b. GTRC: Categories translated from GTRC of G_{gtrc}. Represented as $\langle x \rangle$.

c. Wrap: Categories obtained by Wrapping. They may include wrapped argument represented as $\langle x \rangle$ and/or passed argument $\{\mathbb{P}\}$.

d. BGTRC: Categories obtained by Bounded GTRC.

Note that 'Const$_2$' in this classification stands in relation to 'Const' in G_{gtrc} and that all the categories in $G_{std-sim}$ are *constant*. We will drop the subscript on Const$_2$ where no confusion arises.

The proof is by induction on the height h of a derivation in $G_{std-sim}$. The primary recursion is on Const and we introduce **Lemma 7** to have a mutually-recursive situation on wrapped categories.

Lemma 6 The direction $L(G_{gtrc}) \supseteq L(G_{std-sim})$ of the Main Lemma.

Base case ($h = 0$): By the definition of Const$_2$ above, there is a corresponding constant lexical category with the same terminal string in G_{gtrc}.

Induction hypothesis (IH6): The lemma holds for $h' \leq h - 1$.

Induction step ($h \geq 1$): We consider the following cases that result in Const.

(31) *a*. Const+Const: Apply the induction hypothesis (IH6) to the functor and input categories. Then, the same strings can be generated from the corresponding categories in G_{gtrc}. Since we can apply the same rule in G_{gtrc}, we generates the same string from the same category.

b. Const+BGTRC, BGTRC+Const, and BGTRC+BGTRC: By the simulation, any bounded instance of GTRC in $G_{std-sim}$ has the corresponding GTRC in G_{gtrc}. Apply IH6 to the Const. Then, this case has the corresponding derivation. Note that there is no formal distinction between BGTRC and Const. Thus, there may be ambiguous case where a single derivation may need be considered for both cases, where only one of them may apply.

c. Const+Wrap, Wrap+Const, Wrap+BGTRC, BGTRC+Wrap: These cases do not apply. Regardless of the position of the indication of wrapping (either $\langle x \rangle$ or passed argument), either they fail to unify with the other category or would remain in the result category.

d. GTRC+<any class>, BGTRC+GTRC, Const+GTRC: Not applicable. GTRC-translated category $\langle x \rangle$ can only combine with the identical argument of a wrapped category.

e. Wrap+Wrap: The only applicable case is the following: "$a/b^{\{\mathbb{P}\}} \quad b^{\{\mathbb{P}\}}\mathbb{C} \implies a\mathbb{C}$" (other instances of wrapping are not applicable for the same reason as (3)). Apply **Lemma 7** to both categories.

f. Wrap+GTRC: The rule application takes the form: "$a/\langle b \rangle \quad \langle b \rangle \implies a$". By the simulation, the same string can be generated by the corresponding categories in G_{gtrc}.

■

For the case where the result category is Wrap, consider the following lemma.

Lemma 7 For a wrapped category c in $G_{std-sim}$, there is a constant category c' in G_{gtrc} that generates the same terminal string.

Proof: By induction on the height h of derivation.

Base case ($h = 0$): c is a lexical category in $G_{std-sim}$. There must be a category c' in G_{gtrc} by wrapping (24).

Induction hypothesis: The lemma holds for $h' \leq h - 1$.

Induction step ($h \geq 1$):

(32) *a*. Wrap+Wrap: The derivation takes the form: "$f^{\{\mathbb{O}\}}\mathbb{A}/b^{\{\mathbb{P}\}} \quad b^{\{\mathbb{P}\}}\mathbb{C}|d^{\{\mathbb{I}\}} \implies f^{\{\mathbb{O}\}}\mathbb{A}\mathbb{C}|d^{\{\mathbb{I}\}}$". Either $\mathbb{O}$ or $\mathbb{I}$ is non-nil. Apply the induction hypothesis to both categories. We have the corresponding $f\mathbb{O}\mathbb{A}'/b$ and $b\mathbb{P}\mathbb{C}'|d$ where $\mathbb{A}'$ and $\mathbb{C}'$ are the result of removing $\mathbb{P}$ and $\mathbb{I}$ from $\mathbb{A}$ and $\mathbb{C}$, respectively. They can derive: "$f\mathbb{O}\mathbb{A}'/b \quad b\mathbb{P}\mathbb{C}'|d \implies f\mathbb{O}\mathbb{A}'\mathbb{P}\mathbb{C}'|d = f\mathbb{O}\mathbb{A}\mathbb{C}'|d$".

b. Const+Wrap, Wrap+Const: Apply IH6 to Const and the induction hypothesis to Wrap. The rest is similar to the above.

c. No other case can result in Wrap.

■

Example of Simulation

Example 1 English heavy NP-shift

"John gave the book to Mary."

"[John gave to Mary] **the book which ....**"

$$f_{gtrc} = \left\{ \begin{array}{l} (\texttt{john}, NP), \left(\texttt{john}, \mathsf{T}\left\langle \left(\mathsf{T}\right\rangle NP\right)\right), (\texttt{the book}, NP), \left(\texttt{the book}, \mathsf{T}\left\langle \left(\mathsf{T}\right\rangle NP\right)\right), \\ (\texttt{to mary}, PP), (\texttt{to mary}, \mathsf{T}\backslash(\mathsf{T}/PP)), (\texttt{gave}, S\backslash NP/PP/NP), \ldots \end{array} \right\}$$

$$f_{std}^{base} = \left\{ \begin{array}{l} \textit{replace the GTRCs with } (\texttt{john}, \langle NP\rangle), (\texttt{the book}, \langle NP\rangle), (\texttt{to mary}, \langle PP\rangle), \\ \textit{the rest is the same} \end{array} \right\}$$

$$f_{std}^{sim} = \left\{ \begin{array}{l} \textit{add the following to the above} \\ (\texttt{gave}, S\backslash NP/NP/\langle PP\rangle), (\texttt{gave}, S/NP\backslash\langle NP\rangle/\langle PP\rangle), \ldots \end{array} \right\}$$

$$\dfrac{\dfrac{\text{John} \quad \text{gave} \quad \text{to Mary} \quad \text{the book which ....}}{\langle NP\rangle \quad S/NP\backslash\langle NP\rangle/\langle PP\rangle \quad \langle PP\rangle \quad NP}}{S/NP\backslash\langle NP\rangle}$$

$$S/NP$$

Example 2 Japanese long-distance extraction

"Mary-nom John-nom Mary-acc helped-comp thought."

"**Mary-acc** [Mary-nom John-nom helped-comp thought]."

$$f_{gtrc} = \left\{ \begin{array}{l} (\texttt{john-nom}, NP_{nom}), (\texttt{mary-nom}, NP_{nom}), (\texttt{john-acc}, NP_{acc}), (\texttt{mary-acc}, NP_{acc}), \\ (\texttt{john-nom}, \mathsf{T}/(\mathsf{T}\backslash NP_{nom})), (\texttt{mary-nom}, \mathsf{T}/(\mathsf{T}\backslash NP_{nom})), \\ (\texttt{john-acc}, \mathsf{T}/(\mathsf{T}\backslash NP_{acc})), (\texttt{mary-acc}, \mathsf{T}/(\mathsf{T}\backslash NP_{acc})), \\ (\texttt{helped}, S\backslash NP_{nom}\backslash NP_{acc}), (\texttt{comp}, S'\backslash S), (\texttt{thought}, S\backslash NP_{nom}\backslash S'), \ldots \end{array} \right\}$$

$$f_{std}^{base} = \left\{ \begin{array}{l} \textit{replace the GTRCs with } (\texttt{john-nom}, \langle NP_{nom}\rangle), (\texttt{mary-nom}, \langle NP_{nom}\rangle), \\ (\texttt{john-acc}, \langle NP_{nom}\rangle), (\texttt{mary-acc}, \langle NP_{nom}\rangle), \\ \textit{the rest is the same} \end{array} \right\}$$

$$f_{std}^{sim} = \left\{ \begin{array}{l} \textit{add the following to the above} \\ (\texttt{helped}, S\backslash NP_{acc}\backslash\langle NP_{nom}\rangle), \left(\texttt{helped}, S^{\{\backslash NP_{acc}\}}\backslash\langle NP_{nom}\rangle\right), \\ \left(\texttt{comp}, S'^{\{\backslash NP_{nom}\}}\backslash S^{\{\backslash NP_{nom}\}}\right), \\ \left(\texttt{thought}, S\backslash NP_{nom}\backslash NP_{acc}\backslash S'^{\{NP_{acc}\}}\right), \\ \left(\texttt{thought}, S\backslash NP_{acc}\backslash\langle NP_{nom}\rangle\backslash S'^{\{NP_{acc}\}}\right), \ldots \end{array} \right\}$$

Mary-acc	Mary-nom	John-nom	helped	-comp	thought
NP_{acc}	$\langle NP_{nom}\rangle$	$\langle NP_{nom}\rangle$	$S^{\{\backslash NP_{acc}\}}\backslash\langle NP_{nom}\rangle$	$S^{\prime\{NP_{acc}\}}\backslash S^{\{NP_{acc}\}}$	$S\backslash NP_{acc}\backslash\langle NP_{nom}\rangle\backslash S^{\prime\{NP_{acc}\}}$
				$S\backslash NP_{nom}\backslash\langle NP_{nom}\rangle\backslash S^{\{NP_{acc}\}}$	
			$S\backslash NP_{acc}\backslash\langle NP_{nom}\rangle\backslash\langle NP_{nom}\rangle$		

A.3 Worst-Case Polynomial Recognition Algorithm

This section presents a worst-case polynomial recognition algorithm for a subclass of CCG-GTRC (Poly-GTRC) by extending the polynomial algorithm of Vijay-Shanker and Weir [1990] for CCG-Std (Poly-Std). We will observe below that the crucial property of CCG-Std employed by Poly-Std can be extended to the subclass of CCG-GTRC with an additional condition. Let us start with a brief review of the intuition behind Poly-Std and then move on to Poly-GTRC. Note that Poly-Std has the second stage of structure building but we concentrate on the more critical part of recognition.

Polynomial Algorithm for CCG-Std

First, observe the following properties of CCG categories:

(33) *a*. The length of a category in a cell can grow proportionally to the input size.

b. The number of categories in a cell may grow exponentially to the input size.

For example, consider a lexicon $f = \{(\mathtt{a}, S/NP/S),\ (\mathtt{a}, S/PP/S)\}$. Then, for the input "$\underbrace{\mathtt{a}....\mathtt{a}}_{n}$", the top CKY-cell includes 2^n combinations of categories like $S\underbrace{\left\{\begin{smallmatrix}/NP\\/PP\end{smallmatrix}\right\}\ldots\left\{\begin{smallmatrix}/NP\\/PP\end{smallmatrix}\right\}}_{n}/S$ derived by functional composition. Thus, we have exponential worst-case performance with respect to the input size.

The idea behind Poly-Std is to store categories as if they were some kind of linked list. Informally, a long category $F|a_n...|a_2|a_1$ is stored as 'F this portion is *linked* in a cell $[p,q]$ with *index* $|a_2$ $|a_1$'. A crucial point here is that the instances of target category F and arguments a_2 and a_1 are *bounded*. We will come back to this point in the next subsection. The pair $[p,q]$ can be represented as a n^2 matrix. Thus, by setting up n^2 subcells in each CKY-table cell, we can represent a category in a finite manner.

The effectiveness of such a representation comes from the fact that CCG rule application does not depend on the entire category. Namely, in order to verify "$F\underbrace{|a_n...|a_2}_{\bigstar}|\underline{a_1} \quad \underline{b_0}|b_k...|b_1 \Longrightarrow F\underbrace{|a_n...|a_2}_{\bigstar}|b_k...|b_1$", the sequence marked by '★' does not need to be examined. Thus, for the functor category, we only need to check F and a_1 available in the current cell. In addition, since b_0 must be unified with a bounded a_1, and k is also bounded by k_{max}, the entire input category is bounded and thus can be stored in the current cell. Therefore, the proposed representation does not slow down this type of process. When the result category exceeds a certain limit, we leave the excessive portion right in the original cell and set up a link to it.

One complication is that when an argument (e.g., a_1 in the above example) is canceled, we may have to restore a portion of the category from the linked cell (as the 'index' for the cell is required). We need to scan the linked cells and find the categories with the same index from n^2 subcells. Even though there may be multiple such categories, all of them can be restored in one of n^2 subcells associated with the result category. This case dominates the computational complexity but can be done in $O\left(n^4\right)$. Since this is inside i, j, k of CKY-style loop, the overall complexity is $O\left(n^7\right)$, which can be improved to $O\left(n^6\right)$ by rearranging the loops. The following is an informal description of the algorithm:

(34) **Poly-Std algorithm**:

a. Initialize: set up lexical categories

b. Main loop: for $1 \leq i < j \leq n$,

for $i \leq k < j$, apply rule schemata as follows:

Conditions		Case	Intuition
$\lvert result \rvert$	Link info		
$< limit$	none	No link	$F\lvert a_n...\lvert a_1$
$< limit$	available	Pass link info	$F\boxed{\ }\lvert a_i...\lvert a_1 \rightarrow F\boxed{\ }\lvert a_i...\lvert a_1$
$\geq limit$	either	Set up a new link	$F\boxed{\lvert a_n...\lvert a_i+1}\lvert a_i...\lvert a_1$ ↓ $\boxed{\ }$
$= 0$	available	Restore the linked info	$\boxed{\lvert a_i...\lvert a_1}$ ↓ $F\boxed{\ }$

Polynomial Algorithm for CCG-GTRC

We first note that there are cases where a crucial property of CCG-Std cannot be maintained in CCG-GTRC. The property is that arguments of derived categories are bounded. Although there might be a polynomial algorithm for CCG-GTRC that does not depend on this property, we pursue a straightforward extension of Poly-Std with an additional condition on the rules.[16] In the rest of this section, we will concentrate on the subclass of CCG-GTRC constrained by the Bounded Argument Condition.

Poly-GTRC is an extension to Poly-Std. The basic organization of the algorithm is analogous to Poly-Std. We use the same $n^2 \times n^2$ CKY-style table and a similar representation for constant categories. But we need to deal with GTRCs in polynomial time as well. First, let us examine two representative cases of rule applications since this reveals the necessary conditions for polynomial parsing.

The inner sequence of GTRC can grow as a result of Case (9*e*) , "GTRC+GTRC", repeated below:

(35) $\mathsf{T}/\underset{\substack{\uparrow \\ (k \geq 1)}}{(\mathsf{T}|a_m...|a_2\backslash a_1)}$ $\quad \underline{\mathsf{U}}|(\mathsf{U}|c_p...|c_1)|d_{k-1}...|d_1 \implies \mathsf{T}|(\mathsf{T}|a_m...|a_1|c_p...|c_1)|d_{k-1}...|d_1$

The only information needed to determine if the rule is applicable is the directionality of the slash indicated by '↑'. Thus, we do not actually need to know the inner sequence of the functor or input categories. The inner sequence of the result GTRC can thus be represented as two links to the functor and input categories. This link information virtually encodes a kind of grammar for deriving the inner sequence and is thus considered an application of structure sharing [Billot and Lang, 1989; Dymetman, 1997]. The outer sequence can be represented in a way similar to the argument of constant category. Although there may be exponentially-many GTRCs associated with each CKY cell, the number of cell entries is bounded by the link destinations of the inner sequence and the finite representation for the outer sequence.

Next, consider Case (7*b*), "GTRC+Const". We need to show that the unification process of the underlined portions can be done in polynomial time. As the first approximation, consider this

[16]The length of the argument is still bounded by $O(n)$ in CCG-GTRC since the only source of unboundedness is GTRC inner sequences. If every argument can be represented in some finite manner with link information similar to the one used for the Poly-Std, polynomial recognition might be possible.

process as an iteration of (backward) functional application of the form "$\underline{a_i}$ $c_0|c_m...|\underline{c_i}$" for $i = 1$ to m where a_i and c_i are canceled. But recall that in general, we only store a finite portion of both the functor and the input categories in the current cell and the remaining information must be restored through the links. The restoration of the information could cost exponential time since there may be multiple links to lower locations at any point. Therefore it is crucial that we proceed from $i = 1$ to m so that no enumeration of all the instances of $c_i, ..., c_1$ and $a_i, ..., a_1$ in (7) is actually generated. The traversal of the link from c_1 and a_1 may introduce sets of categories C_i and A_i for each position of $i \geq 2$, as schematically shown below.

(36)
$$\begin{array}{ccccc} C_m & +s & C_2 & & \{c_1\} \\ \updownarrow & & \updownarrow & \longleftarrow & \updownarrow \\ A_m & +s & A_2 & & \{a_1\} \end{array}$$

Note that each set C_i and A_i are bounded. This is the crucial point we needed the Bounded Argument Condition. Now, suppose that an element in C_i is canceled with some elements in A_i. We can proceed to the next set C_{i+1} where the elements in C_{i+1} are obtained by traversing the links from the canceled elements in C_i. Notice that the recovery process may encounter GTRCs as a part of derivation. There are three such cases: (i) (7*b*): GTRCs can be ignored since they do not affect the recovery process, (ii) (8*a*), (9*a*): GTRCs are bounded, and (iii) (9*d*ii): process shifts to GTRC recovery shown below.

Once we move from C_i to C_{i+1}, the history of cancellation can be forgotten, as in the case of iterative functional application in Poly-Std. Thus, even though we have potentially exponential instances of $c_i, ..., c_1$, the traversal of this side can be done step-by-step without suffering the exponential effect.

The traversal of A_i's is more challenging. The availability of a_i for cancellation with some c_i depends on the history of the cancellation of $a_{i-1}, ..., a_1$. Actually, it depends on the *tree structure* exactly encoded by the structure sharing technique. The 'GTRC recovery algorithm' will be introduced below to handle this situation in polynomial time.

The other cases are variation of the previous one. The "Const+Const" case can be processed as in Poly-Std. Next, consider the "GTRC+Const" case.

(37) $\mathsf{T}|(\mathsf{T}|a_m...|a_1)|b_n...|b_2/\underline{b_1}$ $\underline{c}|d_k...|d_1 \implies \mathsf{T}|(\mathsf{T}|a_m...|a_1)|b_n...|b_2|d_k...|d_1$

This case can actually be handled in a way similar to the "Const+Const" case. The only point is that the representation of GTRC must be bounded to avoid exponential combination of a_i's. We will come back to the representation of GTRC below.

The "Const+GTRC" cases are simpler.

(38) *a*. $a/\underline{b} \quad \underline{\mathsf{T}|(\mathsf{T}|c_m...|c_1)|d_n...|d_{k+1}}|d_k...|d_1 \implies a|d_k...|d_1$

b. $a/\underline{b} \quad \underline{\mathsf{T}}|(\mathsf{T}|c_m...|c_1)|d_{k-1}...|d_1 \implies a|(b|c_m...|c_1)|d_{k-1}...|d_1 \quad (k \geq 1)$

We need to recover the contents of the GTRC in both cases. The GTRC in (38*a*) is bounded since category *b* in the functor category is bounded. The one in (38*b*) is bounded by k_{max} (for $|d_{k-1}...|d_1$) and the Bounded Argument Condition (for $b|c_m...|c_1$). Thus, the recovery process for both cases are bounded.

Recall the following cases for "GTRC+GTRC":

(39) *a*. $\mathsf{T}|(\mathsf{T}|a_m...|a_1)|b_n...|b_2/\underline{b_1} \quad \underline{\mathsf{U}|(\mathsf{U}|c_p...|c_1)|d_n...|d_{k+1}}|d_k...|d_1$

$\implies \mathsf{T}|(\mathsf{T}|a_m...|a_1)|b_n...|b_2|d_k...|d_1$

b. $\mathsf{T}|(\mathsf{T}|a_m...|a_1)|b_n...|b_2/\underline{b_1} \quad \underline{\mathsf{U}}|(\mathsf{U}|c_p...|c_1)|d_{k-1}...|d_1$

$\implies \mathsf{T}|(\mathsf{T}|a_m...|a_1)|b_n...|b_2|(b_1|c_p...|c_1)|d_{k-1}...|d_1 \quad (k \geq 1)$

*d*i. $\mathsf{T}/\underline{(\mathsf{T}|a_m...|a_2\backslash a_1)} \quad \underbrace{\mathsf{U}|(\mathsf{U}|c_p...|c_1)|d_n...|d_{k+m+1}|\overbrace{d_{k+m}...|d_{k+1}}^{|a_m...\backslash a_1}}_{\mathsf{T}}|d_k...|d_1$

$\implies \mathsf{U}|(\mathsf{U}|c_p...|c_1)|d_n...|d_{k+m+1}|d_k...|d_1$

*d*ii. $\mathsf{T}/\underbrace{(\mathsf{T}|a_m...|\overbrace{a_{m-j}...|a_2\backslash a_1)}^{\|}}_{a_{m-j,0}|a_{m-j,p}...|a_{m-j,1}} \quad \underbrace{\underbrace{\mathsf{U}_0}_{\mathsf{T}}|\overbrace{\mathsf{U}_j...|\mathsf{U}_1|(\mathsf{U}_0|\mathsf{U}_j...|\mathsf{U}_1|c_p...|c_1)}^{|a_m... \quad a_{m-j}=F|a_{(m-j,q)}...|a_{(m-j,1)}}|\overbrace{d_n...|d_{k+1}}^{...\backslash a_1}}_{}|d_k...|d_1$

$\implies F|a_{(m-j,q)}...|a_{(m-j,q-j-p)}|d_k...|d_1 \qquad \text{where } q \geq j+p$

The cases (a) and (b) are analogous to (38*a*) and (38*b*). For the case $(d\text{i})$, we start comparing the outer sequence of the input category and the inner sequence of the functor category. The outer sequence of the input category can be treated as if it were the arguments of a constant category. Since T spans greater than U, the entire inner sequence of the functor category must be exhausted by comparing with the outer sequence of the input category. The result category can be obtained

Initialization:

- Create an n^2 GTRC recovery table, R

Table setup (Stage 1):

- For each cell (top-down) $O(n^2)$
  - For each entry (depending on the midpoint) $O(n)$
    - Restore the derivation info from CKY table and store the children in the appropriate cells $O(n^2)$

Recovery (Stage 2):

- For each cell in the bottom row (right-to-left) $O(n)$
  - For each entry $O(n)$
    - If there is a matching category in the target category set
      - Mark the current entry as 'success'

      Otherwise
      - Mark the current entry as 'fail'
    - Do **status percolation**

Status percolation (subprocedure):

- For each cell (bottom-up) $O(n^2)$
  - For each entry $O(n)$
    - For each parent $O(n^2)$
      - If the parent is marked as 'fail'
        - Mark the current entry as 'fail'

        Otherwise
        - If the current entry is the right branch *and* marked as 'fail' *and* all the right branch siblings are marked as 'fail,
          - Mark the parent as 'fail'
        - If the current entry is the left branch *and* marked as 'success'
          - Mark the parent as 'success'

Figure A.1: GTRC Recovery Algorithm

from the remaining part of the inner sequence of the input category with the remaining sequence "$|d_k \ldots |d_1$".

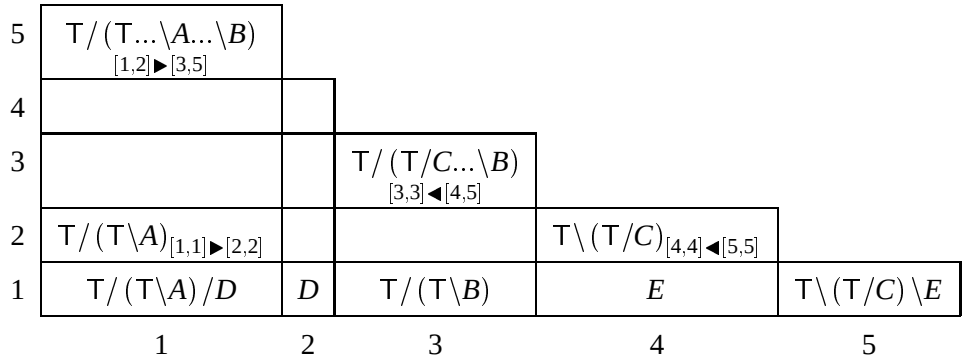
The case $(d\text{ii})$ is slightly different from the previous one in that the inner sequence of the functor category is excessive. We need a process of comparing the inner sequences of the functor and the input categories. Two GTRC recovery processes must be run in parallel.

Through the examination, we conclude that the polynomial parsability of CCG-GTRC depends on recovery of GTRCs. We present the polynomial GTRC recovery algorithm in Figure A.1. An example of GTRC recovery is given in Appendix B of Komagata [1997b].

The GTRC recovery algorithm takes advantage of the encoded shared structure, and utilizes an additional n^2 GTRC recovery table to restore possibly ambiguous GTRC derivations in polynomial

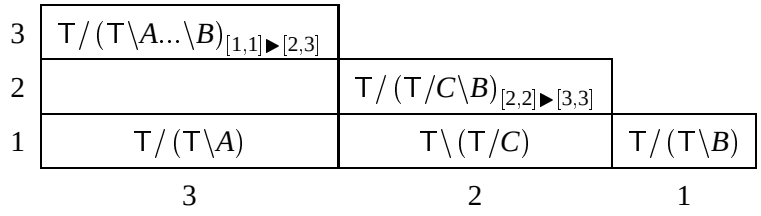
time. The first stage (*table setup*) is to represent the derivational structure available in the CKY table in a slightly different way. Suppose the following partial CKY table starting from a GTRC in question (here '▶' and '◀' indicate the direction of combination):

(40) Partial CKY table:



$T/(T...\backslash A...\backslash B)_{[1,2]\blacktriangleright[3,5]}$ represents the derivation "$T/(T\backslash A)_{[1,1]\blacktriangleright[2,2]}$ $T/(T/C...\backslash B)_{[3,3]\blacktriangleleft[4,5]} \Longrightarrow T/(T...\backslash A...\backslash B)$" at the designated string positions. Only the last argument of the link is stored in the current cell to avoid exponential number of entries. A GTRC recovery table can be used to store the same derivational structure with the bottom row corresponding to *the order of the inner sequence* of the GTRC rather than the string position. This is the order to process the inner sequence for C_i-A_i comparison shown in (36). Since GTRC recovery process only concerns the inner sequence of the GTRCs, the recovery table may have a dimension smaller than the corresponding portion of the CKY table, as seen in the following example:

(41) GTRC recovery table:



The categorial ambiguities originally aligned at string positions are now aligned in the order of processing.

In the second stage (*recovery stage*), the comparison with the target categories is done while the above-mentioned dependency among LTRCs in the bottom row is checked. The comparison proceeds from right to left in the bottom row. The decision on the cancellation of the argument under consideration, a_i, depends on (i) if it is unifiable with some target category (in C_i)

and (ii) if the corresponding sequence to the right of a_i was successfully canceled. This latter condition can be checked by observing the status of the first right branch from the current position since all the processes up to that point must have been completed. For the later processing (for the positions to the left), the success/failure status of the current category must also be percolated to the relevant higher nodes (*status percolation*). Note that even though the algorithm needs to check *all* the right branch siblings, the number of the siblings is bounded by the number of categories and directionalities. The total complexity turns out to be a rather daunting $O\left(n^{10}\right) = O(\underset{i,j,k}{n^3} \times \underset{\text{recovery}}{n^2} \times \underset{\text{percolation}}{n^5})$.

A.4 Progress Towards a Practical Parser for CCG-GTRC

This section investigates the performance of the experimental parser and demonstrates that it runs polynomially in practice. For both the practical parser and the theoretical algorithm, we use CKY-style parsing scheme [Aho and Ullman, 1972]. In addition to the use of CKY-table for recognition of the start category, we associate semantic representation for each category and derive the semantics in a single pass. We will focus on 'category-only' case for purely syntactic analyses but it should be noted that the parser can derive semantics and is not just a recognizer. Discussion of spurious ambiguity is included in Subsection 6.2.2.

We now look at the results of a pilot experiment done on Sun Ultra E4000 2x167MHz Ultraspacs with 320MB memory running SunOS 5.5.1. The program (100KB approx., about a half is the grammar) was written in Sicstus Prolog Ver. 3 and CPU time was measured by Sicstus' built-in predicate `statistics`. We parsed 22 contiguous sentences (6 paragraphs) in Japanese in a section arbitrarily chosen from "Anata-no Byouin-no Kusuri (Your Hospital Drugs) 1996" by Tadami Kumazawa and Ko-ichi Ushiro. The romanized sentences are partially-segmented to the word level but the verb inflection and suffixes are considered a part of the word. The average number of words in a sentence is 20 and the longest sentence contains 41 words. The sentences are realistically difficult, and include complex clauses (relative and complement), coordination (up to 4 conjuncts), nominal/verbal modifications (adjectives/adverbs), scrambling, and verb argument dropping.

The parser is based on a CKY algorithm equipped with Karttunen's equivalence check for

spurious ambiguity elimination but without the worst-case polynomial algorithm introduced in the previous section.[17] LTRCs are assigned to words by lexical rules and GTRCs are restricted to unidirectional forms. Coordination is handled by special trinomial rules [Steedman, 1996] with a few categorial features added to limit the coordination involving multiple constituents only to the left-branching structure. Verb argument dropping is handled by lexical rules that change the verb valency. Morphological analysis is a complete substring match and the results are dynamically 'asserted' among the code. About 200 lexical entries are asserted after parsing the 22 sentences. At the time of Komagata [1997a], morphological analysis takes about 0.2 seconds per word on average and needs improvement.[18] The output of a parse is an enumeration of the final result categories associated with the features and the semantics, as seen below (Sentence 7).

(42)
```
itumo 95mmHg o koeru baai_wa tiryou ga hituyoudesu.
always num(95mmHg) -ACC exceed in_case treatment [-NOM,-CONJv] necessary
  Cat:
    SS: s
    PA: (in_case always((exceed num(95mmHg) $1)) (necessary treatment))
  Cat:
    SS: s
    PA: always((in_case (exceed num(95mmHg) $2) (necessary treatment)))
CPU time: 280 ms  Elapsed: 320 ms  Words: 8  Solutions: 2
```

The unresolved pronoun is shown as '$\$n$' where $n \in \mathbb{N}$. The ambiguity regarding adverbial modification is left unresolved. The implementation has a simplified treatment of quantifiers and scope ambiguity too is left unresolved.

We consider the following two cases: (i) category-only and (ii) category+semantics. As we have discussed in the previous subsection, the application of equivalence check to the category-only case not only eliminates spurious ambiguities but also provides a result without genuine ambiguities.

Let us start with the analysis of the category-only case.[19] This case corresponds to the situation involving syntactic methods and also the polynomial algorithms introduced in the previous section. The results are shown in Figures A.2 (linear scale) and A.3 (log scale). Both exponential ($y =$

[17] Earlier applications of a CKY-style algorithm to CCG parsing include [Pareschi and Steedman, 1987].

[18] Whitelock [1988] has worked on morpho-syntax of Japanese in categorial framework. Some recent work on morphology includes [Hisamitsu and Nitta, 1994], [Tanaka et al., 1993].

[19] Category-only is the case also corresponding to the spurious ambiguity check of syntactic methods.

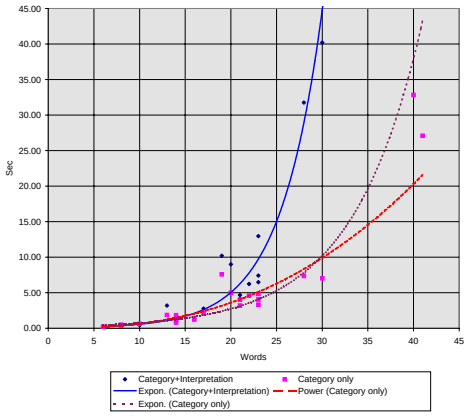


Figure A.2: Basic Data Set (linear scale)

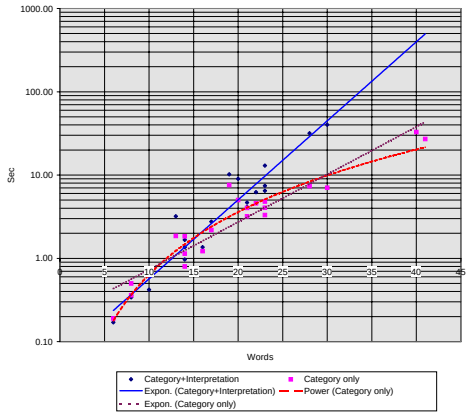


Figure A.3: Basic Data Set (log scale)

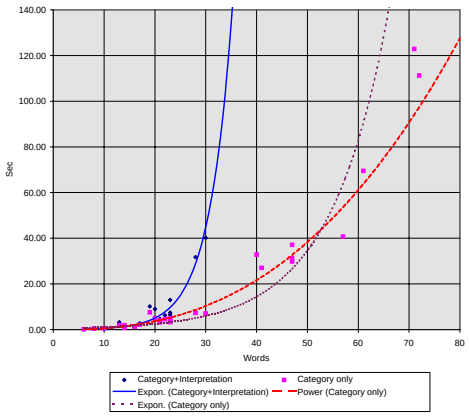


Figure A.4: Extended Data Set (linear scale)

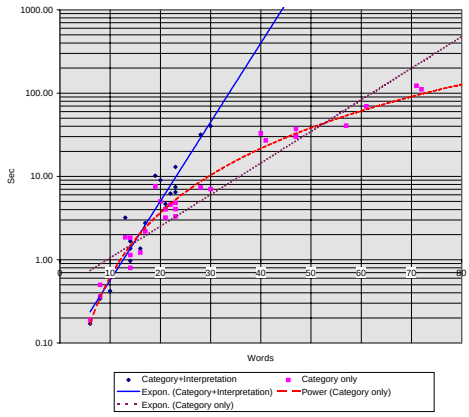


Figure A.5: Extended Data Set (log scale)

0.1963×1.14^n) and polynomial ($y = 0.002 \times n^{2.496}$) regression lines calculated by Microsoft Excel are provided. Although it is often easier to fit either an exponential or a polynomial curve on a log-scale graph, the data do not seem to be enough for such a conclusion. To see how the experiment might extend to the case with words longer than 45 words, we parsed pseudo-long sentences. That is, some of the test sentences are conjoined to form long sentences. Although these are semi-fabricated data, most long sentences are in fact the results of coordination. Thus, natural data are expected to behave similarly rather than differently from our pseudo-long sentences. The results are shown in Figures A.4 and A.5. The polynomial curves ($y = 0.0017 \times n^{2.5616}$) seem to represent the data better than the exponential curve ($y = 0.4375 \times 1.09^n$), especially on the log-scale graph. Since the data is sparse, we do not attempt to obtain a significant statistic analysis for these and

simply eye-fit the data. With these qualifications, we conclude that the performance appears no worse than n^3. The result also shows that categorial ambiguities still present in the parses are in practice within this bound.

A few remarks are in order. We compared our results with the following experiment to see how the figures stand. Tanaka and Ueki [1995] report that the LR-based syntactic analysis of a 19-word sentence in Japanese took 3.240sec.[20] The range of CPU times for our sentences with 19-23 words is between 3 to 8sec (category-only case). The performance of our parser seems to be within a comparable range.

Another point is that the effect of spurious ambiguity check is immediate. Without the check, only the sentences with 10 or fewer words were parsed. Under this condition, the maximum number of cell entries easily exceeds 300 for longer sentences, which resulted in out-of-space errors. We thus confirmed that the exponential effect of spurious ambiguity is well controlled by semantic equivalence check.

The above conclusion naturally remains qualified by the small scale of the experiment reported here. But, the test sentences are reasonably representative and relatively challenging. They vary in sentence length and complexity and span the space we may typically encounter. With additional data, it is reasonable to expect that the missing points will be filled and statistic significance will be obtained. It is also reasonable to believe that the experiment with pseudo-long sentences characterizes the kind of complexity that will be found in natural data.

Since one of the advantages of CCG parsers is the ability to derive semantics along with syntactic structure, the results of the category+semantics case is of special interest.[21] The situation naturally looks quite different. The exponential regression line $t = 0.0638 \times 1.24^n$ (Figures A.2 and A.3) seems to fit the data closely. In fact, the two longest sentences with 40 and 41 words results in out-of-space errors. Since spurious ambiguities are eliminated by equivalence check and categorial ambiguity is only polynomial, as shown in the category-only experiment, the exponential slow down is due exclusively to genuine ambiguity. Genuine ambiguity is a major problem for our parser as it is for any parser. We noticed that the longest two sentences become parsable if modifications across the top-level coordination are prohibited by assigning a special category to

[20] The word count is based on our criteria. Other details are ignored for now.

[21] The current implementation *enumerates* the derivations.

the sentential conjunctive. This kind of technique improves the performance, usually without affecting the completeness. But, of course, we need a more principled idea of how to deal with such a case. The following example (adopted from Sentence 8) shows how easily genuine ambiguities can grow:

(43) *a*. Modification ambiguities:

n-TOP $\underbrace{\mathbf{adjn}_1\text{-GEN}\ \mathbf{n}_2\text{-NOM}}_{\overrightarrow{\ }\text{2-way ambiguous}}$ v-coord, $\mathbf{adv}_1$ $\underbrace{\mathbf{adv}_2\mathbf{v}_1\text{-COMP-TOO}}_{\overrightarrow{\ }\text{3-way ambiguous}}$ v_2

b. Coordination ambiguities:

n-TOP adj n-GEN n-NOM v-**coord**, adv adv v-COMP-TOO v

Each case involves adjective modification ambiguity (2 cases each).

n-TOP adj n-GEN n-NOM v-**coord**, adv adv v-COMP-TOO v

Each case involves both adjective and adverb modification ambiguities (2×3 cases each).

c. Total ambiguities: $2 \times 2 + (2 \times 3) \times 2 = 16$

The worst case (Sentence 4) resulted in 96 parses. Since semantics is not considered by Poly-GTRC, Poly-GTRC does not affect the above situation.

A.5 Conclusion

Through the investigation of CCG-GTRC in detail, a subclass of CCG-GTRC is shown to be equivalent to CCG-Std. This is done by way of simulating unbounded, but restricted 'permutations' of CCG-GTRC by lexical wrapping and argument-passing across categories. This contrasts with the formalisms involving 'doubly'-unbounded scrambling, which are strictly more powerful than CCG-Std. Thus, CCG-GTRC can be used in place of CCG-Std to account for non-traditional constituents including the ones shown in the introduction without proliferation of type-raised categories with the same computational properties.

The most restrictive condition for the choice of the studied subclass seems to be 'no outer sequence'. This is also associated with the limitation that the instances of GTRCs are finite. Naturally, we want $\mathsf{T}\big/\!\!\backslash\left(\mathsf{T}\big\backslash\!\!/ PP\right)/NP$ for English prepositions and $\mathsf{T}/(\mathsf{T}\backslash NP)\backslash NP$ for Japanese particles and to derive categories freely. Inclusion of outer sequence seems to increase the power since

that class cannot be simulated by CCG-Std due to the fact that CCG-Std cannot simulate certain 'island'-like behavior of GTRCs. But, in practice, CCG-Std calls for additional mechanism such as conditions on rule application for various linguistic reasons. These conditions cannot be in general expressed in CCG-Std proper either. We are thus at the borderline of LTAG-equivalence. To find out where exactly we are is another question we want to ask.

We have also shown that the extension of CCGs including GTRCs can be parsed polynomially in theory and in practice, with some qualifications and conditions. These polynomial results support the proposed grammar that can describe non-traditional constituency widely observed across languages, without resorting to a special mechanism for each case. We expect that the grammar is also useful for practical applications.

The practical and the theoretical polynomial results are due to distinct factors. The former comes from a practical bound on the number of cell entries and spurious ambiguity elimination. The latter (for both Poly-Std and Poly-GTRC) is achieved by efficiently representing and processing the potentially exponentially-many entries in a cell. This is possible even with the presence of spurious/genuine ambiguities. But what the polynomial algorithms do is eliminate a possibility that rarely occurs in practice. The additional cost for Poly-Std/GTRC of managing n^2 subcells and links to cover all the cases of exponential factors including spurious/genuine ambiguities is thus considered overkill for the practical case. Although it may be possible to add spurious ambiguity check to Poly-Std/GTRC, we are better off with a simple CKY-style parser with equivalence check, without the overhead of the Poly-Std/GTRC.

For practical applications of the parser, though, we have an agendum for future research. A larger-scale experiment is necessary to obtain statistical significance for varying domains. We may want to consider potentially faster algorithms. For example, GLR-style algorithm may be extended to the proposed case. The most critical problem remains to be that of genuine ambiguity. We may explore a more compact representation of the derived semantics, e.g. polynomial shared structure algorithm of Dörre [1997].[22] Alternatively, we may try to disambiguate early during the recognition stage by a probabilistic method or contextual information (e.g., use of information structure). We expect that these techniques will be applicable to the presented parser and will

[22]Applicability of this technique to our parser needs to be carefully examined because semantic equivalence check will be required to traverse the shared structures. It is not clear if the traversal can be done in polynomial time. This concern is shared by the situation of applying structure sharing technique to conceptual dependency [Bröker et al., 1994]. Other reports on shared structure on semantic representation include [Nagao, 1994] and [Schiehlen, 1996].

improve the performance to a really practical level.

Bibliography

List of Abbreviations:

ACL Association for Computational Linguistics, Proceedings of the Annual Meeting of the

BLS Berkeley Linguistics Society, Proceedings of the Annual Meeting of the

CLS Chicago Linguistic Society, Papers from the Regional Meeting of the

COLING International Conference on Computational Linguistics, Proceedings of the

EACL European Chapter of the Association for Computational Linguistics, Proceedings of the Conference of the

ECAI European Conference on Artificial Intelligence, Proceedings of the

IJCAI International Joint Conference on Artificail Intelligence, Proceedings of the

INLG International Workshop on Natural Language Generation, Proceedings of the

IWPT International Workshop on Parsing Technologies, Proceedings of the

NELS North Eastern Linguistic Society, Proceedings of the Annual Meeting of the

RNLP Recent Advances in Natural Language Processing, Proceedings of International Conference on

Anthony Ades and Mark J. Steedman. 1982. On the Order of Words. *Linguistics and Philosophy*, 4:517–558.

Alfred V. Aho and J. D. Ullman. 1972. *The Theory of Parsing, Translation, and Compiling. Vol. 1: Parsing*. Englewood Cliffs, NJ: Prentice-Hall.

Reiko Aoki. 1992. *Gendaigo Joshi "Ha"-no Koubunron-teki Kenkyu (Syntactic Analysis of "Wa" in Modern Japanese)*. Tokyo: Kasama Shoin.

Jennifer Arnold, Antonio Losongco, Ryan Ginstrom, Amy Brynolfson, and Thomas Wasow. 1997. Save the worst for last: The effects of Syntactic Complexity and Information Structure on Constituent Ordering. In *Proceedings of Linguistic Society of America Annual Meeting, Chicago, January 1997*.

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Dordrecht: Kluwer.

Jay David Atlas. 1991. Topic/Comment, Presupposition, Logical Form and Focus Stress Implicatures: The Case of Focal Particles *only* and *also*. *Journal of Semantics*, 8:127–147.

J. L. Austin. 1962. *How to Do Things with Words*. Oxford: Clarendon Press.

Emmon Bach. 1979. Control in Montague Grammar. *Linguistic Inquiry*, 10(4):515–531.

Jason Baldridge. 1998. *Local Scrambling and Syntactic Asymmetries*. Masters thesis, University of Pennsylvania.

David Ian Beaver. 1997. Presupposition. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Language*, pages 939–1008. Cambridge, MA: MIT Press.

Tilman Becker, Aravind Joshi, and Owen Rambow. 1991. Long-Distance Scrambling and Tree Adjoining Grammars. In *EACL 5, Berlin, April 1991*, pages 21–26.

Sylvie Billot and Bernard Lang. 1989. The Structure of Shared Forests in Ambiguous Parsing. In *ACL 27, Vancouver, Canada, June 1989*, pages 143–151.

Betty J. Birner. 1994. Information status and word order: an analysis of English inversion. *Language*, 70(2):233–259.

Betty J. Birner. 1997. The Linguistic Realization of Inferrable Information. *Language and Communication*, 17(2):133–148.

Betty J. Birner and Gregory Ward. 1999. *Information status and noncanonical word order in English*. Amsterdam: John Benjamins.

Wayne C. Booth, Gregory G. Colomb, and Joseph M. Williams. 1995. *The Craft of Research*. Chicago: University of Chicago Press.

Johan Bos, Paul Buitelaar, and Anne-Marie Mineur. 1995. Bridging as coercive accommodation. In *Proceedings of the Workshop: Computational Logic for Natural Language Processing,*

Edinburgh, April 1995.

Norbert Bröker, Udo Hahn, and Susanne Schacht. 1994. Concurrent Lexicalized Dependency Parsing: The ParseTalk Model. In *COLING-94, Kyoto, August 1994*, pages 379–385.

Gillian Brown. 1995. *Speakers, listeners and communication: Explorations in discourse analysis*. Cambridge: Cambridge University Press.

Gillian Brown and George Yule. 1983. *Discourse Analysis*. Cambridge: Cambridge University Press.

Daniel Büring. 1997a. The Great Scope Inversion Conspiracy. *Linguistics and Philosophy*, 20:175–194.

Daniel Büring. 1997b. *The Meaning of Topic and Focus: The 59th Street Bridge Accent*. London: Routledge.

Greg N. Carlson. 1980. *Reference to kinds in English* (originally a PhD thesis in 1977). New York: Garland Publications.

Bob Carpenter. 1991. The Generative Power of Categorial Grammars and Head-Driven Phrase Structure Grammars with Lexical Rules. *Computational Linguistics*, 17(3):301–313.

Bob Carpenter. 1992. *The Logic of Typed Feature Structures*. Cambridge: Cambridge University Press.

Wallace L. Chafe. 1976. Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View. In Charles Li, editor, *Subject and Topic*, pages 25–55. New York: Academic Press.

Wallace Chafe. 1994. *Discourse, Consciousness, and Time*. Chicago: University of Chicago Press.

Gennaro Chierchia. 1989. Structured Meanings, Thematic Roles and Control. In Gennaro Chierchia, Barbara H. Partee, and Raymond Turner, editors, *Properties, Types, and Meaning. Vol. 2: Semantic Issues*, pages 131–166. Dordrecht: Kluwer.

Gennaro Chierchia and Sally McConnell-Ginet. 1990. *Meaning and Grammar: An Introduction to Semantics*. Cambridge, MA: MIT Press.

Hye-Won Choi. 1996. *Optimizing Structure in Context: Scrambling and Information Structure*.

PhD thesis, Stanford University.

Hye-Won Choi. 1997. Topic and Focus in Korean: The Information Partition by Phrase Structure and Morphology. In Ho min Sohn and John Haig, editors, *Japanese/Korean Linguistics, Vol. 6*, pages 545–561. Stanford, CA: CSLI Publications.

Noam Chomsky. 1971. Deep structure, surface structure, and semantic interpretation. In Danny D. Steinberg and Leon A. Jakobovits, editors, *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics, and Psychology*, pages 183–216. Cambridge: Cambridge University Press.

Patricia M. Clancy and Pamela Downing. 1987. The Use of Wa as a Cohesion Marker in Japanese Oral Narratives. In John Hinds and Shoichi Iwasaki, editors, *Perspectives on Topicalization: The Case of Japanese 'WA'*, pages 3–56. Amsterdam: John Benjamins.

Herbert H. Clark. 1996. *Using language*. Cambridge: Cambridge University Press.

Herbert H. Clark and Susan E. Haviland. 1977. Comprehension and the Given-New Contract. In Roy O. Freedle, editor, *Discourse Production and Comprehension*, pages 1–40. Norwood, NJ: Ablex.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Peter C. Collins. 1991. *Cleft and pseudo-cleft constructions in English*. London: Routledge.

Heles Contreras. 1976. *A theory of word order with special reference to Spanish*. Amsterdam: North Holland.

Ron Cowan. 1995. What are discourse principles made of? In Pamela Downing and Michael Noonan, editors, *Word Order in Discourse*, pages 29–49. Amsterdam: John Benjamins.

M. J. Cresswell. 1985. *Structured Meanings: The Semantics of Propositional Attitudes*. Cambridge, MA: MIT Press.

Peter W. Culicover and Michael Rochemont. 1983. Stress and Focus in English. *Language*, 59(1):123–165.

Deborah Dahl, Martha Palmer, and Rebecca Passoneau. 1987. Nominalization in PUNDIT. In *ACL 25, Stanford, CA, June 1987*.

Frantisek Daneš. 1974. Functional Sentence Perspective and the Organization of the Text. In Frantisek Daneš, editor, *Papers on functional sentence perspective*, pages 106–128. Prague: Academia and The Hague: Mouton.

Charles M. De Wolf. 1987. WA in Diachronic Perspective. In John Hinds and Shoichi Iwasaki, editors, *Perspectives on Topicalization: The Case of Japanese 'WA'*, pages 265–290. Amsterdam: John Benjamins.

Judy Delin. 1995. Presupposition and Shared Knowledge in *It*-Clefts. *Language and Cognitive Process*, 10(2):97–120.

Molly Diesing. 1992. *Indefinites*. Cambridge, MA: MIT Press.

Christy Doran, Dania Egedi, Beth Ann Hockey, B. Srinivas, and Martin Zaidel. 1994. XTAG system – A Wide Coverage Grammar for English. In *COLING-94, Kyoto, August 1994*, pages 922–928.

Jochen Dörre. 1997. Efficient Construction of Underspecified Semantics under Massive Ambiguity. In *ACL 35/EACL 8, Madrid, July 1997*, pages 386–393.

David R. Dowty. 1979. Dative 'Movement' and Thomason's Extensions of Montague Grammar. In Steven Davis and Marianne Mithun, editors, *Linguistics, Philosophy, and Montague Grammar*, pages 153–222. Austin, TX: University of Texas Press.

David R. Dowty. 1988. Type Raising, Functional Composition, and Non-Constituent Conjunction. In Richard Oehrle, Emmon Bach, and Deirdre Wheeler, editors, *Categorial Grammars and Natural Language Structures*, pages 153–197. Dordrecht: D. Reidel.

Matthew S. Dryer. 1996. Focus, Pragmatic presupposition, and activated proposition. *Journal of Pragmatics*, 26:475–523.

Marc Dymetman. 1997. Charts, Interaction-Free Grammars, and the Compact Representation of Ambiguity. In *IJCAI 97, Nagoya, Japan, August 1997*, pages 1002–1007.

Jan van Eijck. 1996. Presupposition and Information Updating. In M. Kanazawa et al., editors, *Quantifiers, Deduction, and Context*, pages 87–110. Stanford, CA: CSLI Publications.

Jason Eisner. 1996. Efficient Normal-Form Parsing for Combinatory Categorial Grammar. In

ACL 34, Santa Cruz, CA, June 1996, pages 79–86.

Martin Emms. 1993. Parsing with polymorphism. In *EACL 6, Utrecht, April 1993*, pages 120–129.

Nomi Erteschik-Shir. 1997. *The dynamics of focus structure*. Cambridge: Cambridge University Press.

Nomi Erteschik-Shir. 1998. The Syntax-Focus Structure Interface. In Peter W. Culicover and Louise McNally, editors, *Syntax and Semantics, Vol. 29: The limits of syntax*, pages 211–240. New York: Academic Press.

Alice Nancy Finn. 1984. Intonational Accompaniments of Japanese Morphemes *WA* and *GA*. *Language and Speech*, 27:47–57.

Jan Firbas. 1964. On Defining the Theme in Functional Sentence Analysis. *Travaux Linguistiques de Prague*, 1:267–280.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

W. A. Foley. 1994. Information Structure. In R. E. Asher, editor, *The Encyclopedia of Language and Linguistics, Vol. 3*, pages 1678–1685. Oxford: Pergamon Press.

Joyce Friedman and Ramarathnam Venkatesan. 1986. Categorial and Non-Categorial Languages. In *ACL 24, New York, NY, June 1986*, pages 75–77.

Peter H. Fries. 1994. On Theme, Rheme and discourse goals. In Malcolm Coulthard, editor, *Advances in Written Text Analysis*, pages 229–249. London: Routledge.

LTF Gamut. 1991. *Logic, Language, and Meaning. Vol. 2*. Chicago: University of Chicago Press.

Claire Gardent and Michael Kohlhase. 1996. Focus and Higher-Order Unification. In *COLING-96, Copenhagen, August 1996*.

Gerald Gazdar. 1979. *Pragmatics: Implicature, Presupposition, and Logical Form*. New York: Academic Press.

Gerald. Gazdar, Ewan Klein, Geoffrey K. Pullum, and Ivan A. Sag. 1985. *Generalized Phrase Structure Grammar*. Cambridge, MA: Harvard University Press.

H. P. Grice. 1975. Logic and Conversation. In Peter Cole and Jerry Morgan, editors, *Syntax and Semantics, 3: Speech Acts*, pages 305–315. New York: Academic Press.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2):203–226.

Jeanette K. Gundel. 1985. 'Shared Knowledge' and Topicality. *Journal of Pragmatics*, 9:83–107.

Jeanette K. Gundel. 1996. Relevance Theory Meets the Givenness Hierarchy An Account of Inferrables. In T. Fretheim and J. Gundel, editors, *Reference and Accessibility*, pages 141–153. Amsterdam: John Benjamins.

Takao Gunji. 1987. *Japanese Phrase Structure Grammar: A Unification-Based Approach*. Dordrecht: D. Reidel.

Carsten Günther, Claudia Maienborn, and Andrea Schopp. 1999. The Processing of Information Structure. In Peter Bosch and Rob A. van der Sandt, editors, *Focus: Linguistic, Cognitive, and Computational Perspectives*, pages 18–42. Cambridge: Cambridge University Press.

Carlos Gussenhoven. 1999. On the Limits of Focus Projection in English. In Peter Bosch and Rob A. van der Sandt, editors, *Focus: Linguistic, Cognitive, and Computational Perspectives*, pages 43–55. Cambridge: Cambridge University Press.

Liliane Haegeman. 1991. *Introduction to Government and Binding Theory*. Oxford: Blackwell Publishers.

Udo Hahn. 1995. Distributed Text Structure Parsing – Computing Thematic Progression in Expository Texts. In Gert Rickheit and Christopher Habel, editors, *Focus and Coherence in Discourse Processing*, pages 214–250. Berlin: Walter de Gruyter.

Udo Hahn, Katja Markert, and Michael Strube. 1996. A Conceptual Reasoning Approach to Textual Ellipsis. In *ECAI 96, Budapest, August 1996*, pages 572–576.

Eva Hajičová. 1991. Dependency-Based Parser for Topic and Focus. In Masaru Tomita, editor, *Current Issues in Parsing Technologies*, pages 127–138. Dordrecht: Kluwer.

Eva Hajičová, Petr Sgall, and Hana Skoumalová. 1993. Identifying Topic and Focus by an Automatic Procedure. In *EACL 6, Utrecht, April 1993*, pages 178–182.

Eva Hajičová, Hana Skoumalová, and Petr Sgall. 1995. An Automatic Procedure for Topic-Focus Identification. *Computational Linguistics*, 21(1):81–94.

Michael A. K. Halliday. 1967. Notes on Transitivity and Theme in English (Part II). *Journal of Linguistics*, 3:199–244.

Chung-hye Han. 1998. Asymmetry in the Interpretation of '-(n)un' in Korean. In Noriko Akatsuka et al., editors, *Japanese/Korean Linguistics 7*, pages 1–15. Stanford, CA: CSLI Publications.

Yoko Hasegawa. 1996. *A Study of Japanese Clause Linkage*. Stanford, CA: CSLI Publications.

John A. Hawkins. 1978. *Definiteness and indefiniteness : a study in reference and grammaticality prediction*. London: Croom Helm.

John A. Hawkins. 1991. On (in)definite articles: implicatures and (un)grammaticality prediction. *Journal of Linguistics*, 27:405–442.

John A. Hawkins. 1994. *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Irene Heim. 1982. *The Semantics of Definite and Indefinite Noun Phrases*. PhD thesis, University of Massachusetts, Amherst.

Julia E. Heine. 1998. Definiteness Predictions for Japanese Noun Phrases. In *COLING-ACL 98, Montreal, August 1998*, pages 519–525.

Herman Hendriks. 1993. *Studied Flexibility: Categories and Types in Syntax and Semantics*. PhD thesis, Universiteit van Amsterdam.

Herman Hendriks and Paul Dekker. 1996. Links without Locations. In *Tenth Amsterdam Colloquium*, pages 339–358.

Mark Hepple. 1987. Methods for Parsing Combinatory Grammars and the Spurious Ambiguity Problem.

Mark Hepple. 1990. *The Grammar and Processing of Order and Dependency: a Categorial*

Approach. PhD thesis, University of Edinburgh.

Mark Hepple and Glyn Morrill. 1989. Parsing and Derivational Equivalence. In *EACL 4, Manchester, April 1989*, pages 10–18.

Caroline Heycock. 1994. Focus Projection in Japanese. In *NELS 24, Amherst, MA, November 1994*, pages 157–171.

Roger Higgins. 1979. *The Pseudocleft Construction in English.* New York: Garland Publications.

John Hinds. 1987. Thematization, Assumed Familiarity, Staging, and Syntactic Binding in Japanese. In John Hinds, Senko K. Maynard, and Shoichi Iwasaki, editors, *Perspectives on Topicalization: The Case of Japanese 'wa'*, pages 83–106. Amsterdam: John Benjamins.

Toru Hisamitsu and Yoshihiko Nitta. 1994. An Efficient Treatment of Japanese Verb Inflection for Morphological Analysis. In *COLING-94, Kyoto, August 1994*, pages 194–200.

Jerry R. Hobbs. 1979. Coherence and Coreference. *Cognitive Science*, 3:67–90.

Beth Ann Hockey. 1998. *The Interpretation and Realization of Focus: An Experimental Investigation of Focus in English and Hungarian*. PhD thesis, University of Pennsylvania.

Beryl Hoffman. 1993. The Formal Consequences of Using Variables in CCG Categories. In *ACL 31, Columbus, OH, June 1993*.

Beryl Hoffman. 1994. Generating Context-Appropriate Word Orders in Turkish. In *INLG 94*.

Beryl Hoffman. 1995. *The Computational Analysis of the Syntax and Interpretation of "Free" Word Order in Turkish*. PhD thesis, University of Pennsylvania.

Beryl Hoffman. 1996. Translating into Free Word Order Languages. In *COLING-96, Copenhagen, August 1996*, pages 556–561.

Beryl Hoffman. 1998. Word Order, Information Structure, and Centering in Turkish. In Marilyn Walker et al., editors, *Centering in Discourse*. Oxford: Oxford University Press.

Laurence R. Horn. 1981. Exhaustiveness and the Semantics of Clefts. In *NELS 11*, pages 125–142.

Kei Huruta. 1982. Kokugo-de-no Hitei Hyougen-no Imi: Zokuchou (Afterthought of "Meaning of Negation in the Japanese Language"). *Mathematical Linguistics*, 13(7):296–315.

Shoichi Iwasaki. 1987. Identifiability, Scope-Setting, and the Particle Wa: A Study of Japanese Spoken Expository Discourse. In John Hinds and Shoichi Iwasaki, editors, *Perspectives on Topicalization: The Case of Japanese 'WA'*. Amsterdam: John Benjamins.

Ray S. Jackendoff. 1972. *Semantic interpretation in Generative Grammar*. Cambridge, MA: MIT Press.

Joachim Jacobs. 1986. The Syntax of Focus and Adverbials in German. In Werner Abraham and Sjaak de Meij, editors, *Topic, Focus, and Configurationality*, pages 103–127. Amsterdam: John Benjamins.

Gerhard Jäger. 1996. *Topics in Dynamics Semantics*. PhD thesis, Humboldt-Universität zu Berlin.

Gerhard Jäger. 1999. Topic, Focus and Weak Quantifiers. In Peter Bosch and Rob A. van der Sandt, editors, *Focus: Linguistic, Cognitive, and Computational Perspectives*, pages 187–212. Cambridge: Cambridge University Press.

Kristiina Jokinen and Tsuyoshi Morimoto. 1997. Topic Information and Spoken Dialogue Systems. In *Proceedings of Natural Language Processing Pacific Rim Symposium (NLPRS), Phuket, Thailand, December 1997*.

Aravind K. Joshi. 1985. Tree Adjoining Grammars. In David R. Dowty, Lauri Karttunen, and Arnold M. Zwicky, editors, *Natural Language Parsing*, pages 206–250. Cambridge: Cambridge University Press.

Aravind K. Joshi, Leon Levy, and Masako Takahashi. 1975. Tree adjunct Grammars. *Journal of Computer Systems Science*, 21(2):136–163.

Aravind K. Joshi, K. Vijay-Shanker, and David Weir. 1991. The Convergence of Mildly Context-Sensitive Grammatical Formalisms. In Peter Sells et al., editors, *Foundational Issues in Natural Language Processing*, pages 31–81. Cambridge, MA: MIT Press.

Aravind K. Joshi, Tilman Becker, and Owen Rambow. 1994. Complexity of Scrambling: A New Twist to the Competence - Performance Distinction. In *3e Colloque International sur les grammaires d'Arbres Adjoints*.

Hans Kamp. 1981. A theory of truth and semantic representation. In J. Groenendijk et al., editors, *Formal Methods in the Study of Language*, pages 277–322. Amsterdam: Mathematisch

Centrum.

Lauri Karttunen. 1973. Presuppositions of Compound Sentences. *Linguistic Inquiry*, 4(2):169–193.

Lauri Karttunen. 1976. Discourse referents. In J. McCawley, editor, *Syntax and Semantics, Vol. 7*, pages 363–385. New York: Academic Press.

Lauri Karttunen. 1986. Radical Lexicalism. Technical Report 86-68, Stanford, CA: CSLI Publications.

Lauri Karttunen and Stanley Peters. 1979. Conventional implicature. In *Syntax and Semantics, Vol. 11*, pages 1–56. New York: Academic Press.

Masashi Kawashima. 1989. Topic 'WA': Its Generation and Licensing. *Sophia Linguistica*, 27:57–69.

Martin Kay. 1975. Syntactic Processing and Functional Sentence Perspective. In *Proceedings of the Conference on Theoretical Issues in Natural Language Processing (TINLP)*, pages 12–15.

Martin Kay, Jean Mark Gawron, and Peter Norvig. 1994. *Verbmobil : a translation system for face-to-face dialog*. Stanford, CA: CSLI Publications.

Tracy Holloway King. 1995. *Configuring Topic and Focus in Russian*. Stanford, CA: CSLI Publications.

Kataline É. Kiss. 1981. On the Japanese 'Double Subject' Construction. *The Linguistic Review*, 1:155–170.

Kataline É. Kiss. 1987. *Configurationality in Hungarian*. Dordrecht: Kluwer.

Kataline É. Kiss. 1995. Introduction. In Kataline É. Kiss, editor, *Discourse Configurational Languages*. Dordrecht: Kluwer.

Nobo Komagata. 1997a. Efficient Parsing for CCGs with Generalized Type-Raised Categories. In *IWPT 97, Cambridge, MA, September 1997*, pages 135–146.

Nobo Komagata. 1997b. Efficient Parsing for CCGs with Generalized Type-Raised Categories. Technical report (IRCS-97-16), University of Pennsylvania.

Nobo Komagata. 1997c. Generative Power of CCGs with Generalized Type-Raised Categories.

In *ACL 35/EACL 8, Madrid, July 1997 (Student Session)*, pages 513–515.

Nobo Komagata. 1997d. Generative Power of CCGs with Generalized Type-Raised Categories. Technical report (IRCS-97-15), University of Pennsylvania.

Nobo Komagata. 1998a. Computer-Assisted Writing System: Improving Readability with Respect to Information Structure. In *Proceedings of International Conference on Natural Language Processing and Industrial Applications (NLP+IA98), Moncton, Canada, August 1998*, pages 200–204.

Nobo Komagata. 1998b. *Contrastive Function of Japanese Particle WA*. Presented at the First Annual Northeast Cognitive Science Society Graduate Conference (NECSS98), Ithaca, NY, May 1998. Available via http://www.cis.upenn.edu/ komagata/papers.html.

Esther König. 1994. A Hypothetical Reasoning Algorithm for Linguistic Analysis. *Journal of Logic and Computation*, 4(1):1–19.

Angelica Kratzer. 1995. Stage-level and Individual-level Predicates. In Gregory N. Carlson and Francis Jeffry Pelletier, editors, *The Generic Book*, pages 125–175. Chicago: University of Chicago Press.

Manfred Krifka. 1992. A Compositional Semantics for Multiple Focus. In Joachim Jacobs, editor, *Informationsstruktur und Grammatik (Linguistische Berichte, Sonderheft 4/1991-92)*, pages 17–53. Opladen: Westdeutscher Verlag.

Haruo Kubozono. 1993. *The Organization of Japanese Prosody*. Tokyo: Kuroshio Publishers.

Susumu Kuno. 1972. Functional Sentence Perspective: A Case Study from Japanese and English. *Linguistic Inquiry*, 3(3):269–320.

Susumu Kuno. 1973. *The Structure of the Japanese Language*. Cambridge, MA: MIT Press.

Susumu Kuno. 1976. Gapping: a functional analysis. *LI*, 7:300–318.

Susumu Kuno. 1978. *Danwa-no Bunpou (Discourse Grammar)*. Tokyo: Taishukan Shoten.

Jan van Kuppevelt. 1995. Discourse structure, topicality and questioning. *Journal of Linguistics*, 31:109–147.

Jan van Kuppevelt. 1996. Inferring from Topics: Scalar Implicatures as Topic-Dependent inferences. *Linguistics and Philosophy*, 19:393–443.

Akira Kurahone. 1983. *A Categorial Analysis of Derived Verbs in Japanese*. PhD thesis, University of Texas, Austin.

Sadao Kurohashi and Makoto Nagao. 1994. Automatic Detection of Discourse Structure by Checking Surface Information in Sentences. In *COLING-94, Kyoto, August 1994*, pages 1123–1127.

Robert Ladd. 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

Joachim Lambek. 1988. The Mathematics of Sentence Structure. In Wojciech Buszkowski, Witold Marciszewski, and Johan Van Benthem, editors, *Categorial Grammar*. Amsterdam: John Benjamins.

Knud Lambrecht. 1994. *Information Structure and Sentence Form: Topic, focus, and the mental representations of discourse referents*. Cambridge: Cambridge University Press.

Fred Landman. 1996. Plurality. In Shalom Lappin, editor, *The Handbook of Contemporary Semantic Theory*, pages 425–457. Oxford: Blackwell Publishers.

John Laver. 1994. *Principles of Phonetics*. Cambridge: Cambridge University Press.

Yae-Sheik Lee. 1993. Exhaustivity, The Scalar Principle, and Focus Semantics. In *Proceedings of the 13th West Coast Conference on Formal Linguistics (WCCFL)*.

Stephen C. Levinson. 1983. *Pragmatics*. Cambridge: Cambridge University Press.

David Lewis. 1979. Scorekeeping in a Language Game. In Ariner Bäuerle et al., editors, *Semantics from Different Points of View*, pages 172–187. Berlin: Springer-Verlag.

Susann LuperFoy. 1997. Discourse Processing for Voice-to-voice Machine Translation. In Christa Hauenschild and Susanne Heizmann, editors, *Machine Translation and Translation Theory*, pages 223–250. Berlin: Mouton de Gruyter.

Arman Maghbouleh. 1996. A Logistic Regression Model for Detecting Prominences. In *Proceedings of the Fourth International Conference on Spoken Language Processing (ICSLP), Philadelphia, PA, October, 1996*.

William Mann and Sandra Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Mitchell P. Marcus. 1980. *A Theory of Syntactic Recognition for Natural Language*. Cambridge, MA: MIT Press.

Vilém Mathesius. 1975. *A Functional Analysis of Present Day English on a General Linguistic Basis* (edited by Josef Vachek). The Hague: Mouton.

Christian M. I. M. Matthiessen and John A. Bateman. 1991. *Text Generation and Systemic-Functional Linguistics: Experiences from English and Japanese*. London: Pinter Publishers.

Senko K. Maynard. 1987. Thematization as a Staging Device in the Japanese Narrative. In John Hinds and Shoichi Iwasaki, editors, *Perspectives on Topicalization: The Case of Japanese 'WA'*, pages 57–82. Amsterdam: John Benjamins.

David B. McDonald. 1981. Compound: A Program that Understands Noun Compounds. In *IJCAI 81, Vancouver, Canada, August 1981*, page 1061.

Naomi Hanaoka McGloin. 1987. The Role of *Wa* in Negation. In John Hinds and Shoichi Iwasaki, editors, *Perspectives on Topicalization: The Case of Japanese 'WA'*, pages 165–184. Amsterdam: John Benjamins.

Kathleen R. McKeown. 1985. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge: Cambridge University Press.

Louise McNally. 1998. On the Linguistic Encoding of Information Packaging Instructions. In Peter W. Culicover and Louise McNally, editors, *Syntax and Semantics, Vol. 29: The limits of syntax*, pages 161–183. New York: Academic Press.

George Miller and Noam Chomsky. 1963. Finitary Models of Language Users. In R.R. Luce et al., editors, *Handbook of Mathematical Psychology, Vol. II*, pages 419–491. John Wiley and Sons.

Shigeru Miyagawa. 1987. Wa and the WH Phrase. In John Hinds and Shoichi Iwasaki, editors, *Perspectives on Topicalization: The Case of Japanese 'WA'*, pages 185–217. Amsterdam: John Benjamins.

Shigeru Miyagawa. 1997. Against Optional Scrambling. *Linguistic Inquiry*, 28(1):1–25.

Richard Montague. 1974. The Proper Treatment of Quantification in Ordinary English. In Richard H. Thompson, editor, *Formal Philosophy*, pages 247–270. New Haven, CT: Yale University Press.

Michael Moortgat. 1988. *Categorial Investigations: Logical and Linguistic Aspects of the Lambek Calculus*. Dordrecht: Foris.

Glyn V. Morrill. 1994. *Type logical grammar: categorial logic of signs*. Dordrecht: Kluwer.

Megan Moser and Johanna D. Moore. 1995. Investigating cue selection and placement in tutorial discourse. In *ACL 33, Cambridge, MA, June 1995*, pages 130–135.

Robert B. Most and Eli Saltz. 1979. Information Structure in Sentences: New Information. *Language-and-Speech*, 22(1):89–95.

Masaki Murata and Makoto Nagao. 1998. An Estimate of Referent of Noun Phrases in Japanese Sentences. In *COLING-ACL 98, Montreal, August 1998*, pages 912–916.

Makoto Nagao. 1989. *Machine Translation: How Far Can It Go? Translated by Norman D. Cook.* Oxford: Oxford University Press.

Katashi Nagao. 1994. A Preferential Constraint Satisfaction Technique for Natural Language Analysis. *IEICE Trans. Information & Systems*, E77-D(2).

Hisashi Noda. 1996. *WA-to GA (WA and GA)*. Tokyo: Kuroshio Publishers.

Elena V. Paducheva. 1996. Theme-rheme structures: Its exponents and its semantic interpretation. In Barbara H. Partee and Petr Sgall, editors, *Discourse and Meaning*, pages 273–287. Amsterdam: John Benjamins.

Martha S. Palmer. 1990. *Semantic Processing for Finite Domains*. Cambridge: Cambridge University Press.

Martha S. Palmer, Rebecca J. Passonneau, Carl Weir, and Tim Finin. 1993. The KERNEL text understanding system. *Artificial Intelligence*, 63:17–68.

Bart Papegaaij and Klaus Schubert. 1988. *Text Coherence in Translation*. Dordrecht: Foris.

Remo Pareschi and Mark Steedman. 1987. A Lazy Way to Chart-Parse with Categorial Grammars.

In *ACL 25, Stanford, CA, June 1987*, pages 81–88.

Jong Cheol Park. 1996. *A Lexical Theory of Quantification in Ambiguous Query Interpretations*. PhD thesis, University of Pennsylvania.

Jong C. Park, Martha Palmer, and Gay Washburn. 1997. An English Grammar Checker as a Writing Aid for Students of English as a Second Language. In *Descriptions of System Demonstrations and Videos of the Conference on Applied Natural Language Processing (ANLP5), Washington, DC, March-April 1997*.

Barbara H. Partee. 1996. Allegation and local accommodation. In Barbara H. Partee and Petr Sgall, editors, *Discourse and meaning: papers in honor of Eva Hajičová*, pages 65–86. Amsterdam: John Benjamins.

Barbara H. Partee. 1999. Focus, Quantification, and Semantics-Pragmatics Issues, Preliminary Version. In Peter Bosch and Rob A. van der Sandt, editors, *Focus: Linguistic, Cognitive, and Computational Perspectives*, pages 213–231. Cambridge: Cambridge University Press.

Lawrence C. Paulson. 1991. *ML for the Working Programmer*. Cambridge: Cambridge University Press.

Jaroslav Peregrin. 1996. Topic and focus in a formal framework. In Barbara H. Partee and Petr Sgall, editors, *Discourse and Meaning*, pages 235–254. Amsterdam: John Benjamins.

Fernando C.N. Pereira and Stuart M. Shieber. 1987. *Prolog and Natural-Language Analysis*. Stanford, CA: CSLI Publications.

Janet B. Pierrehumbert and Mary E. Beckman. 1988. *Japanese Tone Structure*. Cambridge, MA: MIT Press.

Janet Pierrehumbert and Julia Hirschberg. 1990. The Meaning of Intonational Contours in the Interpretation of Discourse. In Philip Cohen, Jerry Morgan, and Martha Pollack, editors, *Intentions in Communication*. Cambridge, MA: MIT Press.

Massimo Poesio and Renata Vieira. 1998. A Corpus-Based Investigation of Definite Description Use. *Computational Linguistics*, pages 183–216.

Massimo Poesio, Renata Vieira, and Simone Teufel. 1997. Resolving Bridging Descriptions in

Unrestricted Text. In *ACL Workshop on Operational Factors in Robust Anaphora Resolution, Madrid, July 1997*.

Carl Pollard and Ivan Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Paul Porter and Katsuhiko Yabushita. 1998. The Semantics and Pragmatics of Topic Phrases. *Linguistics and Philosophy*, 21:117–157.

Scott Prevost. 1995. *A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation*. PhD thesis, University of Pennsylvania.

Scott Prevost. 1996. An Information Structural Approach to Spoken Language Generation. In *ACL 34, Santa Cruz, CA, June 1996*, pages 294–301.

Scott Prevost and Mark Steedman. 1993. Generating Contextually Appropriate Intonation. In *EACL 6, Utrecht, April 1993*, pages 332–340.

Gary D. Prideaux. 1979. A Psycholinguistic Perspective on English Grammar. *Glossa*, 13(2):123–157.

Ellen F. Prince. 1978. A Comparison of Wh-clefts and *It*-clefts in Discourse. *Language*, 54(4):883–906.

Ellen F. Prince. 1981. Toward a Taxonomy of Given-New Information. In Peter Cole, editor, *Radical Pragmatics*, pages 223–256. New York: Academic Press.

Ellen F. Prince. 1984. Topicalization and Left-Dislocation: A Functional Analysis. In Sheila J. White and Virginia Teller, editors, *Annals of New York Academy of Sciences, Vol. 433: Discourses in Reading and Linguistics*, pages 213–225. New York: The New York Academy of Sciences.

Ellen F. Prince. 1986. On the Syntactic Marking of Presupposed Open Propositions. In *CLS22, Part 2*, pages 208–222.

Ellen F. Prince. 1992. The ZPG letter: subjects, definiteness, and information-status. In S. Thompson and W. Mann, editors, *Discourse description: diverse analyses of a fund raising text*, pages 295–325. Amsterdam: John Benjamins.

Ellen F. Prince. 1998. On the Limits of Syntax, with Reference to Left-Dislocation and Topicalization. In Peter W. Culicover and Louise McNally, editors, *Syntax and Semantics, Vol. 29: The limits of syntax*. New York: Academic Press.

Stephen G. Pulman. 1997. Higher Order Unification and the Interpretation of Focus. *Linguistics and Philosophy*, 20:73–115.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Owen Rambow. 1994. *Formal and Computational Aspects of Natural Language Syntax*. PhD thesis, University of Pennsylvania.

Owen Rambow and Aravind K. Joshi. 1994. A Processing Model for Free Word Order Languages. In Jr. C. Clifton, L. Frazier, and K. Rayner, editors, *Perspectives on Sentence Processing*, pages 267–301. Lawrence Erlbaum.

Random House. 1993. The Random House Unabridged Electronic Dictionary.

Keith Rayner and Alexander Pollatsek. 1987. Eye movements in reading: A tutorial review. In *Attention and performance 12: The psychology of reading*, pages 327–362. Hillsdale, NJ: Lawrence Erlbaum.

Tanya Reinhart. 1982. Pragmatics and linguistics: an analysis of sentence topics. *Philosophica*, 27:53–94.

Jeffrey C. Reynar. 1998. *Text Structuring: Algorithms, Applications and Evaluation.* PhD thesis, University of Pennsylvania.

Craige Roberts. 1996. Information Structure in Discourse: Towards an Integrated Formal Theory of Pragmatics. In *OSU Working Papers in Linguistics 49, Papers in Semantics*, pages 91–136.

Craige Roberts. 1998. Focus, the Flow of Information, and Universal Grammar. In Peter W. Culicover and Louise McNally, editors, *Syntax and Semantics, Vol. 29: The limits of syntax*, pages 109–160. New York: Academic Press.

Michael S. Rochemont. 1986. *Focus in generative grammar*. Amsterdam: John Benjamins.

Michael S. Rochemont. 1998. Phonological Focus and Structural Focus. In Peter W. Culicover

and Louise McNally, editors, *Syntax and Semantics, Vol. 29: The limits of syntax*, pages 337–363. New York: Academic Press.

Mats E. Rooth. 1985. *Association with Focus*. PhD thesis, University of Massachusetts, Amherst.

Mats Rooth. 1992. A Theory of Focus Interpretation. *Natural Language Semantics*, 1:75–116.

Mats Rooth. 1996. Focus. In Shalom Lappin, editor, *The Handbook of Contemporary Semantic Theory*, pages 271–297. Oxford: Blackwell Publishers.

Bertrand Russell. 1948. *Human knowledge, its scope and limits*. New York: Simon and Schuster.

Stuart J. Russell and Peter Norvig. 1995. *Artificial Intelligence: A Modern Approach*. Englewood Cliffs, NJ: Prentice-Hall.

Kumi Sadakane and Masatoshi Koizumi. 1995. On the nature of the "dative" particle ni in Japanese. *Linguistics*, 33:5–33.

Walter J. Savitch. 1987. Context-Sensitive Grammar and Natural Language Syntax. In Walter J. Savitch et al., editors, *The Formal Complexity of Natural Language*. Dordrecht: D. Reidel.

Yves Schabes. 1990. *Mathematical and Computational Aspects of Lexicalized Grammars*. PhD thesis, University of Pennsylvania.

Michael Schiehlen. 1996. Semantic Construction from Parse Forests. In *COLING-96, Copenhagen, August 1996*, pages 907–912.

Satoshi Sekine. 1996. Modeling Topic Coherence for Speech Recognition. In *COLING-96, Copenhagen, August 1996*, pages 913–918.

Elisabeth Selkirk. 1984. *Phonology and Syntax: The Relation Between Sound and Structure*. Cambridge, MA: MIT Press.

Petr Sgall. 1975. Focus and the Question Test. *Folia Linguistica*, 7(3-4):301–305.

Petr Sgall, Eva Hajičová, and Jarmila Panevova. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. Dordrecht: D. Reidel.

Masayoshi Shibatani. 1990. *The Languages of Japan*. Cambridge: Cambridge University Press.

Stuart M. Shieber. 1986. *An Introduction to Unification-Based Approaches to Grammar*. Stanford, CA: CSLI Publications.

Mitsuaki Shimojo. 1995. *Focus Structure and Morphosyntax in Japanese: WA and GA, and Word Order Flexibility*. PhD thesis, SUNY Buffalo.

Ken-ichiro Shirai. 1986. Japanese Noun-phrases and Particles wa and ga. In Jeroen Groenendijk et al., editors, *Studies in Discourse, Representation Theory and the Theory of Generalized Quantifiers*, pages 63–80. Dordrecht: Foris.

Melanie Siegel. 1999. The Syntactic Processing of Particles in Japanese Spoken Language. In *Proceedings of 13th Pacific Asia Conference on Language, Information, and Computation*.

Sidney Siegel and N. John Castellan, Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences, 2nd ed.* New York: McGraw-Hill.

K. Sparck Jones. 1983. So what about parsing compound nouns? In Karen Sparck Jones and Yorick Wilks, editors, *Automatic Natural Language Parsing*, pages 164–168. Chichester, England: Ellis Horwood.

Richard Sproat, editor. 1998. *Multilingual text-to-speech synthesis: the Bell Labs approach*. Dordrecht: Kluwer.

Robert Stalnaker. 1978. Assertion. In P. Cole, editor, *Syntax and Semantics, Vol. 9: Pragmatics*, pages 315–322. New York: Academic Press.

Arnim von Stechow. 1981. Topic, Focus, and Local Relevance. In Wolfgang Klein and Willem J. M. Levelt, editors, *Crossing the Boundaries in Linguistics*, pages 95–130. Dordrecht: D. Reidel.

Arnim von Stechow. 1991. Focusing and backgrounding operators. In Werner Abraham, editor, *Discourse Particles*, pages 37–84. Amsterdam: John Benjamins.

Mark J. Steedman. 1985. Dependency and Coordination in the Grammar of Dutch and English. *Language*, 61:523–56.

Mark J. Steedman. 1988. Combinators and Grammars. In Richard Oehrle, Emmon Bach, and

Deirdre Wheeler, editors, *Categorial Grammars and Natural Language Structures*, pages 417–442. Dordrecht: D. Reidel.

Mark J. Steedman. 1990. Gapping As Constituent Coordination. *Linguistics and Philosophy*, 13:207–2.

Mark Steedman. 1991a. Structure and Intonation. *Language*, 67:260–296.

Mark Steedman. 1991b. Type-Raising and Directionality in Combinatory Grammar. In *ACL 29, Berkeley, CA, June 1991*, pages 71–78.

Mark Steedman. 1996. *Surface Structure and Interpretation*. Cambridge, MA: MIT Press.

Mark Steedman. 1997. Making Use of Intonation in Interactive Dialogue Translation (invited talk). In *IWPT 97, Cambridge, MA, September 1997*, page xix.

Mark Steedman. 1999. *The Syntactic Process (to appear)*. Cambridge, MA: MIT Press.

Michael Strube. 1998. Never Look Back: An Alternative to Decentering. In *COLING-ACL 98, Montreal, August 1998*, pages 1251–1257.

Malgorzata E. Styś and Stefan S. Zemke. 1995. Incorporating Discourse Aspects in English-Polish MT: Towards Robust Implementation. In *Proceedings of Recent Advances in NLP 1995*. Available as cmp-lg/9510006.

Henriëtte de Swart. 1999. Position and Meaning: Time Adverbials in Context. In Peter Bosch and Rob A. van der Sandt, editors, *Focus: Linguistic, Cognitive, and Computational Perspectives*, pages 336–361. Cambridge: Cambridge University Press.

Anna Szabolcsi. 1981. Compositionality in Focus. *Folia Linguistica*, XV(1-2):141–161.

Anna Szabolcsi. 1983a. Focusing Properties, or the Trap of First Order. *Theoretical Linguistics*, 10(2/3):125–145.

Anna Szabolcsi. 1983b. The Semantics of Topic-Focus Articulation. In et. al J. A. G. Goenendijk, editor, *Formal Methods in the Study of Language*, pages 513–540. Amsterdam: Mathematisch Centrum.

Hozumi Tanaka and Masahiro Ueki. 1995. Japanese Parsing System using EDR Dictionary. `http://www.icot.or.jp/AITEC/PUBLICATIONS/Itaku/95/catalogue18-E.html`.

Hozumi Tanaka, Takenobu Tokunaga, and Michio Aizawa. 1993. Integration of Morphological and Syntactic Analysis Based on the LR Parsing Algorithm. In *IWPT 93, Tilburg, August, 1993*.

Koichi Tateishi. 1994. *The Syntax of 'Subject'*. Stanford, CA: CSLI Publications.

Hideo Teramura. 1991. *Nihongo-no Sintakusu-to Imi (Japanese Syntax and Meaning), Vol. 3*. Tokyo: Kuroshio Publishers.

Simon Thompson. 1991. *Type Theory and Functional Programming*. Reading, MA: Addison-Wesley.

Noriko Fujii Ueno. 1987. Functions of the Theme Marker *Wa* from Synchronic and Diachronic Perspectives. In John Hinds and Shoichi Iwasaki, editors, *Perspectives on Topicalization: The Case of Japanese 'WA'*, pages 221–263. Amsterdam: John Benjamins.

Yoich Uetake. 1992. Analysis of the Theme and Rheme Structure of a Japanese Sentence. *Lingua Posnaniensis*, 34:125–134.

Enric Vallduví. 1990. *The informational component*. PhD thesis, University of Pennsylvania.

Enric Vallduví and Elisabet Engdahl. 1994. Information Packaging and Grammar Architecture. In *NELS 25, Philadelphia, PA, October 1994*, pages 519–533.

Enric Vallduví and Elisabet Engdahl. 1996. The Linguistic Realization of Information Packaging. *Linguistics*, 34:459–519.

Enric Vallduví and Maria Vilkuna. 1998. On Rheme and Kontrast. In Peter W. Culicover and Louise McNally, editors, *Syntax and Semantics, Vol. 29: The limits of syntax*, pages 79–108. New York: Academic Press.

K. Vijay-Shanker. 1988. *A Study of Tree Adjoining Grammars*. PhD thesis, University of Pennsylvania.

K. Vijay-Shanker and David J. Weir. 1990. Polynomial Time Parsing of Combinatory Categorial Grammars. In *ACL 28, Pittsburgh, PA, June 1990*, pages 1–8.

K. Vijay-Shanker and David J. Weir. 1993. Parsing Constrained Grammar Formalisms. *Computational Linguistics*, 19(4):591–636.

K. Vijay-Shanker and D. J. Weir. 1994. The Equivalence of Four Extensions of Context-Free

Grammars. *Mathematical Systems Theory*, 27:511–546.

K. Vijay-Shanker, D.J. Weir, and Aravind Joshi. 1986. Tree Adjoining and Head Wrapping. In *COLING-86, Bonn, August 1986*, pages 202–207.

Marilyn A. Walker. 1992. Redundancy in Collaborative Dialogue. In *COLING-92, Nantes, August 1992*, pages 345–351.

Gregory L. Ward. 1990. The discourse functions of VP preposing. *Language*, 66:742–763.

Yasuko Watanabe. 1989. *The Function of "WA" and "GA" in Japanese Discourse*. PhD thesis, University of Oregon.

Bonnie Lynn Webber. 1983. So What Can We Talk About Now? In M. Brady and R. Berwick, editors, *Computationial Models of Discourse*, pages 331–371. Cambridge, MA: MIT Press.

Bonnie Webber. 1991. Structure and ostension in the interpretation of discourse deixis. *Natural Language and Cognitive Process*, 6(2):107–135.

Hae-Kyung Wee. 1995. Meaning and Intonation Associated with the Subject Marker and the Topic Marker in Korean. In *Harvard Studies in Korean Linguistics 6*. Cambridge, MA: Harvard University.

David Weir. 1988. *Characterizing Mildly Context-Sensitive Grammar Formalisms*. PhD thesis, University of Pennsylvania.

David J. Weir and Aravind K. Joshi. 1988. Combinatory Categorial Grammars: Generative Power and Relationship to Linear Context-Free Rewriting Systems. In *ACL 26, Buffalo, NY, June 1988*, pages 278–285.

Pete J. Whitelock. 1988. A Feature-Based Categorial Morpho-Syntax for Japanese. In U. Reyle and C. Rohrer, editors, *Natural Language Parsing and Linguistic Theories*, pages 230–261. Dordrecht: Kluwer.

Susanne Winkler. 1997. *Focus and Secondary Predication*. Berlin: Mouton de Gruyter.

Terry Winograd. 1972. *Understanding Natural Language*. New York: Academic Press.

Kent Barrows Wittenburg. 1986. *Natural Language Parsing with Combinatory Categorial Grammar in a Graph-Unification-Based Formalism*. PhD thesis, University of Texas, Austin.

Kent Wittenburg. 1987. Predictive Combinators: A Method for Efficient Processing of Combinatory Categorial Grammars. In *ACL 25, Stanford, CA, June 1987*.

Kent Wittenburg and Robert E. Wall. 1991. Parsing with Categorial Grammar in Predictive Normal Form. In Masaru Tomita, editor, *Current Issues in Parsing Technology*, pages 65–83. Dordrecht: Kluwer.

Mary McGee Wood. 1993. *Categorial Grammars*. London: Routledge.

Kei Yoshimoto. 1992. "Wa" to "Ga": Sorezore-no Kinoosuru Reberu-no Tigai-ni Tyuumoku-site ("Wa" and "Ga": Focusing on their different function levels). *Gengo Kenkyu*, 81:1–17.

Henk Zeevat. 1988. Combining Categorial Grammar and Unification. In Uwe Reyle and Chirstian Rohrer, editors, *Natural Language Parsing and Linguistic Theories*, pages 202–229. Dordrecht: D. Reidel.

Maria Lusa Zubizarreta. 1998. *Prosody, Focus, and Word Order*. Cambridge, MA: MIT Press.

Index

Page numbers in **boldface** indicate the locaion where the term is defined or discussed extensively.