

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*i*) might look like the following:

- (2) *Kotososyou_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 eikyouaremasu.
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) *Kotososyou_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)

¹The 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 (5ix) 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)	<i>wa</i>	(vi)	<i>wa</i>
(iii)	<i>ga</i>	(vii)	<i>wa</i>
(iv)	<i>ga</i>	(viii)	<i>ga</i>
(v)	<i>wa</i>	(ix)	<i>wa</i>

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		Hypothesis 2	
		{ Disc-old → <i>wa</i> Disc-new → <i>ga</i>		{ Definite → <i>wa</i> Otherwise → <i>ga</i>	
(ii)	<i>wa</i>	Old	✓	Indefinite	*
(iii)	<i>ga</i>	New	✓	Indefinite	✓
(iv)	<i>ga</i>	New	✓	Indefinite	✓
(v)	<i>wa</i>	Old	✓	Indefinite	*
(vi)	<i>wa</i>	New	*	Definite	✓
(vii)	<i>wa</i>	New	*	Definite	✓
(viii)	<i>ga</i>	New	✓	Indefinite	✓
(ix)	<i>wa</i>	New	*	Definite	✓

✓ : 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

²Japanese 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 (*8a*) and (*8b*) 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:

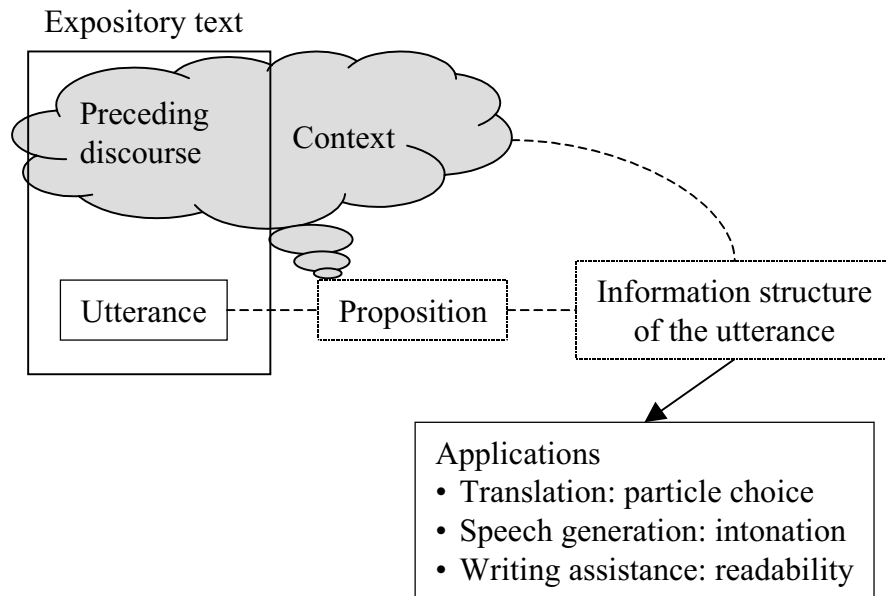


Figure 1.1: The Phenomenon under Investigation

(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

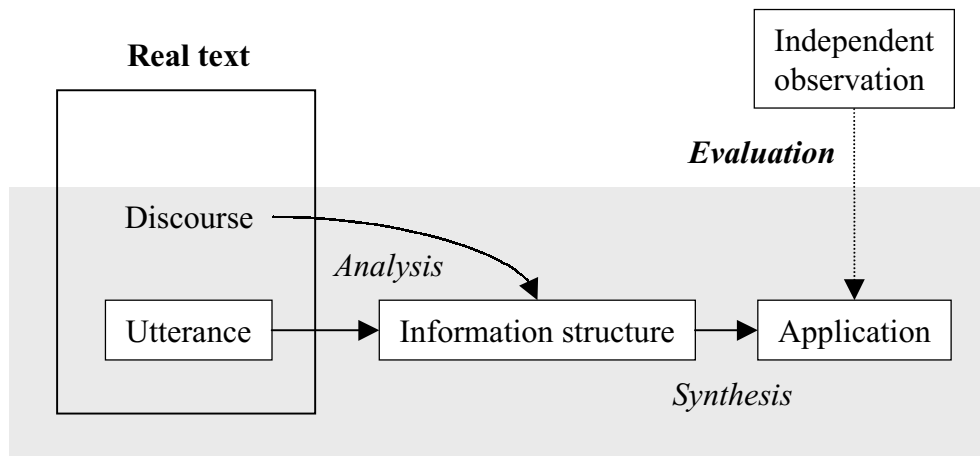


Figure 1.2: Limitations of Previous Approaches to the Identification Problem

Nagao, 1994; Hajičová et al., 1995; Hahn, 1995; Styš 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], Styš 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 Categorical 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

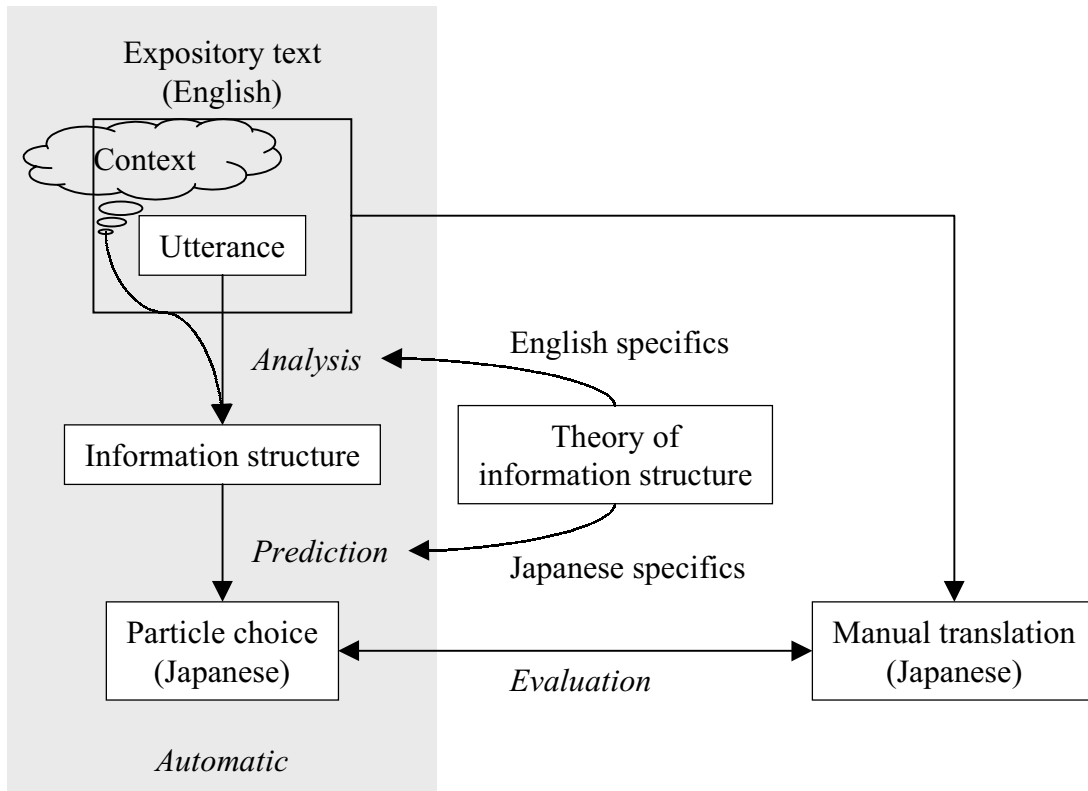


Figure 1.3: Overview of the Project

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 discontinuous 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 Categorical Grammar (CCG) [Ades and Steedman, 1982]. We show that specification of ‘contextual link’ can be formalized within the framework, and analysis of discontinuous 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 Categorical 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.