# 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.