

## 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  
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.<sup>1</sup>

<sup>1</sup>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 discontinuous. The process of derivation and information-structure analysis are shown below.

- (220) a. [Whether patients are younger or older,]<sub>CL1</sub> [*the social costs of osteoporosis*]<sub>CL2</sub> [are enormous.]<sub>NL</sub>
- b. [Whether patients are younger or older,]<sub>CL1</sub> [*the social costs of osteoporosis are enormous.*]<sub>(CL2,NL)</sub>
- c. [Whether patients are younger or older, *the social costs of osteoporosis are enormous.*]<sub>(CL1+CL2, NL)</sub>
- $$\begin{array}{ccc} \downarrow & & \downarrow \\ \text{Theme} & & \text{Rheme} \end{array}$$

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.<sup>2</sup> 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)

---

<sup>2</sup>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.<sup>3</sup>

- (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.<sup>4</sup> They are all native speakers of Japanese (two male and two female). Three of them have some

---

<sup>3</sup>The texts used in this preliminary experiment are taken from the same journal but not included in the training nor the test data.

<sup>4</sup>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 discontinuous 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				<i>wa/ga</i> choice		
	N	A	F	I	<i>wa</i>	<i>ga</i>	n/a
(ii)	<i>wa</i>	<i>wa</i>	<i>wa</i>	<i>wa</i>	4	0	0
(iii)	<i>ga</i>	n/a	n/a	n/a	0	1	3
(iv)	<i>ga</i>	<i>ga</i>	<i>ga</i>	n/a	0	3	1
(v)	<i>wa</i>	<i>wa<sub>drop</sub></i>	<i>wa</i>	<i>wa<sub>drop</sub></i>	4	0	0
(vi)	<i>wa</i>	n/a	<i>wa</i>	n/a	2	0	2
(vii)	<i>wa</i>	<i>wa</i>	<i>wa</i>	<i>wa</i>	4	0	0
(viii)	<i>ga</i>	<i>ga</i>	<i>ga</i>	<i>ga</i>	0	4	0
(ix)	<i>wa</i>	<i>wa</i>	<i>wa</i>	<i>wa</i>	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	<i>wa</i>	<i>ga</i>	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  $\kappa$  statistic, following the procedure described in Siegel and Castellan [1988].<sup>5</sup> The  $\kappa$  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  $\kappa$  statistic distributes approximately normally. Therefore, it is possible to estimate the significance of a  $\kappa$  statistic in terms of, in our case, a *z* score. Since the  $\kappa$  statistic simply scales from chance to perfect agreement, comparing  $\kappa$  statistics

<sup>5</sup>The standard reference for the  $\kappa$  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  $\kappa$  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*.<sup>6</sup> For example, in Table 7.2, Utterances (iii), (iv), and (vi) are no longer available for the four-rater comparison.

First, the  $\kappa$  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 ( $p < .05$ ),<sup>7</sup> 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	–	$\kappa$ $z$ <b>0.59/2.69</b>	0.42/1.56	<b>0.46/2.31</b>
A	–	–	<b>0.46/1.65</b>	<b>0.39/1.71</b>
F	–	–	–	0.19/0.77
I	–	–	–	–

Table 7.4: Agreement between Two Translators (Training Data)

The  $\kappa$  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:

<sup>6</sup>We still include the dropping case, though.

<sup>7</sup>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** =  $\frac{\text{number of correctly-predicted target data}}{\text{number of total target data}}$
- b. **Precision** =  $\frac{\text{number of correctly-predicted data}}{\text{number of total predicted data}}$

Recall/precision is calculated for several alternative hypotheses, as shown in Table 7.5.

Hypothesis	<i>wa</i> (Target = 82)				<i>ga</i> (Target = 9)			
	Predicted		Recall (%)	Precision (%)	Predicted		Recall (%)	Precision (%)
	Correct	Total			Correct	Total		
All <i>wa</i>	82	91	$\frac{82}{82}=100$	$\frac{82}{91}=90$	0	0	$\frac{0}{9}=0$	$\frac{0}{0}=n/a$
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
<b>Proposed</b>	<b>73</b>	<b>73</b>	<b>89</b>	<b>100</b>	<b>9</b>	<b>18</b>	<b>100</b>	<b>50</b>

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  $\kappa$  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 training 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]<sub>*wa*</sub> 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]<sub>*Rheme*</sub> [at risk.]<sub>*Theme*</sub> (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]<sub>*NL*</sub> [many active patients]<sub>*CL*</sub> [at risk.]<sub>*CL*</sub>

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.]<sub>*ga*</sub>

**Japanese:** *Among cheerleading routines, there are gymnastic elements, tumbling runs, partner stunts, pyramid formations, and dance routines.*

b. (System) [Cheerleading routines]<sub>*Theme*</sub> [can include gymnastic elements, tumbling runs, partner stunts, pyramid formations, and dance routines.]<sub>*Rheme*</sub>

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]<sub>*wa*</sub> 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]<sub>*Theme*</sub> [is on injury assessment and detection.]<sub>*Rheme*</sub>

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]<sub>ga</sub> 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]<sub>Theme</sub> [may result in disability years later.]<sub>Rheme</sub>

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,]<sub>wa</sub> 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]<sub>Theme</sub> [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.]<sub>Rheme</sub>

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 “ $\langle CL_1, - \rangle + \langle CL_2, NL_2 \rangle_{NL-CL} \Rightarrow \langle CL', NL_2 \rangle_{CL-NL}$ ”<sup>8</sup>

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  $\kappa$  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  $\kappa$  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

---

<sup>8</sup>Here,  $NL'$  is a composition of  $CL_2$  and  $NL_2$ , and  $CL'$  is a composition of  $CL_1$  and  $CL_2$ .

Translator	<i>wa</i>	<i>ga</i>	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	–	<sup>k</sup> <b>0.60/2.50</b> <sub>z</sub>	<b>0.48/1.67</b>	0.28/1.14	<b>0.55/2.36</b>	<b>0.44/1.83</b>
A	–	–	0.25/0.89	0.26/1.07	0.54/2.09	<b>0.48/1.93</b>
F	–	–	–	0.16/0.60	<b>0.47/1.66</b>	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.

Hypothesis	<i>wa</i> (Target = 44)				<i>ga</i> (Target = 7)			
	Predicted		Recall (%)	Precision (%)	Predicted		Recall (%)	Precision (%)
	Correct	Total			Correct	Total		
All <i>wa</i>	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
<b>Proposed</b>	<b>36</b>	<b>37</b>	<b>82</b>	<b>97</b>	<b>6</b>	<b>14</b>	<b>86</b>	<b>43</b>
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  $\kappa$  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.