

## Chapter 5

### An Extended Centering Model for Thai

In Chapter 3, we have shown that the necessary condition for the use of a zero pronoun in Thai is that the referent must have the cognitive status of ‘in focus’, and Centering Theory can be used for keeping track of the focused entities. In this chapter, we will argue that Centering Theory, which has been used for other languages that have zeroes, is also applicable to Thai. However, since the model does not take the hierarchical structure of discourse into account, it cannot be used to resolve zero pronouns with distant antecedents. In this chapter, we will propose an extended centering model that incorporates the hierarchical structure of discourse for zero pronoun resolution in Thai.

#### **5.1 Centering in Thai**

As discussed in Chapter 3, Centering Theory explicates coherence in a discourse segment in terms of centers. Centers are discourse entities that serve to link utterances in a segment. The theory assumes that an utterance contains one backward looking center (Cb) and a set of forward looking centers (Cf). The Cb is regarded as the center of attention of the utterance while Cf is an ordered list of discourse entities realized in the utterance. Constraints and rules as used in the model will be presented first. Then, we will discuss the problem of center establishment and how to order

entities in the Cf for Thai. The last part of this section discusses how the centering model can be used for resolving zero pronouns in Thai and how the problem of distant antecedents can be resolved.

### 5.1.1 Constraints and rules

Constraints and rules for the centering model used for the analyses Japanese and Italian are similar to the original version in Grosz et al. (1995). But Rule 1 accounts for zero pronouns instead of overt pronouns because zero pronouns are the correspondent form of focused entities in these languages. Constraints and rules in the theory are stated below:

(104)

Constraints:

For each utterance  $U_i$  in a discourse segment  $U_1, \dots, U_n$ :

1. There is precisely one backward-looking center Cb.
2. Every element of the forward centers list, Cf( $U_i$ ), must be realized in  $U_i$ .
3. The center, Cb( $U_i$ ), is the highest-ranked element of Cf( $U_{i-1}$ ) that is realized in  $U_i$ .

(105)

RULE 1: If any element of Cf( $U_n$ ) is realized by a zero pronoun in  $U_{n+1}$ , then the Cb( $U_{n+1}$ ) must be realized by a zero pronoun also.

RULE 2: Sequences of continuation are preferred over sequences of retaining; and sequences of retaining are to be preferred over sequences of shifting. In particular, a pair continuous across  $U_n$  and across  $U_{n+1}$ , represented as Cont( $U_n, U_{n+1}$ ) and Cont( $U_{n+1}, U_{n+2}$ ) respectively, is preferred over a pair of Retainings, Ret( $U_n, U_{n+1}$ ) and Ret( $U_{n+1}, U_{n+2}$ ). This case is analogous for pairs of Retainings and pair of shifts.

The first and the second constraints are the basic assumptions of the model. The third constraint restricts the  $C_b$  to the highest-ranked entity in the previous  $C_f$  that is contained in the current utterance.  $C_f$  is a list of entities in the utterance that are ordered according to their discourse saliency. One entity in the  $C_f$  is expected to be the  $C_b$  of the utterance. It is defined as the highest ranked entity of the immediately preceding utterance's  $C_f$  that is realized in the current utterance. This reflects the assumption that the more salient the discourse entity, the more likely it will be the center of attention ( $C_b$ ) of the next utterance. Thus, the highest-ranked entity in the current  $C_f$  is regarded as the preferred center for the next utterance ( $C_p$ ).

A transition state represents the transition of attention from one utterance to the next. It is determined by comparing the  $C_b$  of the preceding utterance with the  $C_b$  of the current utterance and by comparing the  $C_b$  and  $C_p$  of the current utterance. The list of transition states is shown below (Walker et al. 1994:200):

- (106) Continuation:  $C_b(U_i) = C_b(U_{i-1})$  and  $C_b(U_i) = C_p(U_i)$   
 Retaining:  $C_b(U_i) = C_b(U_{i-1})$  and  $C_b(U_i) \neq C_p(U_i)$   
 Smooth-shift:  $C_b(U_i) \neq C_b(U_{i-1})$  and  $C_b(U_i) = C_p(U_i)$   
 Rough-shift:  $C_b(U_i) \neq C_b(U_{i-1})$  and  $C_b(U_i) \neq C_p(U_i)$

A pair of utterances,  $U_i$  and  $U_{i-1}$ , is continuous when both utterances have the same  $C_b$  and  $C_b$  of  $U_i$  is the same as  $C_p$  of  $U_i$ . Continuation represents a transition state in which the center of attention is the same in both utterances. When  $C_b$  of  $U_i$  differs from  $C_p$  of  $U_i$  while both  $U_{i-1}$  and  $U_i$  have the same  $C_b$ , the transition state is

called ‘retaining’. Retaining represents a transition state in which the center of attention is retained in the current utterance ( $U_i$ ) but it is likely to be changed in the next utterance. When  $C_b$  of  $U_i$  is different from  $C_b$  of  $U_{i-1}$ , shifting of attention occurs. There are two kinds of shifting. A smooth-shift is the transition state in which the  $C_p$  of  $U_i$  is the same as the  $C_b$  of  $U_i$ . A rough-shift is the transition in which the  $C_p$  of  $U_i$  is different from the  $C_b$  of  $U_i$ .

### 5.1.2 Preferences of transition states

Preferences of transition states are used to determine which transition state is preferred to others at a certain point. In other words, they are used to determine whether the center of attention ( $C_{b_i}$ ) should remain the same as the previous one ( $C_{b_{i-1}}$ ) and whether the preferred center ( $C_{p_i}$ ) should be the same as the center ( $C_{b_i}$ ). As discussed in Chapter 3, centering algorithms use these preferences of transition states to determine preferred referents for zero pronouns. The preferences of transition states that are generally accepted in the centering literature are as follows (Brennan et al.1987):

(107) Continuation >> Retaining >> Smooth-shift >> Rough-shift

It is generally assumed that this hierarchy is always applicable. However, there is some evidence that preferences are sensitive to the previous transition state. According to a reading comprehension experiment conducted by Gordon et al.

(1993:340), ‘shifting’ is preferred to ‘continuation’ when the previous transition state is ‘retaining’. Therefore, in this study we propose preferences of transition state with respect to previous transition state.

The proposal here is adopted from Strube and Hahn’s (1996) discussion of ‘cheap’ and ‘expensive’ transition pairs. According to Strube and Hahn, the costs (‘cheap’ and ‘expensive’) are to be understood in terms of human sentence processing effort. We shall assume that, all others things being equal, an option with a cheaper processing cost will be forward over a more expensive one. Therefore, we will use  $Cb_i = Cp_{i-1}$  as the main criterion for setting the transition state preferences. Any transition in which the current Cb is the same as the previous Cp is considered more preferred than others. With this criterion, ‘continuation’ and ‘retention’ are preferred over both shifting, when the previous transition is ‘continuation’ or ‘smooth-shift’. To dertermine the preference between ‘continuation’ and ‘retention’, we also assume that the cost of human sentence processing effort is lower when the Cb is likely to be continued in the next utterance. Thus, we will use  $Cb_i = Cp_i$  as the second criterion for setting the transition state preferences. Any transitions in which the current Cp is the same as the current Cb is preferred to others. The preferences of transition states are listed in Figure 15 below.

Previous state	Preferences of transition state			
Continuation $Cb_{i-1} = Cb_{i-2}$ $Cb_{i-1} = Cp_{i-1}$	Continuation >>	Retaining >>	Smooth-shift >>	Rough-shift
	$Cb_i = Cb_{i-1}$	$Cb_i = Cb_{i-1}$	$Cb_i \neq Cb_{i-1}$	$Cb_i \neq Cb_{i-1}$
	$Cb_i = Cp_i$	$Cb_i \neq Cp_i$	$Cb_i = Cp_i$	$Cb_i \neq Cp_i$
	$Cb_i = Cp_{i-1}$	$Cb_i = Cp_{i-1}$	$Cb_i \neq Cp_{i-1}$	$Cb_i \neq Cp_{i-1}$
Retaining $Cb_{i-1} = Cb_{i-2}$ $Cb_{i-1} \neq Cp_{i-1}$	Smooth-shift >>	Rough-shift >>	Continuation >>	Retaining >>
	$Cb_i \neq Cb_{i-1}$	$Cb_i \neq Cb_{i-1}$	$Cb_i = Cb_{i-1}$	$Cb_i = Cb_{i-1}$
	$Cb_i = Cp_i$	$Cb_i \neq Cp_i$	$Cb_i = Cp_i$	$Cb_i \neq Cp_i$
	$Cb_i \neq Cp_{i-1}$	$Cb_i \neq Cp_{i-1}$	$Cb_i \neq Cp_{i-1}$	$Cb_i \neq Cp_{i-1}$
Smooth-shift $Cb_{i-1} \neq Cb_{i-2}$ $Cb_{i-1} = Cp_{i-1}$	Continuation >>	Retaining >>	Smooth-shift >>	Rough-shift
	$Cb_i = Cb_{i-1}$	$Cb_i = Cb_{i-1}$	$Cb_i \neq Cb_{i-1}$	$Cb_i \neq Cb_{i-1}$
	$Cb_i = Cp_i$	$Cb_i \neq Cp_i$	$Cb_i = Cp_i$	$Cb_i \neq Cp_i$
	$Cb_i = Cp_{i-1}$	$Cb_i = Cp_{i-1}$	$Cb_i \neq Cp_{i-1}$	$Cb_i \neq Cp_{i-1}$
Rough-shift $Cb_{i-1} \neq Cb_{i-2}$ $Cb_{i-1} \neq Cp_{i-1}$	Smooth-shift >>	Rough-shift >>	Continuation >>	Retaining >>
	$Cb_i \neq Cb_{i-1}$	$Cb_i \neq Cb_{i-1}$	$Cb_i = Cb_{i-1}$	$Cb_i = Cb_{i-1}$
	$Cb_i = Cp_i$	$Cb_i \neq Cp_i$	$Cb_i = Cp_i$	$Cb_i \neq Cp_i$
	$Cb_i \neq Cp_{i-1}$	$Cb_i \neq Cp_{i-1}$	$Cb_i \neq Cp_{i-1}$	$Cb_i \neq Cp_{i-1}$

Figure 15: Preferences of transition states

When the previous transition state is ‘continuation’, ‘continuation’ is preferred to ‘retaining’; ‘retaining’ is preferred to ‘smooth-shift’; and ‘smooth-shift’ is preferred to ‘rough-shift’. In this case, ‘continuation’ and ‘retaining’ are preferred over both shiftings because  $Cb_i = Cp_{i-1}$ . (Since  $Cb_i = Cb_{i-1}$  and  $Cb_{i-1} = Cp_{i-1}$ , thus,  $Cb_i = Cp_{i-1}$ ). And ‘continuation’ is preferred to ‘retaining’ because  $Cp_i = Cb_i$ . With the same reason, ‘smooth-shift’ is preferred to ‘rough-shift’.

When the previous transition state is ‘retaining’, ‘smooth-shift’ is preferred to ‘rough-shift’; ‘rough-shift’ is preferred to ‘continuation’; and ‘continuation’ is preferred to ‘retaining’. In this case, both shiftings are preferred to ‘continuation’ and

‘retaining’ because in ‘continuation’ and ‘retaining’  $Cb_i \neq Cp_{i-1}$ . (Since  $Cb_i = Cb_{i-1}$  and  $Cb_{i-1} \neq Cp_{i-1}$ , thus,  $Cb_i \neq Cp_{i-1}$ ). But in ‘smooth-shift’ and ‘rough-shift’,  $Cb_i$  may equal or differ from  $Cp_{i-1}$ . (We only know that  $Cb_i \neq Cb_{i-1}$  and  $Cb_{i-1} \neq Cp_{i-1}$ .) Therefore, continuation and retaining are less likely to be preferred to both shiftings. In addition, that  $Cb_{i-1} \neq Cp_{i-1}$  signals that the shifting of center is likely to occur. Thus, shiftings should be preferred to continuation and retaining. As for ‘continuation’ and ‘retaining’, ‘continuation’ is preferred to ‘retaining’ because of the second criterion. ‘Smooth-shift’ is also preferred to ‘rough-shift’ for the same reason.

When the previous transition is ‘smooth-shift’, the preferences of transition states are the same as those when the previous transition is ‘continuation’. The preferences can be explained as described above. And when the previous transition is ‘rough-shift’, the preferences of transition states are the same as those when the previous transition is ‘retaining’. The preferences of transition as stated in (108) will be used in the rest of our discussion as Rule 2.

(108)

RULE 2:

If  $\text{Cont}(U_{n-1}, U_n)$  or  $\text{Smooth-Shift}(U_{n-1}, U_n)$ , then

$\text{Cont}(U_n, U_{n+1}) > \text{Ret}(U_n, U_{n+1}) > \text{Smooth-Shift}(U_n, U_{n+1}) > \text{Rough-Shift}(U_n, U_{n+1})$

If  $\text{Ret}(U_{n-1}, U_n)$  or  $\text{Rough-Shift}(U_{n-1}, U_n)$  then

$\text{Smooth-Shift}(U_n, U_{n+1}) > \text{Rough-Shift}(U_n, U_{n+1}) > \text{Cont}(U_n, U_{n+1}) > \text{Ret}(U_n, U_{n+1})$

### 5.1.3 Cb establishment

Since a Cb is determined from the previous Cf, the Cb of the first utterance will be undetermined. This will result in a problem of undetermined transition state between the first and the second utterances. Since preferences of transition states are used in our focusing algorithms for selecting preferred referents of zero pronouns, the problem of undetermined transition state will likely affect the zero pronoun resolution.

Not only that the Cb is undefined for the first utterance, the Cb in some utterances may be undefined if no entity in the previous Cf is realized in the utterance. A solution to this problem is suggested by Brennan et al. (1987), as discussed in Passonneau (1998:334). Under the circumstance where no entity from the previous Cf is realized in the current utterance, Brennan et al. set the Cb to be the highest-ranked entity in the current Cf.

In this study, we follow Brennan et al (1987) in setting the Cb as the highest-ranked entity in the Cf when it cannot be determined from the previous Cf. Note that the process of setting up the Cb here is similar to Sidner's expected focus algorithm (Sidner 1983). Sidner uses the expected focus algorithm to determine the expected focus from the first utterance. The expected focus, then, will be confirmed or rejected later in the next step of the process.



We believe that setting up a Cb of the first utterance does not contradict the basic assumption about the Cb. Cb is determined from the immediately preceding utterance because it is believed that, usually, there is a cohesive link between two contiguous utterances. Cb and Cf are entities serving to link the previous utterance to the current one. The Cb of the current utterance is determined from the highest-ranked entity of the previous Cf because that entity is assumed to be the center of attention and should be continued from the preceding to the current utterance. When there is no preceding utterance to look for, we may assume that the most focused entity of the first utterance is the center of attention at that moment.

#### 5.1.4 Ranking of Cf

Cf is defined as an ordered list of discourse entities. However, how the entities in the Cf are ranked and what factors could affect the ranking are issues of active study. Cf ranking is believed to vary from language to language (Walker et al. 1990:2). The following are some examples of proposals for Cf ranking in English and Japanese:

(109)

Grosz et al. (1995:214) English

SUBJECT >> OBJECT(S) >> OTHER

Kameyama (1985:115) Japanese

TOPIC >> SUBJECT >> OBJECT(2) >> others

Walker et al. (1990:3) Japanese

TOPIC >> EMPATHY<sup>39</sup> >> SUBJ >> OBJ2 >> OBJ

---

<sup>39</sup> Empathy-loaded verbs are verbs that are sensitive to speakers' perspective. In Japanese, a verb can become empathy-loaded by being used as an auxiliary. (Walker et al. 1990:3).

In these proposals, the ranking of Cf is usually determined by grammatical function. However, it has been claimed that other factors should be considered as well. For example, Walker et al. (1990:3) show that an ‘empathy-loaded’ verb could affect centering in such a way that an entity realized as the ‘empathy-locus’ becomes more salient than an entity realized as the subject. Strube and Hahn (1996) argue that the ranking of Cf in free word order languages like German should be determined from functional relations rather than from grammatical relations.

In this study, we will follow most of the studies in Centering Theory in using grammatical relations for determining the ranking of Cf. Other factors that might affect the ranking of Cf in Thai will be left for further research. We hypothesize that the ordering of Cf in Thai is similar to that proposed for Japanese.<sup>40</sup> The Cf ranking for Thai used in this study is in (110).

(110) TOPIC >> SUBJ >> OBJ(s) >> OTHERS

In the absence of any obvious principle for ranking multiple objects (as in serial verb constructions), we shall assume that object entities are ranked based on linear order in the sentence. For example, in (111-112), obj1 is ranked before obj2, and obj2 is ranked before obj3.

---

<sup>40</sup> Thai is a topic prominent language.

- (111) khăw bòok sùdaa hâj ʔaw năŋsǎu hâj júphaa  
 he tell Suda(obj1) give take book(obj2) give Yupa(obj3)  
 ‘He told Suda to give Yupa a book.’

- (112) khăw hâj ɲəən sùdaa paj sǎu ʔaahǎan  
 he give money(obj1) Suda(obj2) go buy food(obj3)  
 ‘He gave Suda some money to buy food.’

Since an embedded clause is analyzed as a part of an utterance (see section 4.2.2), an utterance may have more than one subject/object in different syntactic levels. To code the difference in syntactic level, the number 1 or 2 will be added before the grammatical function of the embedded clause depending on the level of embedding. For example, in (113) ‘1subj’ is used to code the subject of an embedded clause at the first level and ‘2subj’ is for the subject at the second level of embedding.

- (113)  
 năŋsǎu thîi ʔaacaan khon thîi dɛɛŋ nápthǎu chôɔp ʔàan  
 khuu “The Little Prince”  
 book(subj) COMP teacher(1subj) CL COMP Daeng(2subj) admire like  
 read be “The Little Prince”  
 ‘The book that the teacher whom Daeng admires likes to read is “The Little Prince”’.

When an utterance has more than one entity with the same grammatical function, those entities are assumed here to be ranked with respect to linear order. In addition, embedded arguments are assumed to be less salient than arguments in the higher clause. With the coding of embedding clauses, below are the abbreviations of grammatical function and the preference order used in this study. ‘Other’ is used to

code grammatical functions other than subject, object, and topic. The ranking in (114) is a first approximation of Cf ranking for Thai. It should be noted that further research is necessary to confirm the ranking proposed here.

- (114)  
 top >> subj >> obj(s) >> 1subj(s) >> 1obj(s) >> 2subj >> 2obj(s) >> other

### **5.1.5 Zero pronoun resolution**

This section demonstrates how centering can be used for zero pronoun resolution in Thai. An algorithm for resolving zero pronouns in Thai can be implemented in the same way as Brennan et al.'s centering algorithm (1987). The algorithm has to keep track of the Cb and the Cf of each utterance. In an utterance with one zero, if we assume that its referent could be found in the immediately preceding utterance, according to Rule 1 (given in (105)), it is not possible for the Cb to be any other entity else besides the referent of the zero. In an utterance with more than one zero, if we assume that all the referents could be found in the immediately preceding utterance, one of the referents must be the Cb. The centering algorithm will try suggesting entities from  $Cf_{i-1}$  that are not yet referred to as the referents of these zeroes. There may be more than one possible assignment of referents to these zeroes. But the one that observes constraints and rules and the preferred transition state will be selected as the preferred interpretation. For example, following Rule 1, the referent of one zero must be the Cb (i.e. the highest ranked entity from the previous Cf). Following Rule 2

(given in (108)), continuation is preferred to retaining to shifting when the previous transition state is continuation or smooth-shift.

The following example demonstrates how the Centering Theory can be used for resolving zero pronouns in Thai.

(115)

- a. dɛɛŋ miɪ panhǎa nàk  
Daeng have problem severe  
'Daeng had a severe problem'.
- b. Z1 cʉŋ thoo paj prùksǎa dam  
Z1 thus call go consult Dam  
'Thus, (he) called to consult Dam'.
- c. thǎŋthǎŋthīi muâaɛwɛwnǐi Z2 phêɛŋ thamhâj dam ramkhaan  
though recently Z2 just make Dam annoyed  
'Though he just annoyed Dam recently'.
- d. chêen Z3 thoo paj hǎa Z4 tɔɔn tiisǐi  
for-example Z3 call go find Z4 at 4-am.  
'For example, (he) called (him) at 4 a.m.'.

- (116)
- a. Cb = 'Daeng' Cf = (Daeng, Problem1)  
Cp = Daeng
  - b. Cb = 'Daeng' Cf1 = (Daeng, Dam)  
Cp1 = Daeng Continue  
Cb = 'Problem1' Cf2 = (Problem1, Dam)  
Cp2 = Problem1 Smooth-shift
  - c. Cb = 'Daeng' Cf = (Daeng, Dam)  
Cp = Daeng Continue
  - d. Cb = 'Daeng' Cf1 = (Daeng, Dam)  
Cp1 = Daeng Continue  
Cf2 = (Dam, Daeng)  
Cp2 = Dam Retain

In example (115), there are two entities in (115a). Both are not yet referred to in (115b). Thus, one of them should be the referent of Z1 in (115b). If Z1 is resolved with

‘Daeng’, the transition state between (115a) and (115b) is continuation. If Z1 is resolved with ‘problem1’, the transition state between (115a) and (115b) is smooth-shift. Thus, the centering algorithm will suggest ‘Daeng’ as the preferred referent of Z1 because the continuation state is preferred to smooth-shift. In (115c), ‘Daeng’ is the only entity in the Cf of (115b) that is not yet referred to in (115c). Thus, it is selected as the preferred referent of Z2. In (115d), there are two zeroes. Since there are two entities in the Cf of (115c), there will be two possible interpretations: ‘Daeng called Dam at 4 am’ and ‘Dam called Daeng at 4 am’. For both interpretations, ‘Daeng’ is the Cb of (115d) because it is the highest ranked entity in (115c) that is realized in (115d). The transition state of the first interpretation is continuation while the transition state of the latter is retention. In this example, the first interpretation is preferred since the continuation state is preferred to retention state after continuation state.

Example (115) above shows that Centering Theory can be used for zero pronoun resolution in Thai. However, we have seen when an antecedent of a zero is in a distant utterance, Centering Theory cannot be used for zero pronoun resolution. To handle this problem, the centering model must be extended to work with the hierarchical structure of discourse. This will be discussed in the next section.

## **5.2 An extended centering model for Thai**

The motivation of the extension was discussed in section 3.4 and a model of extended centering was proposed in section 3.4.3. Details of constraints and rules of the extended model will be discussed in section 5.2.1. Since the extended model has to work with hierarchical structure of discourse, the unit in the constraints and rules of the extended model will be a discourse unit rather than a single utterance. A discourse unit can be either a single utterance or a group of utterances. When a unit is larger than a single utterance, it is necessary to discuss how entities in those utterances are combined and ordered as the Cf of the unit. The ordering of Cf of multiple utterances will be discussed in section 5.2.2. We will then show how Extended Centering can resolve zeroes with distant antecedents in section 5.2.3.

### **5.2.1 Constraints and rules**

Extended centering uses the same constraints and rules as defined in the existing centering in sections 5.1.1 and 5.1.2 with one exception, which is that the unit in our extended model can be either an utterance or a group of utterances. Following Centering Theory, we propose that a discourse unit contains one backward-looking center (Cb) and a set of forward-looking centers (Cf) which are all discourse entities in the discourse unit. Cb is assumed to be the entity that links the current unit with the

preceding one. Cf is an ordered list of entities in the unit. Constraints and rules for the extended centering are restated below:

(117)

Constraints:

For each discourse unit  $U_i$  in a discourse:

1. There is precisely one backward-looking center Cb.
2. Every element of forward centers list, Cf( $U_i$ ), must be realized in  $U_i$ .
3. The center, Cb( $U_i$ ), is the highest-ranked element of Cf( $U_{i-1}$ ) that is realized in  $U_i$ .

RULE 1: If any element of Cf( $U_i$ ) is realized by a zero pronoun in  $U_{i+1}$ , then the Cb( $U_{i+1}$ ) must be realized by a zero pronoun also.

RULE 2:

If Cont( $U_{n-1}, U_n$ ) or Smooth-Shift( $U_{n-1}, U_n$ ), then

Cont( $U_n, U_{n+1}$ ) > Ret( $U_n, U_{n+1}$ ) > Smooth-Shift( $U_n, U_{n+1}$ ) > Rough-Shift( $U_n, U_{n+1}$ )

If Ret( $U_{n-1}, U_n$ ) or Rough-Shift( $U_{n-1}, U_n$ ) then

Smooth-Shift( $U_n, U_{n+1}$ ) > Rough-Shift( $U_n, U_{n+1}$ ) > Cont( $U_n, U_{n+1}$ ) > Ret( $U_n, U_{n+1}$ )

Unlike the existing centering model, the constraints and rules in the extended model require a different interpretation of variables. Given that  $U_i$  is the current discourse unit,  $U_{i-1}$  is the immediately preceding discourse unit and  $U_{i+1}$  is the following discourse unit. The definition of precedence is defined as follow:

(118)

Precedence

$U_{i-1}$  precedes unit of  $U_i$  iff either

- a.  $U_{i-1}$  is the left adjacent unit of  $U_i$ .
- b.  $U_{i-1}$  is the left adjacent unit of  $U_k$  and  $U_i$  is the left most unit under  $U_k$

Left adjacency

$U_i$  is the left adjacent unit of  $U_j$  if  $U_i$  and  $U_j$  have the same parent,  $U_k$ , and there is no other unit between  $U_i$  and  $U_j$ .



It should be noted that the extended centering model proposed in this study is a minimal extension of Centering Theory. The modification only enables Centering Theory to work with the hierarchical structure of discourse. We have set the Cf of a discourse unit to be all entities in the unit, but it may not be necessary for the Cf to include all discourse entities. Further research may prove that more refinement of the model may be necessary.

As stated above, the main difference between the extended centering model and the existing centering model lies in the definition of the immediately preceding unit. While centering compares the current utterance with the immediately preceding utterance, extended centering compares the current unit with the immediately preceding unit. Therefore, when an utterance contains a zero pronoun, the scope in which the extended centering algorithm looks for the referents of zero pronouns will depend on the structure of the discourse. For example, in a discourse structure given in Figure 16, the preceding unit of (u7) is (u6), while the preceding unit of (u8) is (h).

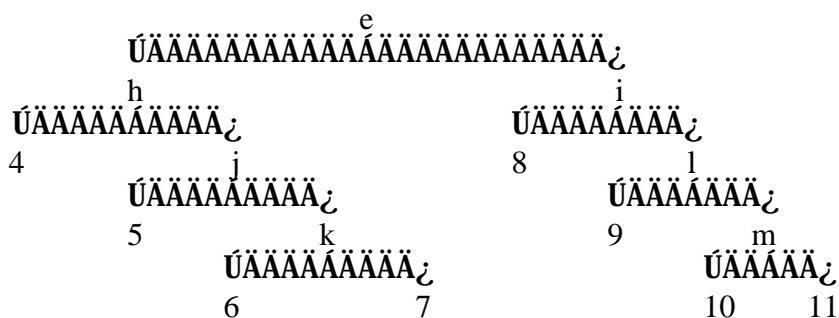


Figure 16: An example of discourse structure

### 5.2.2 Cf of multiple utterances

As stated in the last section, we assume that all entities in the discourse unit are accessible for resolving the referent of a zero pronoun in the next utterance, and the Cf of a discourse unit is defined as a combination of the Cf of its parts. How entities in this Cf are ranked and what factors affect the salience of discourse entities are open questions. For the purpose of this study, we rank entities in the Cf of a discourse unit on the basis of three factors: frequency, recency, and hierarchy. We hypothesize that frequency is relevant for the ranking of Cf because we think that an entity that is referred to more often in the discourse unit is likely to be more salient than other entities. Therefore, it should have a higher rank than other entities. In addition, we hypothesize that recency is also relevant for the ranking of Cf because we think that entities in the most recently unit are likely to be more salient than entities in farther units. Therefore, they should have a higher rank than other entities. We also hypothesize that entities in a unit at the lower level of hierarchical structure (embedded structure) are likely to be less salient than entities in a unit at the higher level. Thus, they should have a lower rank than other entities.

To verify whether the hypotheses above are applicable for the corpus of this study, we implemented a program to examine the three factors by considering zero pronouns and their antecedents in the corpus. For frequency factor, we counted the

number of zeroes whose referents are the most frequently referred entity in the preceding unit and the number of zeroes with the least frequently referred entity. We found that the more often the entities are referred to the higher the chance of their being the referents of zero pronouns. The probability that the most frequently referred entity is the referent of a zero pronoun is 0.72, while the probability that the least frequently referred entity is the referent of a zero pronoun is only 0.24. This is also true with the entities that are referred to more recently, though, the latter is not immediately obvious. The probability of finding the antecedent at the most recent unit is 0.55 while the probability of finding the antecedent at the farthest unit is 0.44. On the contrary, the hierarchy factor does not have an effect as hypothesized. The probability of finding the antecedent at the lowest level of structure is 0.71, higher than the probability of finding the antecedent at the highest level, which is 0.38.

	Frequency	Recency	Hierarchy
highest	0.72	0.55	0.38
lowest	0.24	0.44	0.71

Table 8: Probability of referents found with respect to three factors

It is evident from Table 8 that frequency should be the most important factor in ranking entities in the Cf and recency should be the second. As for hierarchy effect, since the results rejects our hypothesis, it might be possible to reverse the hypothesis so that entities in the utterance at the lowest level of structure is more salient than others. However, the greater chance of finding antecedents at the lowest level of structure than

at the highest level may result from the greater number of utterances at the lowest level than at the highest level. Moreover, when implementing the extended centering algorithm with the preference of entities in the utterance at the lower level over others at the higher level, the results were worse than the one which prefers entities at the higher level (see section 6.5.4). Therefore, we will not use the hierarchy effect for ranking entities in the Cf in this study.

To show how the Cf of a higher level unit is constructed based on the hypotheses of frequency and recency, consider example (119) which has four utterances (u8-u11). The structure of discourse is shown in Figure 17.

(119) Text from News1

- #8 kooránii thîi phráthamkoosăacaan rǎu  
thâanphúttháthâatphíkku hēē wátsuăanmôokphálaaraam  
ʔamphēechajjaa caŋwàtsùrâatthaanii ʔaaphâat nàk kòon cà  
thǔŋ wan khláaj wankèet troŋ kàp wanthîi 27  
phríksàphaakhom níi  
a-case COMP Pra-Thamkosajarn or Bhutathat-Bhikhu<sub>Z6</sub> ('Bhikhu') of  
Wat-Suanmokplaram Chaiya-county Suratthanee-province sick heavy  
before will reach day like birthday exactly with date 27 May this  
'Pra Thamkosajarn, Bhutathat Bhikhu of Wat Suanmokplaram at Chaiya  
county Suratthanee, was sick before his birthday in May 27'
- #9 lé khánáphêet rooŋphájaabaansùrâatthaanii nímon [Z6] paj  
ráksăatuaa jùu thîi rooŋphájaabaan  
and doctors<sub>Z7</sub> ('Doctor1') Suratthanee-hospital invite [Z6]<sub>Z8</sub> go treat  
stay at hospital  
'and doctors in Suratthanee hospital has taken (him) for a treatment in the  
hospital'
- #10 sǔŋ lǎŋcàak thîi [Z7] truàatʔaakaan [Z8] léēw  
CONJ after COMP [Z7]<sub>Z9</sub> examine [Z8]<sub>Z10</sub> ASP  
'After (the doctors) have examined (him)'

#11 [Z9] kô dâj hâj [Z10] noonphák ráksăatuaa jùu naj hòon  
 ʔajsiijuu tântèe chăwtrùu wanthîi 25 phríksàphaakhom  
 thîiphàanmaa  
 [Z9] then ASP let [Z10] rest cure stay in room ICU since morning  
 date 25 May last

‘(they) asked (him) to stay in an ICU room since the morning of May 25’  
 Before his birthday in May 27, Pra Thamkosajarn or Than Bhutathat  
 Bhikhu, the abbot of Wat Suanmokplaram at Chaiya county, Suratthanee  
 province, was sent to Suratthanee hospital. After physical examination,  
 the doctors have asked him to stay at the hospital. He is now in the I.C.U  
 room since the morning of May 25.

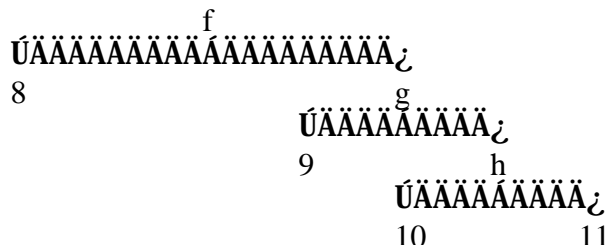


Figure 17: A discourse structure of example (119)

Figure 18 below shows Cfs of utterances (u8-u11) and Cfs of discourse units (h), (g), and (f). In this example, Z6, Z8, and Z10 are resolved with ‘Bhikhu’; Z7 and Z9 are resolved with ‘Doctor1’. The square bracket indicates the frequency of an entity in the unit. For example, in Cf(h), which is a combination of Cf(u10) and Cf(u11), ‘Doctor1’ and ‘Bhikhu’ have the frequency of two while the rest has the frequency of one. Since frequencies of ‘Doctor1’ and ‘Bhikhu’ are equal, their ranks remain the same as in (u11). Cf(g) is a combination of Cf (u9) and Cf (h). Again, the rank of ‘Doctor1’ and ‘Bhikhu’ is preserved unchanged. ‘Hospital1’ has a lower rank than ‘ICURoom1’ and ‘May-25’ because (h) is more recent than (u9).

In Cf(f), which is the combination of Cf(u8) and Cf(g), ‘Bhikhu’ has the highest rank because it has the highest frequency. ‘Doctor1’ is ranked next, because its frequency is three. The rest has frequency of one. ‘WatSuanMok’, ‘Chaiya-County’, ‘Suratthanee-Prov,’ and ‘May-27’ have lower rank than ‘ICURoom1’, ‘May-25’, and ‘Hospital1’, because they are in (u8), which is more distant than (g).

f. (Bhikhu[4], Doctor1[3], ICURoom1[1], May-25[1], Hospital1[1],  
 WatSuanMok[1], Chaiya-County[1], Suratthanee-Prov[1], May-27[1])  
 ÚAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA¿  
 u8: (Bhikhu[1], WatSuanMok[1], g: (Doctor1[3], Bhikhu[3],  
 Chaiya-County[1], Suratthanee-Prov[1], ICURoom1[1], May-25[1],  
 May-27[1]) Hospital1[1])  
 ÚAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA¿  
 u9: (Doctor1[1], Bhikhu[1], h: (Doctor1[2], Bhikhu[2],  
 Hospital1[1]) ICURoom1[1], May-25[1])  
 ÚAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA¿  
 u10: (Doctor1[1], Bhikhu[1]) u11: (Doctor1[1], Bhikhu[1],  
 ICURoom1[1], May-25[1])

Figure 18: Cfs of discourse units in Example (119)

The combination of entities in a Cf can be formalized as follows:

(120)

For  $U_i$  consisting of units  $U_1, U_2, \dots, U_n$ ,  $Cf(U_i)$  is a union list of  $Cf(U_1)$ ,  $Cf(U_2)$ , ..., and  $Cf(U_n)$  such that

$E_a$  and  $E_b$  are entities in  $Cf(U_i)$ ,  $E_a$  is ranked before  $E_b$  iff

Frequency( $E_a$ ) > Frequency( $E_b$ ) , or

Frequency( $E_a$ ) = Frequency( $E_b$ ) , and  $E_a$  is an entity in  $Cf(U_i)$

and  $E_b$  is an entity in  $Cf(U_k)$ , and  $U_k$  precedes  $U_i$ .

### 5.2.3 Zero pronoun resolution

In section 5.1.5, we state that the existing centering model cannot account for zeroes with distant antecedents. In this section, we demonstrate how extended centering can be used to resolve these zero pronouns.

An extended centering model can be implemented in a similar way as the existing centering model. But the extended centering algorithm has to deal with two additional aspects. First, the unit in the extended model can be larger than a single utterance and the immediately preceding unit is not necessary an immediately preceding utterance. Second, when a unit has more than one utterance, the extended centering algorithm has to deal with the combination of Cfs. Otherwise, the extended centering model can be implemented in the same way as the existing model.

The algorithm must keep track of the Cb and the Cf of each unit. In an utterance with one zero, if we assume that its referent could be found in the immediately preceding unit, according to Rule 1, its referent must be the Cb. In an utterance with more than one zero, if we assume that all the referents could be found in the immediately preceding unit, one of the referents must be the Cb. The algorithm will suggest entities that are not yet referred to as the referents of zeroes. There may be more than one possible interpretation. But the one that observes constraints and rules and the preferred transition state will be selected as the preferred interpretation.

To demonstrate how the extended centering model can be used to resolve zero pronouns, consider example (121), which is a discourse from text Health1. Its discourse structure is shown in Figure 19.

(121) Text from Health1

#23 daŋnán thâa suàan tàaŋtàaŋ khǒoŋ hũu dâjráp kaan  
kràthópkràthuaan

therefore if part any of ear<sub>Z11</sub> receive NOM impact<sub>Z17</sub>  
‘Therefore, if any part of ears receives an impact’

#24 rǎu [Z11] dâjráp chuáarôok

or [Z11]<sub>Z12</sub> receive germ  
‘or (it) receives germ’

#25 chêen [Z12] doon kràthêek rɛɛŋrɛɛŋ

such-as [Z12]<sub>Z13</sub> get strike strong  
‘For example, (it) is struck’

#26 [Z13] dâjjin siãaŋ daŋ mâak

[Z13] hear voice loud very  
‘(It) is imposed to very loud noise’

#27 mii nám khâw hũu

there-is water enter ear<sub>Z14</sub>  
‘Water enters ears’

#28 [Z14] dâjráp chuáarôok

[Z14] receive germ  
‘(Ears) receive germ’

#29 phró [Z15] khé hũu duâaj khruâaŋmũu sòkkàpròk

because [Z15] pick ear with instrument dirty  
‘because (we) pick ears with dirty instrument’

#30 [Z16] khé hũu rɛɛŋ kəən paj

[Z16] pick ear strong over ASP  
‘(we) pick ears too hard’

#31 [Z17] ʔàat thamhâj kəət rôok hũu dâj

[Z17] may cause occur disease ear ASP  
‘(The impact like this) could damage ears’

Therefore, if any part of the ear is damaged, we might have an ear disorder. The ear damage can be caused by impact or diseases, such as



being struck, hearing an extremely loud noise, getting wet, and being infected by diseases from dirty ear sticks.

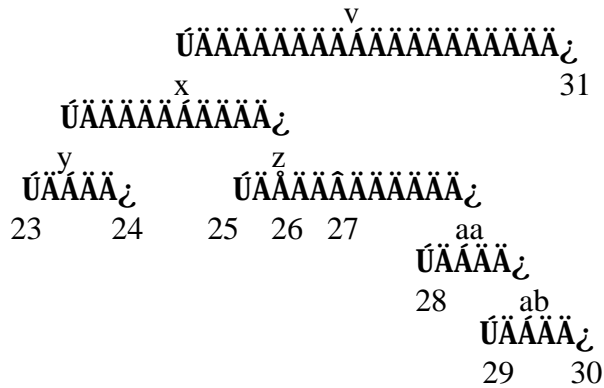


Figure 19: A discourse structure of example (121)

In example (121), Z17 cannot be resolved by the existing centering algorithm because its antecedent is not in the immediately preceding utterance (u30). Rather, its antecedent is in a distant utterance (u23). The extended centering algorithm can resolve this zero because it will look for an antecedent of Z17 in (uu), which is composed of (u23-u30).

(122) below is the list of the Cb and Cf of each unit created by the extended centering algorithm. When an utterance contains a zero pronoun, the extended centering algorithm will suggest each entity from the immediately preceding unit that is not yet referred to in the utterance as the referent of the zero. There can be more than one assignment of referents to the zero. Each interpretation will be ranked with respect to the preference of transition state. The highest ranked interpretation is the preferred

one. For example, (u24) has one zero (Z11) and one entity 'Germ'. According to the structure in Figure 19, the immediately preceding unit for (u24) is (u23). There are two entities in (u23) that are not yet referred to in (u24), 'Ear' and 'Impact1'. Thus, there would be two possible interpretations for Z11. The first interpretation in which Z11 is resolved with 'Ear' represents a continuation state, while the second interpretation in which Z11 is resolved with 'Impact1' represents a smooth-shift state. Since continuation state is preferred to smooth-shift in this case, the extended centering algorithm will suggest 'Ear' as the referent of Z11. And that referent is not rejected by the inference mechanism.

(122)

23 Cb = Ear Cf = (Ear, Impact1)

24 If Z11 is resolved with 'Ear',

Cb = Ear Cf = (Ear, Germ)

Continue

If Z11 is resolved with 'Impact1',

Cb = Impact1 Cf = (Impact1, Germ)

Smooth-shift

Since continuation is preferred to smooth-shift, Z11 is resolved with 'Ear'.

y = 23 + 24

Cb = 'Ear' Cf = (Ear[2], Germ[1], Impact1[1])

Continue

25 If Z12 is resolved with 'Ear'

Cb = 'Ear' Cf = (Ear)

Continue

The referent suggested is accepted by the inference mechanism.

26 If Z13 is resolved with 'Ear'

Cb = 'Ear' Cf = (Ear, LoundNoise)

Continue

27 Cb = 'Ear' Cf = (Water, Ear)

Retain

28 If Z14 is resolved with 'Water'

Cb = 'Water' Cf = (Water, Germ)

Smooth-shift

But the referent is rejected by the inference mechanism.

If Z14 is resolved with 'Ear'

Cb = 'Ear' Cf = (Ear, Germ)

Continue

The referent suggested is accepted by the inference mechanism.

- 29 Cb = Ear Cf = (Ear, Instrument1, Unidentified) Continue  
(The referent of Z15 is an 'unidentified' entity. Thus, it is excluded from the algorithm.)
- 30 Cb = Ear Cf = (Ear:obj, Unidentified:subj) Continue  
(The referent of Z16 is an 'unidentified' entity. Thus, it is excluded from the algorithm.)
- ab = 29 + 30  
Cb = 'Ear' Cf = (Ear[2], Instrument1[1], Unidentified[2]) Continue
- aa = 28 + z  
Cb = 'Ear' Cf = (Ear[3], Instrument1[1], Germ[1], Unidentified[2]) Continue
- z = 25 + 26 + 27 + aa  
Cb = 'Ear' Cf = (Ear[6], Instrument1[1], Germ[1], Water[1], LoudNoise[1], Unidentified[2]) Continue
- x = y + z  
Cb = 'Ear' Cf = (Ear[8], Germ[2], Instrument1[1], Water[1], LoudNoise[1], Impact1[1], Unidentified[2]) Continue
- 31 If Z17 is resolved with 'Ear',  
Cb = 'Ear' Cf = (Ear, Disease) Continue  
But the referent is rejected by the inference mechanism.  
If Z17 is resolved with 'Germ',  
Cb = 'Germ' Cf = (Germ, Disease) Smooth-shift  
But the referent is rejected by the inference mechanism.  
...  
If Z17 is resolved with 'Impact1',  
Cb = 'Impact1' Cf = (Impact1, Disease) Smooth-shift  
The referent suggested is accepted by the inference mechanism.

For unit (y) which is composed of (u23) and (u24), the entity with the highest frequency will be the first entity in the Cf. The square bracket indicates frequency of entities in that unit. In this example, Z12, Z13, and Z14 can be resolved in the same way as Z11. The immediately preceding unit for (u25), (u26), and (u28) is unit (y), (u25), and (u27) respectively. The first entity suggested by the extended centering

model is accepted as the referent of Z12 and Z13. In (u27), the first entity suggested is rejected and Z14 is resolved with the next preferred entity.

For Z17 in (u31), the immediately preceding unit of (u31) is (x). There are six entities in Cf(x) (excluding an ‘unidentified’ referent) which is not yet referred to in (u31). ‘Ear’ is the highest ranked in Cf(x) since its frequency is higher than other entities. ‘Germ’ is ranked next. Extended centering algorithm will suggest these entities in an ordering list as the probable referents of Z17. But the first five entities will be rejected by inference mechanism. The ‘correct’ referent, ‘Impact1’, will be found at the sixth try. Though the extended centering algorithm cannot suggest the ‘correct’ referent at the first try, it limits the scope of referents for zero pronoun resolution. The ‘correct’ referent of Z17 is included in this list or scope.

### **5.3 Conclusion**

In this chapter we have discussed two models of centering for Thai. One is a model as used in other centering literature (Walker, Iida, and Cote 1990, 1994, Iida 1997, and Eugenio 1990, 1996, 1997). The other is the model we have extended by incorporating the hierarchical structure of discourse. The latter is capable of resolving zero pronouns whose antecedents are in distant utterances. In the next chapter, the two models will be tested on the same corpus. The results when applying the existing centering model will reveal cases where antecedents of zeroes are not in the preceding

utterance. Whether these zeroes can be resolved when the hierarchical structure of discourse is considered will be verified by applying the extended centering model to the same corpus.