

Chapter 6

Comparison of Centering Algorithms

This chapter presents a comparison of the extended centering algorithm and the existing centering algorithm. The comparison is based on the tests which are conducted on our corpus of twenty Thai texts to verify whether hierarchical structure of discourse contributes to zero pronoun resolution in Thai. Although the hypothesis can be verified by manually analyzing the corpus, we choose to implement the centering algorithms to work automatically on the corpus. This would increase reliability and consistency of analysis. The results when applying the existing centering algorithm will provide us zero pronouns that do not have antecedents in the preceding utterance. When the extended centering model is applied to the same corpus, we will see whether these zeroes can be resolved. We will discuss whether these zeroes are resolved on the basis of hierarchical structure of discourse.

6.1 Testing centering algorithms

The two centering models were implemented within a simulation program. The simulation program provided necessary settings as if the centering algorithms were implemented in an NLP system. It accepted a discourse structure analyzed by each subject as an input for the extended centering algorithm. A list of discourse entities in each utterance was prepared as an input for both the extended and the existing

centering algorithms. These entities were assumed here to be constructed while a parser was processing an utterance. Both centering algorithms used these input entities to determine a preferred referent for a particular zero pronoun. The referents suggested by the centering algorithms would then be accepted or rejected by another program which would include such process such as selectional restriction. However, in our simulation program, we simply used the list of ‘correct’ referents to accept or reject the suggested referents. This list of ‘correct’ referents was taken from the majority agreement on the referents of zero pronouns of our subjects (see section 4.3.1). The matching of suggested referents against the ‘correct’ referent began with the first preferred referent. If this matching fails, the matching process moved on to the next suggested referent.

6.1.1 Input

To test centering algorithms on each text, three input files had to be manually prepared for the simulation program. These are the entity list file, the discourse structure file, and the referent list file.

6.1.1.1 Entity list file

In the entity list file, each line represents a list of discourse entities referred to in an utterance. This information is supposed to be constructed by a parser which processes an utterance. In this study, the list of entities for each utterance was manually prepared. Each discourse entity was coded with two parts: an identifying name and its

grammatical function. It should be noted that the coding of a discourse entity here is aimed at the convenience of the test. It is not a semantic representation of a discourse entity used in an NLP system. Grammatical function is included because it is used for the ranking of Cf in this study. A zero pronoun was coded as Zn where n is the number of that zero in the texts. Each line in an entity list file has three parts: the number of the utterance, the number of entities being referred to in an utterance and list of the entities in the utterance. An example of data in an entity list file is shown in (123).

(123)

```

1, 2, Visitor1:subj, Bhikhu:obj
2, 3, Doctor1:subj, Z1:obj, Z2:obj2
3, 2, Z3:subj, ICURoom1:pp
4, 1, Z4:subj
5, 1, Bhikhu:subj
6, 2, Doctor1:subj, Z5:obj
8, 5, Bhikhu:subj, WatSuanMok:pp, Chaiya-County:pp, Suratthane-
Prov:pp, May-27:pp
9, 3, Doctor1:subj, Hospital1:pp, Z6:obj
10, 2, Z7:subj, Z8:obj
11, 4, Z9:subj, Z10:obj, ICURoom1:pp, May-25:pp

```

In (123), each entity in a list has two parts: an identifying name and a grammatical function. Duplicated entities are coded only once in the line. A discourse entity is a conceptual representation⁴¹, so the number of entities in each utterance may not equal the number of noun phrases in the utterance. This is because the same entity may be referred to by more than one noun phrase in an utterance. For example, in (124)

⁴¹ Though an abstract referent could be represented as a conceptual object too, it was excluded from this study.

below two noun phrases, Pra-Thamkosajarn and Bhutathat-Bhikhu, are used in the same utterance to refer to the same discourse entity. The former is the royal title granted to the monk while the latter is his pseudonym. In (125), a noun phrase ‘lawyer’ /nákkòtmăaj/ does not evoke a new discourse entity in the utterance. It simply refers to Balladur as a lawyer. So, this utterance has only one discourse entity, ‘Balladur’.

(124) Text from News1

#8 kooránii thîi phráthamkoosăacaan rŭu
thâanphúttháthâatphíkku hēēŋ wátsuāanmôokphálaaraam
ʔamphœchajjaa caŋwàtsùràatthaanii ʔaaphâat nàk kòon cà thŭŋ
wan khláaj wankèet troŋ kàp wanthîi 27 phríksàphaakhom níi
case COMP Pra-Thamkosajarn or Bhutathat-Bhikhu (‘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’

(125) Text from Article2

#9 doojnuáathéē khǒŋ tuaabanlaadœ léew khăw pēn nákkòtmăaj
thîi mii khwaamsùphâap
by-nature of Balladur ASP he be lawyer COMP have politeness
‘By his nature, Balladur is a lawyer who is polite’

6.1.1.2 Discourse structure file

A discourse structure file contains information about the hierarchical structure of a discourse. The structure of discourse in Figure 20 when coded in the discourse structure file looks like rewriting rules as shown in (126).

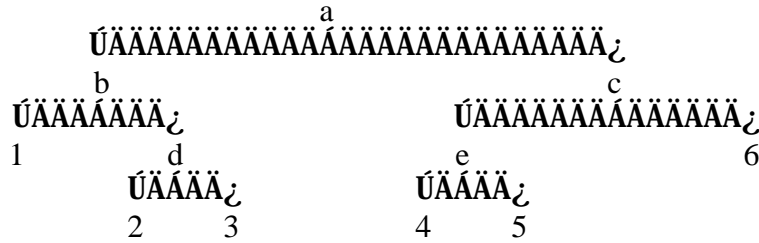


Figure 20: An example of the discourse structure

(126)

a -> b+c
 b -> 1+d
 d -> 2+3
 c -> e+6
 e -> 4+5

Items on the left side of the arrow represent the mother nodes, while items on the right side of the arrow represent child nodes. Numbers in these rules represent utterances while letters represent discourse units. The simulation program will read this input file and create a hierarchical structure of the discourse for the extended centering algorithm.

6.1.1.3 Referent list file

A referent list file contains a pair consisting of a zero pronoun number and its referent. The list was obtained from the analyses of five subjects. The 'correct' referent was taken from the majority agreement (three out of five subjects). If the majority failed to identify the referent of any zero pronoun, 'conflicting' was marked for that zero. Example (127) below is an example of a referent list file. Each line has three

elements. The first represents the number of zero pronouns in the discourse. The second one indicates the probability of agreement among subjects. The last one records the referent of zero pronoun as identified by the majority of subjects.

(127)

- 1, 1.00, Visitor1
- 2, 1.00, Bhikhu
- 3, 1.00, Bhikhu
- 4, 1.00, Bhikhu
- 5, 1.00, Bhikhu
- 6, 1.00, Bhikhu
- 7, 1.00, Doctor1
- 8, 1.00, Bhikhu
- 9, 1.00, Doctor1

6.1.2 Scope of the test

Not all zeroes marked on the corpus were handled by the centering algorithms. First to be excluded were zeroes which the majority of subjects failed to agree on their referents. Next were deictic zeroes, zeroes with abstract referents, and zeroes whose referents are analyzed as ‘unidentified’ referents. In addition, zeroes in embedded clauses were excluded from the test because the centering algorithms are not designed to resolve zeroes whose referents are in the same utterance. Table 9 below shows the number of different types of zeroes in the corpus which were excluded from the test. Of the 719 zeroes marked on the corpus, referents of 40 could not be agreed on by the majority of subjects. 82 zeroes were analyzed as having ‘unidentified’ referents; 133

were used as deictic zeroes; 34 were in embedded clauses; and 12 were used to refer to abstract referents. Therefore, only 418 zeroes were tested by the centering algorithms.

Zeroes which are excluded from this study were coded differently so that the algorithms could recognize them. In an entity list file, unidentified zeroes were coded as *UZn*; deictic zeroes as *DZn*; zeroes in embedding clauses as *IZn*; and zeroes with abstract referents as *AZn*. When these coded zeroes were recognized by the algorithms, they were simply replaced by the ‘correct’ referent by the simulation program. Conflicting zeroes whose referents cannot be agreed upon by the subjects had to be left out because we do not have any ‘correct’ referent for them. These zeroes were marked as *CZn* in an entity list file. They then were eliminated from the entity list when found by the simulation program.

	No of Zero	Conflict (CZn)	Unidentified (UZn)	Deictic (DZn)	Embedding (IZn)	Abstract Zero (AZn)	Normal (Zn)
Total	719	40	82	133	34	12	418
Percent	100.00%	5.56%	11.40%	18.50%	4.73%	1.67%	58.14%

Table 9: Zero pronouns in the corpus

6.1.3 Simulation of centering algorithms

An overview of the simulation is as follows. First, the simulation program reads a referent list file and a discourse structure file to set up an initial environment for the centering process. The list of ‘correct’ referents for zero pronouns in a text is created from the referent list file, and the hierarchical structure of the discourse is created from

the discourse structure file. Then, each utterance is processed. Entities referred to in each utterance is read from the entity list file. These entities are ordered and sent to either the existing centering or the extended centering algorithm. The ordered list of entities represents the Cf of the utterance. If an utterance contains any zero pronouns, the zero number Z_n will appear in the Cf. The centering algorithms' task is to suggest the referents of these zeroes. The simulation program then compares the suggested referents with the 'correct' referents. If they do not match the 'correct' referents, the simulation program rejects the suggested referents and the centering algorithms looks for the next preferred referents. When the 'correct' referent is found, the simulation program substitutes the coding Z_n with the identifying name of the referent. If none of the possible interpretations matches the 'correct' referents, it means that the centering algorithms fail to suggest any preferred referents.

The algorithm of the simulation program is shown below:

(128)

Simulation program:

- Initialization.
 - Read a referent list file.
 - Read a discourse structure file and construct the hierarchical structure.
- Do from the first utterance to the last one
 - Read entities of an utterance from an entity list file.
 - Rank entities in the utterance.
 - Call the extended centering or the existing centering algorithm.
 - Proceed to the next utterance.

Loop

6.2 The two centering algorithms

The extended centering algorithm and the existing centering one are different mainly in that the extended centering uses information of discourse structure to determine the Cb, Cf, and transition state of a discourse unit. Details of the existing centering and the extended centering algorithms will be provided first. These are followed by description of other routines called by the two centerings. These routines include determining Cb, determining transition state, generating all possible interpretations, ranking all possible interpretations, and constructing Cf of multiple utterances.

6.2.1 Existing centering algorithm

The existing centering first looks for zero pronouns in an utterance. If an utterance (U_i) does not have any zero pronouns, the program will proceed to determine the Cb and the transition state. If an utterance has one or more zero pronouns, its preferred referent will be selected from the potential referent list, which is an ordered list of entities from U_{i-1} that are not retained in $Cf(U_i)$.⁴² A preferred referent is taken from the potential referent list in order. If the number of zeroes in the utterance is r and

⁴² The potential referent list is just a temporary structure created during the processing. It is used for the purpose of implementation. It does not directly relate to any data type proposed in the centering models.

the number of possible referents is n , the number of possible interpretations for the zeroes in the utterance will equal to $n!/(n-r)!$ ⁴³.

For example, if there is one zero in the utterance, the number of possible interpretations for that zero will be n where n is the number of possible referents. When there are two zeroes in the utterance, the number of possible interpretations for the two zeroes will be $n \times (n-1)$ where n is the number of possible referents.

The Cb and transition state of each interpretation are determined and used later for ranking the interpretations (see ‘Ranking all possible interpretation’ routine in section 6.2.5). The preferred interpretation is selected according to the ranking of interpretations until the suggested referents match the ‘correct’ referents. If none of the interpretations matches the ‘correct’ referents, the centering algorithm fails.

After zero pronouns are resolved, the program proceeds to update the Cb and the transition state between U_i and U_{i-1} . Below is the algorithm of the existing centering.

⁴³ $n! = n \times (n-1) \times (n-2) \times \dots \times 1$.

(129)

Existing centering:

- Look for the number of zeros.
- If the number of zeros = 0, then
 - Determine Cb; Determine the transition state
- If the number of zeros > 0, then
 - Get potential referents from U_{i-1} .
 - Generate all possible referents for each zero.
 - Determine Cb and determine the transition state for each interpretation.
 - Ranking all possible interpretations.
 - Do until the correct referents are found.
 - Select the highest-ranked interpretation.
 - If not accepted, select the next lower rank interpretation.
 - If no interpretation is accepted, mark the failure.
- Update Cb and the transition state.

6.2.2 Extended centering algorithm

There are two differences between the extended centering algorithm and the existing centering one. First, U_{i-1} in the extended centering can be either the immediately preceding utterance or the immediately preceding discourse unit. It is determined from the hierarchical structure of the discourse as discussed before. Second, the extended centering algorithm has to determine whether the current utterance is the right most node. If the utterance is the right most node in the discourse structure, the program will proceed to construct the Cf of the mother node (see ‘Constructing Cf of multiple utterances’ routine in section 6.2.6) and determine its Cb and transition state. The algorithm of the extended centering is presented below.

(130)

Extended centering:

- Look for the number of zeros.
- If the number of zeros = 0, then
 - Determine Cb; Determine transition state.
- If the number of zeros > 0, then
 - Get potential referents from U_{i-1} .
 - Generate all possible referents for each zero.
 - Determine Cb and determine the transition state for each interpretation.
 - Ranking all possible interpretations.
 - Do until the correct referents are found.
 - Select the highest-ranked interpretation.
 - If not accepted, select the next lower rank interpretation.
 - If no interpretation is accepted, mark the failure.
- Update Cb and the transition state.
- If the unit is the right most unit,
 - Construct the Cf of the mother node.

6.2.3 Algorithm for determining the Cb

This routine determines the Cb of U_i . If an utterance contains any zero pronouns, this routine will be called after the zero pronouns are resolved and their referents are put in $Cf(U_i)$. $Cb(U_i)$ is the first entity in $Cf(U_{i-1})$ that is retained in $Cf(U_i)$. If the Cb cannot be determined from $Cf(U_{i-1})$, it will be set by default as the first entity in $Cf(U_i)$ (as discussed in section 5.1.3). The Cb cannot be determined for two possible reasons: First, U_{i-1} unit does not exist. This is, U_i is the first unit in the structure. Or second, U_{i-1} does not contain any entity in common with U_i . In addition, an

‘unidentified’ referent will never be selected as the Cb since it never serves as a link between U_{i-1} and U_i . Below is the algorithm for determining the Cb.

(131)

Determining Cb:

- Set Cb = the first entity in $Cf(U_{i-1})$ that is also in $Cf(U_i)$.
- If not found or what is found is an ‘unidentified’ entity,
- set Cb = the first entity in $Cf(U_i)$.

6.2.4 Algorithm for determining a transition state

This routine is used to determine the transition state between U_{i-1} and U_i . The transition state is determined after $Cb(U_i)$ has been determined. This is done by comparing $Cb(U_i)$ and $Cb(U_{i-1})$, and $Cb(U_i)$ and $Cp(U_i)$. The algorithm for determining a transition state is as follows:

(132)

Determining transition state:

- If $Cb_i = Cb_{i-1}$ and $Cb_i = Cp_i$, then set transition state = ‘Continuation’.
- If $Cb_i = Cb_{i-1}$ and $Cb_i \neq Cp_i$, then Set transition state = ‘Retaining’.
- If $Cb_i \neq Cb_{i-1}$ and $Cb_i = Cp_i$, then Set transition state = ‘Smooth-shift’.
- If $Cb_i \neq Cb_{i-1}$ and $Cb_i \neq Cp_i$, then Set transition state = ‘Rough-shift’.

6.2.5 Algorithm for generating and ranking all possible interpretations

The two routines in this section are used by both centering algorithms. First, all potential referents from U_{i-1} will be selected from $Cf(U_{i-1})$. Next, all possible referents for each zero in U_i will be generated. The number of possible assignments of referents to zeroes is $n \times (n-1) \times \dots \times (n-r-1)$ where n is the number of the potential referents and

r is the number of zeroes. For example, if the number of possible referents is 4 and the number of zero is 2, the number of possible assignments is $4 \times 3 = 12$. These possible interpretations will be ranked in order by two factors: discourse salience and preferences of transition states.

The first one is succeeded by generating interpretations based on the order of entities in the previous Cf. Interpretations in which the first entity from $Cf(U_{i-1})$ is assigned will have a higher rank than interpretations in which successively entities in $Cf(U_{i-1})$ are assigned. Next, these interpretations are ranked by the preference of transition state. The ordering depends on the previous transition state. Referents assigned to zeroes from the highest-ranked interpretation are regarded as preferred referents. If all the referents in that interpretation do not match the ‘correct’ referents, the next interpretation will be suggested in order until all the referents of zeroes match the ‘correct’ referents.

(133)

Generating all possible interpretations:

- Get a list of potential referents from U_{i-1} .
- Create all possible interpretations by replacing each zero with an entity from the potential referent list.
(The number of possibilities equals to $n \times (n-1) \times \dots \times (n-r-1)$ when n is the number of entities in the potential referent list and r is the number of zeroes.)
- Determine the Cb of each interpretation
- Determine the transition state for each interpretation.

(134)

Ranking all possible interpretations:

- If the transition state of U_{i-1} is continuation,
 - Sort possible interpretations by the preference ‘continuation >> retaining >> smooth-shift >> rough-shift’.
- If the transition state of U_{i-1} is retaining,
 - Sort possible interpretations by the preference ‘smooth-shift >> rough-shift >> continuation >> retaining’.
- If the transition state of U_{i-1} is smooth-shift,
 - Sort possible interpretations by the preference ‘continuation >> retaining >> smooth-shift >> rough-shift’.
- If the transition state of U_{i-1} is rough-shift,
 - Sort possible interpretations by the preference ‘smooth-shift >> rough-shift >> continuation >> retaining’.
- If the transition state of U_{i-1} is undetermined,
 - Sort possible interpretations by the preference ‘continuation >> retaining >> smooth-shift >> rough-shift’.

6.2.6 Algorithm for constructing Cf of multiple utterances

This routine constructs Cf and determine a transition state of a discourse unit. It is used only by the extended centering algorithm. $Cf(U_i)$ is defined as the combination of the child nodes’ Cfs. In short, entities from each Cf will be merged into a new list of entities. Duplicated entities will be removed from the list. Two factors, which are frequency and recency, are used in ranking the Cf here. Since frequency is the most important factor, it will be applied after ordering by recency. Ordering by recency is done by moving entities referred to in the farther utterance to the back of the list. Next, entities in the list is ordered again by frequency of entities. An entity which is referred to more often will be moved to the front of the list. The final list is saved as the Cf of

the mother node. If the mother node is also the right most node, the routine will be applied recursively. The algorithm for constructing the Cf of a discourse unit is presented below.

(135)

Constructing the Cf of multiple utterances:

- Put entities of Cf(U_i) in a list.
- Do until there is no U_{i-1}
 - Put entities of U_{i-1} at the end of the list.
 - Get the next U_{i-1} .
- Loop
 - Sort the list by recency.
 - Sort the list by frequency.
 - Update Cf of U_i 's mother node.
 - Determine Cb and transition state of the mother node.
 - If the mother node is also the right most node,
 - Construct the Cf for its mother node.

6.3 Results

The results in Table 10 below indicates that the extended centering algorithm can resolve more instances of zero pronouns than the existing centering algorithm. The results when applying the existing centering algorithm indicate 159 cases where zero pronouns could not be resolved. These are zeroes whose antecedent are not in the immediately preceding utterance. When the extended centering algorithm is applied, additional 42 instances of zeroes are resolved.

	No of Zero	Success	Fail	%Success	%Fail
Existing centering	1254	1095	159	87.32%	12.68%
Extended centering	1254	1137	117	90.67%	9.33%

Table 10: Results of existing centering and extended centering

However, since the extended centering algorithm only used the immediately preceding unit for resolving zeroes, it did not account for zero pronouns in a controlling pattern as discussed in section 3.4.3. Therefore, we modified the extended and the existing algorithms to search the next preceding unit, i.e. U_{i-2} , when the ‘correct’ referent was not in U_{i-1} . In order to make the two algorithms comparable, the existing centering algorithm was modified to account for frequency of occurrence. Entities in U_{i-2} that also occur in U_{i-1} are more salient than those that do not occur in U_{i-1} . Then we tested the two algorithms again. The results of the new test are as follow:

	No of Zero	Success	Fail	%Success	%Fail
Existing centering	1254	1194	60	95.22%	4.78%
Extended centering	1254	1216	38	96.97%	3.03%

Table 11: Results of existing and extended centerings with two-step lookback

From Table 11, the extended centering could resolve 1216 instances of zeroes, or 79 instances more than its first test, while the existing centering could resolve 99 more instances of zeroes. From these results, we can infer that 1095 instances of zeroes have antecedents in the immediately preceding utterance, 99 instances (1194 -1095) have antecedent in the next preceding utterance, and at least 22 instances (60 - 38) have antecedent in more than two utterances away.

However, the success of the extended centering algorithm is owned to the fact that it has a wider scope of reference resolution than the existing centering. To see

whether the extended centering algorithm is more efficient than the existing centering algorithm, we compared the success of both algorithms in terms of the number of successes at the first-try and the number of attempts before the centering algorithms could suggest the correct referent. Table 12 shows the number of first-try successes and attempts in the extended centering and the existing centering algorithms, (with two-step lookback).

	NoOfZero	First-try Success	Total Success	Attempts	Work-load
Existing centering	1254	819	1194	477	1.40
Extended centering	1254	829	1216	559	1.46

Table 12: First-try success and attempts in existing and extended centerings

The third column indicates the number of times when the first-suggested entities are the ‘correct’ referents. The number of attempts used by the centering algorithms before they could suggest the correct referent is shown in the fifth column. The work-load column indicates the average number of efforts for resolving a zero pronoun. If the work-load is two, it means that the algorithms have to try two times before they can resolve a zero pronoun. It is calculated by adding the number of attempts to that of the total success and dividing the sum by the number of total success.

From Table 12, the extended centering algorithm did not have greater success than the existing centering algorithm. In fact, both approaches had about the same first-try success. The extended model seemed to have more difficulty in resolving zero

pronouns because its work-load was greater than that of the existing model. However, if counting only zero pronouns which were resolved by both algorithms, the work-load of the extended model was about the same as the existing model, as shown in Table 13 below.

	NoOfZero	No of Zero resolved by both algorithms	Attempts	Work-load
Existing centering	1254	1194	477	1.40
Extended centering	1254	1194	465	1.39

Table 13: Attempts counted in existing and extended centerings

The extended centering did not perform better than the existing centering because most of antecedents of zero pronouns in our corpus were found in the immediately preceding utterance. There were few zero pronouns whose antecedents were many preceding utterances away. In this study, antecedents of 1,095 instances from 1,254 instances (87.32%) were found in the immediately preceding utterance; antecedents of 99 instances (7.89%) were found two preceding utterances away; and antecedents of 39 instances (3.11%) were found more than two preceding utterances away. Thus, in natural-occurring text, the existing centering algorithm can resolve most of the zero pronouns. But there are few cases where antecedents of zeroes are in a distant utterance and the existing centering algorithm alone cannot resolve these zeroes. These are cases where the hierarchical structure of discourse may be useful for the

resolution. In the next section, we will investigate whether hierarchical structure of discourse is an important factor for the resolution of these zero pronouns.

6.4 Discussions

Based on the results of zero pronoun resolution, we categorized zero pronouns into four groups as follows: Group I are zeroes that were resolved by both centering algorithms; Group II are zeroes that were resolved only by the existing centering algorithm; Group III are those that were resolved only by the extended centering algorithm; Group IV are those that could not be resolved by neither algorithm. Zero pronouns group I, which are the majority of cases (1194 instances), can be ignored here since they do not indicate differences between the two approaches. In addition, no zero in our study belongs to group II. Therefore, we can limit the discussion on zeroes that belong to groups III and IV.

6.4.1 Zero pronouns group III

Zero pronouns group III are those that were resolved by the extended centering but could not be resolved by the existing centering. There are 22 instances of zeroes on this type. We examined these zeroes to see whether hierarchical structure is helpful for zero pronoun resolution. We found that in most of these 22 cases, antecedents of zeroes

are not in the nearest unit, but in the nucleus part of the preceding unit.⁴⁴ Consider examples (136-139) below.

(136) Text from Health1

#23 *daɲnán thâa suàn tàaɲtàaɲ khǒoɲ hũu dâjráp kaan*
kràthópkràthuaan

therefore if part any of *ear*_{Z11} receive NOM *impact*_{Z17}
 ‘Therefore, if any part of ears receives an impact’

#24 *rũu [Z11] dâjráp chuáarôok*

or *[Z11]*_{Z12} receive germ
 ‘or (it) receives germ’

#25 *chêen [Z12] doon kràthêek rɛɛɲrɛɛɲ*

such-as *[Z12]*_{Z13} get strike strong
 ‘For example, (it) is struck’

#26 *[Z13] dâjjin siãaɲ daɲ mâak*

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

#27 *mii nám khâw hũu*

there-is water enter *ear*_{Z14}
 ‘Water enters ears’

#28 *[Z14] dâjráp chuáarôok*

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

#29 *phró [Z15] khé hũu duâaj khruâaɲmuu sòkkàpròk*

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

#30 *[Z16] khé hũu rɛɛɲ 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’

⁴⁴ Since we only asked our subjects to identify the hierarchical structure of the discourse, the judgement of what is the nucleus part is ours.

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.

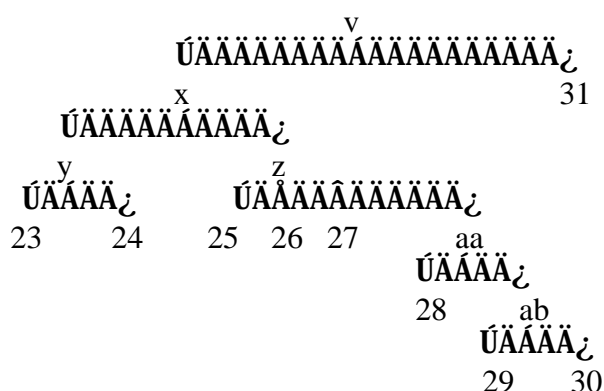


Figure 21: Hierarchical structure of example (136)

In examples (136), the antecedent of Z17 in (u31) is found in (u23), which is eight utterances earlier in linear view. But it is one unit back in structural view because (u23) is a part of the nucleus of unit (x), i.e. (y). The structure here can be viewed as a hierarchical structure of clauses in a sentence consisting of (u23-u31). Utterances (u25-u30) are an illustration part of the if-clause (u23-u24), while (u31) is the main clause of the sentence. Coreference of Z17 then can be viewed as an anaphora between a zero pronoun in the main clause (u31) and its antecedent in the nucleus of the subordinate part (y).

(137) Text from Health1

#65 rôok rŭu ʔantàraaj khǒŋ hŭu nôokcàak cà maa càak ʔantàraaj
thîi hŭu dâjráp doojtrŋ càak hŭu chánnôok
disease or danger of ear₇₄₆ beside will come from danger COMP
ear₇₄₅ receive directly from ear outer

‘Ear disease or damage, beside getting from the damage that ear receives from the outer ear’

#66 chên bajhũu lé kêwhũu dâjráp kaan kràthópkràthuaan

for-example auricle and eardrum receive impact

‘For example, auricle and eardrum receive an impact’

#67 chuáarôok khâw [Z45] phró kaan khé hũu duâaj májkhéhũu thii
mâj sàʔàat léew

germ enter [Z45] because-of NOM pick ear with an-ear-stick
COMP not clean ASP

'Germ enters (ear) because of picking ear with a stick that is not clean'

#68 [Z46] jaŋ mii sǎahèet càak kaan thîi hũu mii thôo tittòo kàp
suàan ʔùunʔùun ʔiik

[Z46] CONJ there-is cause from NOM COMP ear have canal
connect with part other too

‘(The damage) is caused from the fact that ear has canal connecting to other organ’

... Ear damage is not only a result of damages received from the outer ear, such as an impact on ear blade and eardrum, or transmission of germs from dirty ear-sticks. It may also be a result of damage in other organs that are connected to the ear. ...

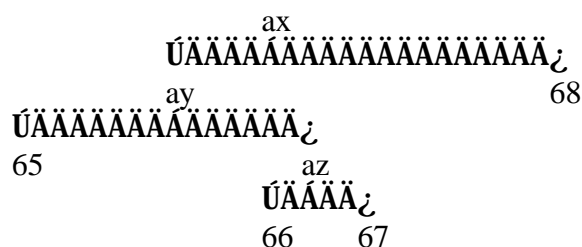


Figure 22: Hierarchical structure of example (137)

In examples (137), the antecedent of Z46 in (u68) is found in (u65), which is three earlier before in linear view. But it is one unit back in structural view because (u65) is the nucleus part of the unit (ay). The hierarchical structure here can be viewed as a complex sentence, in which (u68) is the main clause, (ay) is the subordinate part.

In unit (ay), (u66-u67) are the illustration part of the nucleus (u65). Thus, coreference of Z46 is an anaphora within a sentence, between a zero pronoun in the main clause (u68) and its antecedent in the subordinate clause (u65).

(138) Text from Health1

- #80 cà kèet kaanʔàksèep naj chōŋ hũu chánklaan
 will occur infection_{Z52} in cavity ear_{Z51} middle
 ‘There will be an infection in the cavity of the middle ear’
- #81 [Z51] cà dâjjin siăan nōoj lon
 [Z51] will hear sound decrease ASP
 ‘(Ear) will receive less sound’
- #82 rũu kèet kêewhũu thálú
 or occur eardrum torn
 ‘Or, eardrum is torn’
- #83 mii námnoŋ lăj
 there-is lymph out
 ‘Lymph comes out’
- #84 hàak [Z52] pēn mâak
 if [Z52] be much
 ‘If (the infection) is severe’

... (If the middle ear gets any disease from the outer ear,) the cavity in it will also be infected. As a result, we might lose our hearing, the eardrum might be torn, or there might be lymph coming out of ears. If the infection is severe, ...

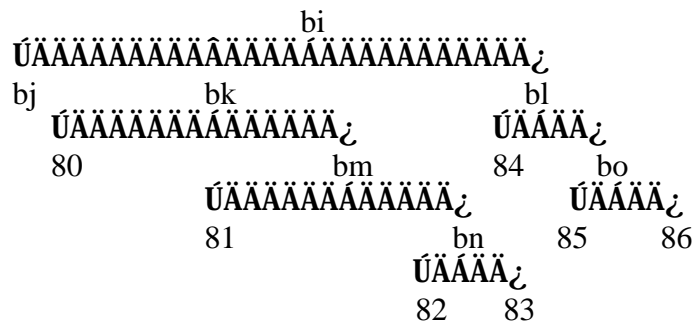


Figure 23: Hierarchical structure of example (138)

In examples (138), the antecedent of Z52 in (u84) is found four utterances earlier in linear view (u80). But it is only one unit back in structural view because (u80) is the nucleus part of the unit (bk). In this example, (bk) can be viewed as a sentence consisting of (u80-u83) and (bl) as the next sentence consisting of (u84-u86). Coreference of Z52 then can be viewed as an anaphora between two contiguous sentences.

(139) Text from Editor2.txt

#5 kaanluâaktân thuâapaj naj kamphuuchaa khrán ní càt khûn maa
dooj khwaamphájaajaam khǝŋ ʔoŋkaansàhàpràchaachâat
election general in Cambodia time this arrange ASP ASP by effort
of UN_{Z1,Z4}

‘The election in Cambodia this time is arranged by the UN’

#6 thîi [Z1] dâj thamhâj klùm khâměensiifàaj loŋnaam naj
khǝŋtòkloŋ sântìphâap thîi krunpaariit pràthêetfàràŋsèet
muâa pii phǝŋsǝ 2534

COMP [Z1] ASP make group four-party_{Z2} sign in agreement peace
at Paris France in year 1991

‘who made the four parties sign the peace agreement in Paris in 1991’

#7 léew [Z2] damnəenkaan taam nɛwthaŋ thîi kamnòt wáj
then [Z2] practice along guideline COMP outline ASP
‘and (the four party) to follow the guideline that is outlined’

#8 phuâa hâj kèet sântìphâap jàaŋ thǎawəen
for give occur peace Adv-Mrk permanent
‘so that peace could occur permanently’

#9 dooj [Z3] cháj kràbuaankaan thaŋ kaanmuaaŋ naj
rábǝppràchaathíppàtaj khuu kaanluâaktân thuâapaj
such-that [Z3] use process of political in democracy be election
general
‘by the use of political process of democracy that is general election.’

#10 [Z4] mǎaj hâj khâměen tɛɛlá klùm sǝŋ tuaathɛen
loŋsàmàkrárápluâaktân

[Z4] want give Khamer each group send representative apply-for-election

‘(The UN) wants each group of Khmer to send representatives in the election’

#11 1έεw [Z5] hâj pràchaachon chaawkhàměen pen phûuluâak

then [Z5] give people Khamer be chooser

‘Then, let Khamer people be the chooser’

The election in Cambodia was arranged by the UN, who made the four parties sign the peace agreement in Paris in 1991. The four parties had to follow the guideline supported by the UN. The election will be hold so that Khamer people will be the one who choose their government.'

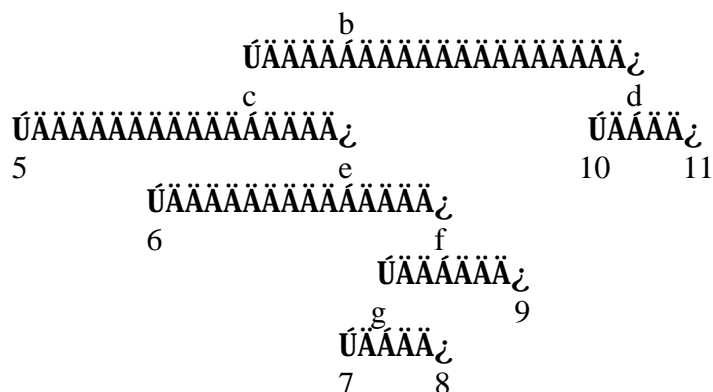


Figure 24: Hierarchical structure of example (139)

In example (139), the antecedent of Z4 is in (u5), which is six utterances earlier in linear view. But it is one unit back in structural view because (u5) is the nucleus of the unit (c). In this example, (c) can be viewed as a complex sentence consisting of (u5-u9). The next sentence is the unit (d) consisting of (u10) and (u11). Thus, coreference of Z4 is an anaphora between a zero pronoun in (u5) and its antecedent in the main clause of the preceding sentence, i.e. (u5).

As seen from examples above, coreferences of these zero pronouns could be described on the basis of hierarchical structure of clauses at the sentence level. This is not surprising. If we consider the nature of the backward-looking center, which is assumed to be the most focused entity or the current attention of the utterance and functions as a cohesive link between the current utterance and the previous one, it is not surprising to find most antecedents of zero pronouns in the immediately preceding utterance. Rather, it should be questioned why few zero pronouns do not have their antecedents in the immediately preceding utterance. How can their referents be resolved? The answer may be inferred from these examples. Though antecedents of these zero pronouns are not in the immediately preceding utterance, they are in the same sentence or in the preceding sentence. The hierarchical structure of clauses in the sentence will help locate the antecedents. Therefore, it will not be too difficult for hearers to infer the correct referents for these zero pronouns. However, the number of examples found in this study is too small to make a strong conclusion.

6.4.2 Zero pronouns group IV

There are 38 instances of zeroes that could not be resolved by neither the existing nor the extended centering algorithms. Of these 38 instances, we found that 15 instances (5 zeroes) failed because they are in the first utterance of the paragraph and their antecedents are in the previous paragraph. Since we did not ask our subjects to

analyze the structure of discourse beyond paragraph boundaries, the previous paragraph was not included in the scope of reference resolution. As for the rest (23 instances), 17 instances failed because the antecedent of zeroes are not in the scope of reference resolution while 6 instances fail because all antecedents of the zeroes are not in the same discourse unit. The centering algorithms in this study are not modeled to search for antecedents of zeroes in separate units. Like Brennan et al.'s algorithm (1987), the model here assumes that all antecedents are found in the same utterance or discourse unit.

As for the 17 instances, 5 instances (5 zeroes) failed in one test but were resolved in the other two tests. And 12 instances (4 zeroes) failed in all three tests because neither structure analyzed provides enough scope of reference resolution. What is interesting here is those four zeroes which failed to be resolved in all three tests. Below are examples in which the antecedents of the four zeroes are not covered by the scope of reference resolution.

(140) Text from Comp3

#49 thaənʔòok ʔan nùŋ khuu kaan thii bàrìsàt lé nuàajŋaan naj
təənpràthêet chēen sàhàráʔàmeeríkkaa ʔəŋkrìt lé jīipùn dāj
mii kaan wícaj lé khítkhón rábòpʔintəəfèet trəŋklaən thii
sǎamāat ráprúu lé khāwcaj phaasǎamánút rǔu
phaasǎathammáchāat

solution CL one be NOM COMP company and organization in
foreign-country such-as USA. England and Japan ASP have NOM
research and invent interface-system central₂₂₁ COMP can recognize
and understand human-language or natural-language

‘One solution is to have companies or organization in other countries such as the US, England, or Japan, to invent a central interface system that can recognize human language or natural language’

#50 thîi phûucháŋnaaŋ cà sàŋnaaŋ [CZ19] dâj càak khaamsàŋ naj rûup
khõŋ ʔèeksăaŋ rûu siăaŋphûut

COMP user will command [CZ19] ASP from instruction in form of text or voice

‘that users can use (it) by instruction in the forms of text or voice’

#51 [Z20] cà phàan kaan plεε lé tiikhwaam

[Z20] will pass NOM translate and interpret

‘(Text or voice command) will be translated and interpreted’

#52 léew [Z21] sâaŋ pən khamsàŋ màj taam kaan pràmuaanphôn
phaasǎathammáchâat (NLP)

then [Z21] build be command new follow NOM process natural-language (NLP)

‘then (the interface system) will build new commands by using natural language processing’

#53 [Z22] pa.j sàŋ chá.jŋaan sópwee lé thǎankhôm̐muun ʔiik thōot nǎn

[Z22] go command use software and database again time one

‘(The new commands) then control software and database system’

One solution is to let companies or organizations in the US, England, or Japan create an interface system that can take natural language commands so that users can use text or voice when using a computer. Text and voice command will then be interpreted by an NLP system into commands of other software or databases.

(a)

af

ÚÄÄÄÄÄÄÄÄÄÄÄÄÄÄÄÄÄ¿

49 50

ag

ÚÄÄÄÄÄÄÄÄ¿

51 52 53

(b)

aj

ÚÄÄÄÄÄÄÄÄÄÄÄ¿

49

ak

ÚÄÄÄÄÄÄÄ¿

50

al

ÚÄÄÄÄÄÄÄÄ¿

51 52 53

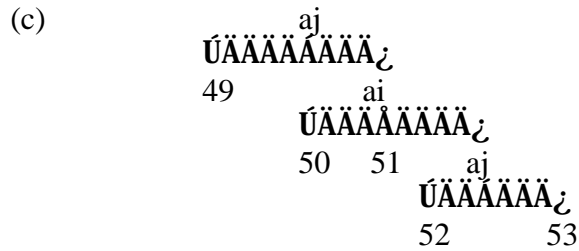


Figure 25: Hierarchical structures of example (140)

In example (140), the antecedent of Z21 in (u52) is in (u49). All subjects analyzed the structure of this example differently, as shown in Figure 25. Although, the extended centering algorithm were not able to resolve this zero, utterances (u49)-(u53) can be viewed as a complex sentence, in which (u49) is the main clause and Z21 is in the subordinate clause.

(141) Text from Article1

#84 tɛɛ [Z41] kô khonj tɔɔj ráwəŋtuaa mâak khûn

but [Z41]_{Z43} then may must care much more

‘But (Mr.Vachara) have to be more careful’

#85 phró lăŋcàaknánmaa mii kaan lâattràween

as after-that there-be NOM observe

‘as there have been illegal people being around my house after that’

#86 khláaj [Z42] cà bòok wâa

like [Z42] will tell that

‘It looks like (they) want to tell that’

#87 tɔɔpaj [Z43] ʔàatcà mâj cəə máj

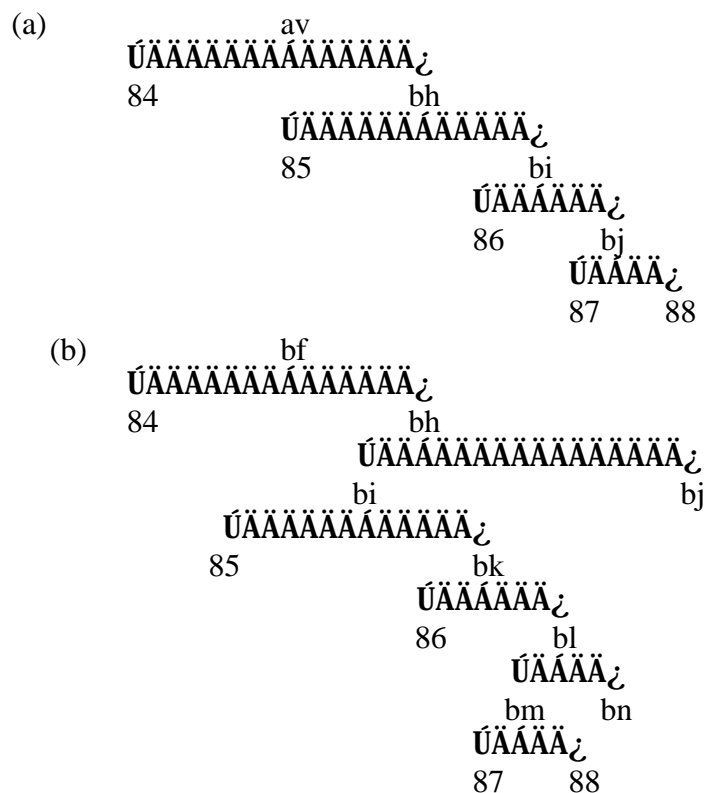
next [Z43]_{Z44} may not meet wood-stick

‘next time, (Mr.Vachara) would not be injured by a wood-stick’

#88 [Z44] cəə puun ləəj ʔàrajjàəŋníi

[Z44] meet gun ASP like-that

‘(Mr.Vachara) would be injured by a gun, something like that’



In example (141), the antecedent of Z43 in (u87) is in (u84). The structures analyzed for this example are shown in Figure 26. Two subjects came up with the same structure as (a) but another subject analyzed the structure as (b). Utterances (u84-u88) can be viewed as a complex sentence, in which the antecedent of Z43 is in the main clause (u84).

(142) Text from Comp4

- #86 baan̄khrán̄ muâa prookreem câwbân tham̄jaan paj
 sometimes when program host_{Z43} work ASP
 ‘Sometimes, when the host program works,’
- #87 [Z43] kô cà paj krâtûn hâj kham̄sân̄ wajrât nîi tham̄jaan taam
 paj duâaj
[Z43]_{Z47} then will go activate let instruction virus this work follow ASP
 too
 ‘(it) will activate the virus program’
- #88 phôn̄ kô khuu man cà rîip kóoppîi tuaaʔeen̄ paj wáj jaŋ
 prookreem tâan̄tân̄
 result then be it_{Z44} will hurry copy itself go keep at program any
 ‘The result is that (the virus program) will copy itself to any program’
- #89 sôn̄ [Z44] ʔaat tham̄hâj kèet khwaamsiãahãaj dâj
 CONJ [Z44]_{Z45} maybe cause occur damage ASP
 ‘It could cause a damage’
- #90 chên̄ [Z45] paj lóp fém̄khôomuun
 for example [Z45]_{Z46} go delete file
 ‘For example, (it) may delete some files’
- #91 rûu [Z46] tham̄hâj [Z47] pluaan̄ weelaa tham̄jaan mâak paj
 doojchâjhèet
 or [Z46] cause [Z47] waste time work much ASP unnecessary
 ‘or (it) could cause (the host program) to waste more time unnecessary.’

When the host program works, those virus instructions will be active. The virus program, then, will copy itself to other programs, destroy data files, or slow down the system.

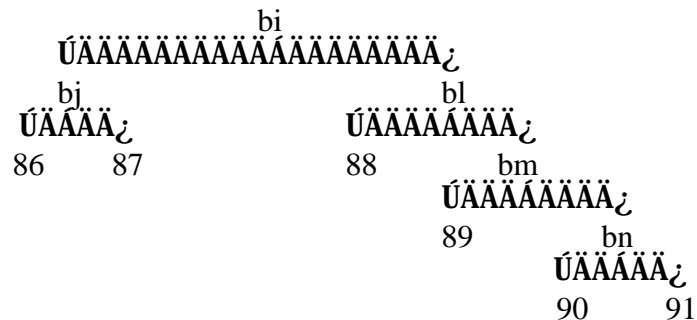


Figure 27: Hierarchical structure of example (142)

In example (142), Z46 and Z47 could not be resolved because the referent of Z47 in (u91) is referred to in (u86-u87) which is four or five utterances earlier. Three subjects provided the same structure for this example, as shown in Figure 27. In this example, (u86-u87) can be viewed as a sentence, and (u88-u90) as another sentence. Thus, antecedent of Z47 is found in the preceding sentence.

Although these zeroes were not resolved by neither algorithms, coreferences of these zeroes can be viewed as anaphora within a sentence or anaphora between two contiguous sentences. Therefore, it would not too difficult for readers to resolve these zero pronouns.

6.5 Further research

In this section, first, we discuss issues in Centering Theory that are areas of active research. Next, we discuss how Centering Theory might be extended in other directions. In section 6.5.2, we argue against Strube and Hahn (1996) who propose a model of ‘functional centering’. In section 6.5.3, we propose that the extended centering model should take into account the differences between nucleus and satellite units. We think that the model would be able to resolve zero pronouns more efficiently if the distinction between nucleus and satellite units is taken into account. In section 6.5.4, we examine whether positions of utterances in the hierarchical structure are useful for zero pronoun resolution.

6.5.1 Open issues on Centering Theory

After Centering Theory was formulated in 1983, it has been applied to other languages, such as Japanese (Walker, Iida, and Cote 1990, 1994, Kameyama 1985, 1986, Iida 1998), Italian (Eugenio 1990, 1996, 1998), German (Strube and Hahn 1996), and Turkish (Turan 1998). However, many issues are still areas of active research. For example, beside grammatical function, what factors are relevant for ranking forward-looking centers? Are those factors language specific? Is the backward-looking center an obligatory part of every utterance? What should the centering process be at the discourse segment boundary? Is intersentential centering different from intrasentential centering? These issues are the current research in Centering Theory (Walker, Joshi, and Prince 1998). Our study is just another direction of extending Centering Theory to be more applicable to naturally-occurring texts. We propose to extend the theory by incorporating hierarchical structure of discourse. But the model here is constructed on the basis of some assumptions which are still an open issue. First, in this study we assume that grammatical functions are relevant for ranking forward-looking centers and the preference order is the same as that of Japanese. In addition, we use frequency and recency factors for ranking forward-looking centers in a unit larger than a single utterance. If different factors were used, the results from the extended centering would be different. Second, we assume the backward-looking center to be an obligatory part

of every unit. When it cannot be determined from the preceding unit, it is set to be the same as the preferred center of the current unit. What we do is similar to what Brennan et al. (1987) did, while Walker et al. (1996) and Kameyama (1993) preferred the other way. Walker et al. (1996) allowed the Cb to be null if it cannot be determined from the preceding utterance. Kameyama (1993) set an additional condition that the Cb must be realized as a pronoun. These differences could affect the results from centering algorithms.⁴⁵ In addition, when we applied hierarchical structure of discourse to the centering model, we assumed that centering process is homogenous to all types of sentences. But Kameyama (1998) used different methods of updating centering when processing different types of complex sentences. Assumptions underlying in this study will certainly affect the process of the extended model. Therefore, it is not surprising if the results from the extended model is not as good as expected.

However, the central question of this study is to investigate whether discourse structure is relevant for zero pronoun resolution. From the analysis of our corpus, most of zero pronouns do not require discourse structure for the resolution. Only a few may need this kind of knowledge, but it seems to be the hierarchical structure of clauses at the sentence level rather than at the discourse level. Then, should centering algorithms be applied between sentences rather than between utterances? We think that centering

⁴⁵ This is because the resolution of pronouns is based on transition states, which is determined from the Cb and the Cp of utterance pairs.

algorithms should apply between utterances rather than between sentences. As discussed in Kameyama (1998), sentence-based centering requires more computational process than clause-based centering. It is preferable to break a complex sentence into a structure of utterances and apply centering between utterances.

6.5.2 Functional centering

Strube and Hahn (1996) introduced a revision of the Centering Theory: a functional centering. They claimed that the new model could account for a free word order language like German better than the original model. The main differences between the functional centering model and the original centering model are the criteria for ranking entities in Cf. While most of the research on Centering Theory use grammatical roles to rank entities in Cf, functional centering uses ‘functional information structure’ in terms of ‘context-boundedness’ to rank entities in Cf. An entity is bounded if it is also in Cf(U_{i-1}). A bounded entity is ranked before an unbounded entity. Strube and Hahn (1996) used the cost of transition states to show the preference of the functional model to the original model. They determined transition cost in relation to a pair of transition states, as shown in Table 14 (Strube and Hahn 1996:7). A transition pair is considered cheap if Cb(U_i) is the same as Cp(U_{i-1}). For example, a sequence continuation-continuation is a cheap transition while a sequence continuation-shifting is an expensive one. Strube and Hahn claimed that the functional

model yield more continuation states at a lower transition cost. Thus, it is better than other models in terms of inference load. (The lower the transition cost, the less of an inference load posted on listeners.) They also claimed that the functional constraint may possibly be universal since it can be verified on all examples used in the centering literature.

	continue	retain	smooth-shift	rough-shift
-	cheap	expensive	-	-
continue	cheap	cheap	expensive	expensive
retain	expensive	expensive	cheap	expensive
smooth-shift	cheap	expensive	expensive	expensive
rough-shift	expensive	expensive	cheap	expensive

Table 14: Costs of transition pairs

However, we did not find context-boundedness to be useful in zero pronoun resolution in Thai. The results when using boundedness is worse in both the existing and the extended centering algorithms. We implemented the functional model by adding a routine to reorder entities in a Cf with respect to the context-boundedness. An entity that is bounded to the immediately preceding unit is set to be more salient than an unbounded entity. The results of the centering algorithms (with two-step look back) when using context-boundedness feature is presented in Table 15 below.

		NoOfZero	First-try Success	Total Success	Attempt	Work-load
existing centering	non- functional	1254	819	1194	477	1.40
	functional	1254	813	1194	489	1.41
extended centering	non- functional	1254	829	1216	559	1.46
	functional	1254	820	1216	556	1.46

Table 15: Results of non-functional and functional centerings

When context-boundedness is used, the first-try success in the existing and the extended centering algorithms dropped from 819 and 829 to 813 and 820 respectively. The attempt and work-load of the functional model were also not better than those of non-functional model. This is because, in many cases where a grammatical role was overridden by a context-boundedness feature, zero pronouns were more difficult to be resolved, as seen from examples (143-144) below:

(143) Text from Comp1

#55 naj pàtcùban níi phaasăa thîi cháj khiăan prookreem

rábòpphûuchiâawchaan khuu phaasăa LISP

in present this language COMP use write program expert-system be language LISP

‘Now, the language that is used for developing an expert system is LISP’

#56 sũ [Z24] pən phaasăa thîi [Z25] khâwcaj jâak nõj

CONJ [Z24] be language COMP [Z25] understand difficult a-bit

‘(It) is a language that is quite difficult to learn’

#57 chòokdii thîi dâj mii phûu plɛɛ phaasăa daŋklàaw

lucky COMP ASP there-be man_{Z26} translate language this

‘Luckily, someone has translated this language’

#58 lé [Z26] camlɔɔŋ rábòpphûuchiâawchaan loŋ bon

khruâaŋmajkhrookhoomphiwtêe

and [Z26] simulate expert-system down on micro-computer

‘and (that person) has imported an expert system into a micro-computer’

The language used for developing an expert system now is LISP. It is a difficult language. Fortunately, the LISP language has been transported to a micro-computer so that an expert system can be developed on a micro-computer.

In example (143), if context-boundedness is used, the centering algorithms will predict ‘LISP’ as the first preferred referent for Z26 in (u58) instead of ‘Person1’, which is the ‘correct’ referent. Since ‘LISP’ is already referred to in (u56), ‘LISP’ will be more salient than ‘Person1’ in (u57). The order of entities in Cf(u57) will be changed from (Person1, LISP) to (LISP, Person1). But if context-boundedness is not used, the centering algorithms will predict correctly that ‘Person1’ is the preferred referent of Z26.

(144) Text from News1

- #8 kooránii thîi phráthamkoosăacaan rǎu
 thāanphúttháthâatphíkkihù hēēŋ wátsuăanmôokphálaaraam
 ʔampheəchajjaa caŋwàtsùrâatthaanii ʔaaphâat nàk kòon cà thǎŋ
 wan khláaj wankèet troŋ kàp wanthîi 27 phríksàphaakhom níi
 a-case COMP Pra-Thamkosajarn or Bhutathat-Bhikhu_{Z6} (‘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 roonphájaabaansùrâatthaanii nímon [Z6] paj
 ráksăatuaa jùu thîi roonphájaabaan
 and doctors_{Z7} (‘Doctor1’) Suratthanee-hospital invite [Z6]_{Z8} go treat
 stay at hospital
 ‘and doctors in Suratthanee hospital has taken (him) for a treatment in the
 hospital’
- #10 sǎŋ lǎŋcàak thîi [Z7] truàatʔaakaan [Z8] léew
 CONJ after COMP [Z7]_{Z9} examine [Z8]_{Z10} ASP

‘After (the doctors) have examined (him)’

#11 [Z9] kô dâj hâj [Z10] nɔɔnphák ráksăatuaa jùu naj hɔɔŋ
 ʔajsiijuu tâŋtɛɛ cháwtrùu 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.

In example (144), when context-boundedness is used, ‘Bhikhu’ in (u9) will be more salient than ‘Doctor1’ because it is also in Cf(u8). The order of entities in Cf(u9) will be changed from (Bhikhu, Doctor1, Hospital1) to (Doctor1, Bhikhu, Hospital1). Then, the continuation state from (u9) to (u10) will be preferred, and ‘Doctor1’ and ‘Bhikhu’ will be the first-suggested referents for Z7 and Z8 respectively. But if context-boundedness is not used, Cf(u9) will be (Bhikhu, Doctor1, Hospital1), and the centering algorithms will correctly predict ‘Bhikhu’ and ‘Doctor1’ as the referents for Z7 and Z8 respectively.

Therefore, in Thai, grammatical functions are an important factor in ordering entities in the Cf. Context-boundedness cannot replace the grammatical functions. Therefore, it is unlikely for context-boundedness to be a universal criteria for ranking entities in a Cf as hypothesized by Strube and Hahn (1996).

As for the lower of transition cost and the higher number of continuation states, which is the reason Strube and Hahn used to argue in favor of the functional centering model, we calculated transition costs as defined in Table 14. When a context-boundedness was implemented, it gave us additional continuation states and lower transition costs as Strube and Hahn contended. Table 16 shows the number of continuation states and the number of cheap and expensive transitions in our comparison.

		No of Continuation	Cheap	Expensive
existing centering	non-functional	2101	1803	2054
	functional	2389	1886	1971
extended centering	non-functional	3918	3054	3172
	functional	4121	3075	3151

Table 16: Number of continuation and transition cost

It is true that the functional model gave us more continuation states and lower transition costs. However, we think that the results are directly related to the use of context-boundedness. Since bounded entities are likely to be on the first rank in the Cf, it is not surprising that continuation states would occur more often in the functional model. When transition states are likely to be a sequence of continuation, transition cost in the functional model will be lower than transition cost in a non-functional model. Therefore, the results of the transition cost should not be used to determine which model is better. We think that it is more important to look at the success in

resolution rather than at the number of continuation states and transition costs. Transition states should be used in determining the degree of coherence of a discourse, rather than in measuring the advantage of an approach. In example (145) below, the discourse would have more continuation states and lower transition costs when context-boundedness were used, but that does not mean that the discourse is easier to understand when context-boundedness were used. The degree of inference load is still the same no matter what approach is used.

(145)

- a. John went to his favorite music store to buy a piano.
- b. It was a store John had frequented for many years.
- c. He was excited that he could finally buy a piano.
- d. It was closing just as John arrived. (Grosz et al. 1995:206)

In sum, the degree of inference load should not be determined from the number of continuation states or transition costs. The higher number of continuation states is a result of the process we have defined. Therefore, while the context-boundedness is useful for anaphora resolution in a free word order language like German, it cannot replace the role of grammatical function in a configurational language like Thai.

6.5.3 Nucleus and satellite units

The extended centering algorithm might perform better if the distinction between nucleus and satellite units is recognized by the algorithm. In this study, we did not distinguish between nucleus and satellite units as used in rhetorical structure theory

(Mann and Thompson 1987) because we followed the conclusion of Fox's analysis (1987:95) that a pronoun can be used when its referent is already referred to in an 'active' or 'controlling' unit (see section 3.4.2). Since an active unit can be either a nucleus or a satellite unit whose structure partner is being processed, we were able to simplify the analysis, and not ask the subjects to distinguish between nucleus and satellite units during the discourse structure analysis. However, we now think that the distinction between nucleus and satellite might be useful for setting saliency of discourse entities. It may be possible that the lower number of first-try success in the extended centering algorithm is a result of not using nucleus-satellite information. For instance, we found that in examples (136-138) repeated below as (146-148), when discourse entities in the nucleus unit were set to be more salient than entities in satellite units, the extended centering algorithm was able to resolve the zero pronouns easily. Instead of using 5, 4, and 4 attempts for resolving zero pronouns Z17, Z46, and Z52 in examples (146-148) respectively, the extended centering algorithm would be able to resolve these zero pronouns at the first-try or with one attempt.

(146) Text from Health1

#23 daɲnán thâa suàan tàaɲtàaɲ khǒɔɲ hũu dâjráp kaan
kràthópkràthuaan

therefore if part any of ear_{Z11} receive NOM impact_{Z17}

'Therefore, if any part of ears receives an impact'

#24 rǎu [Z11] dâjráp chuáarôok

or [Z11]_{Z12} receive germ

'or (it) receives germ'

- #25 chêen [Z12] doon kràthêek rɛɛŋrɛɛŋ
such-as [Z12]_{Z13} 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_{Z14}
'Water enters ears'
- #28 [Z14] dâjráp chuáarôok
[Z14] receive germ
'(Ears) receive germ'
- #29 phró [Z15] khé hũu duâaj khruâaŋmuu sòkkàpròk
because [Z15] pick ear with instrument dirty
'because (we) pick ears with dirty instrument'
- #30 [Z16] khé hũu rɛɛŋ kœn paj
[Z16] pick ear strong over ASP
'(we) pick ears too hard'
- #31 [Z17] ʔàat thamhâj kèet 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.

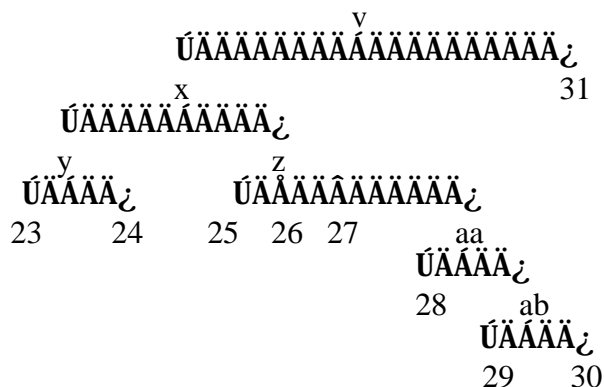


Figure 28: Hierarchical structure of example (146)

In example (146), the antecedent of Z17 is found in (u23) which is a part of the nucleus unit (y). If entities in the nucleus unit were set to be more salient than other entities, the entity ‘Impact1’ would be ranked the second instead of the sixth entity on Cf(x). Thus, the extended centering algorithm would be able to resolve Z17 with one attempt instead of five attempts.

(147) Text from Health1

#65 rôok rǔu ʔantàraaj khǒŋ hǔu nǒkcaak cà maa càak ʔantàraaj
thīi hǔu dâjráp doojtroŋ càak hǔu chánnǒk

disease or danger of ear_{Z46} beside will come from danger COMP
ear_{Z45} receive directly from ear outer

‘Ear disease or damage, beside getting from the damage that ear receives from the outer ear’

#66 chēen bajhǔu lé kēwhǔu dâjráp kaan kràthópràthuaan
for-example auricle and eardrum receive impact

‘For example, auricle and eardrum receive an impact’

#67 chuáarǒok khâw [Z45] phró kaan khé hǔu duâaj májkhéhǔu thīi
mâj sàʔaat léew

germ enter [Z45] because-of NOM pick ear with an-ear-stick
COMP not clean ASP

‘Germ enters (ear) because of picking ear with a stick that is not clean’

#68 [Z46] jaŋ mii sǎahèet càak kaan thīi hǔu mii thǒo tittòo kàp
suàan ʔùunʔùun ʔìik

[Z46] CONJ there-is cause from NOM COMP ear have canal
connect with part other too

‘(The damage) is caused from the fact that ear has canal connecting to other organ’

... Ear damage is not only a result of damages received from the outer ear, such as an impact on ear blade and eardrum, or transmission of germs from dirty ear-sticks. It may also be a result of damage in other organs that are connected to the ear. ...

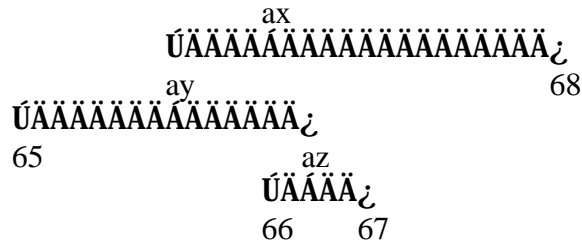


Figure 29: Hierarchical structure of example (147)

In example (147), (u65) is the nucleus part of the active unit (ay). If entities in (u65) is set to be more salient than entities in the satellite unit (az), ‘Ear-disease’ would be the most salient entity in Cf(ay). Then, the extended centering algorithm should successfully resolve Z46 at the first-try.

(148) Text from Health1

- #80 cà kèet kaanʔàksèep naj chôn hũu chánklaan
 will occur infection_{Z52} in cavity ear_{Z51} middle
 ‘There will be an infection in the cavity of the middle ear’
- #81 [Z51] cà dâjjin siăan nój lon
 [Z51] will hear sound decrease ASP
 ‘(Ear) will receive less sound’
- #82 rũu kèet kêwhũu thálú
 or occur eardrum torn
 ‘Or, eardrum is torn’
- #83 mi i námnoŋ lăj
 there-is lymph out
 ‘Lymph comes out’
- #84 hàak [Z52] pən mâak
 if [Z52] be much
 ‘If (the infection) is severe’

... (If the middle ear gets any disease from the outer ear,) the cavity in it will also be infected. As a result, we might lose our hearing, the eardrum might be torn, or there might be lymph coming out of ears. If the infection is severe, ...

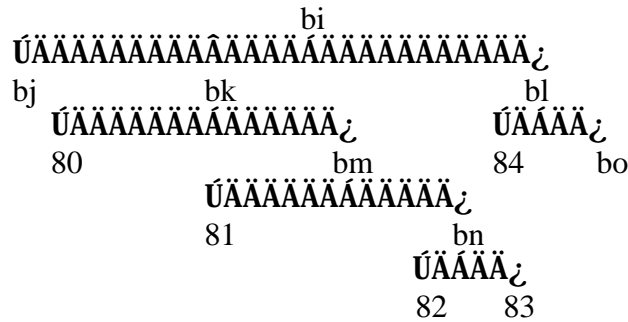


Figure 30: Hierarchical structure of example (148)

In example (148), the antecedent of Z52 is found in (u80), which is the nucleus part of the active unit (uk). If entities in the nucleus part is set to be more salient, the referent of Z52, i.e. ‘Inflection’, would be the highest-ranked entity and the extended centering algorithm would be able to resolve Z52 at the first-try.

Therefore, further testing should be taken to see whether the extended centering algorithm could perform better if the distinction between nucleus and satellite units are included.

6.5.4 Positions on hierarchical structure

In addition to the distinction between nucleus and satellite units, we question whether the position of utterances in the hierarchical structure of discourse is relevant for the resolution. In section 5.2.2, we hypothesized that entities in an utterance at a higher level of structure is more salient than entities in an utterance at a lower level because we think that they should be easier to access than others. However, the hypothesis is not supported by our corpus. The probability of finding antecedents at the

lowest level of structure is higher than the probability of finding antecedents at the highest level. But we think that it may result from the fact that there are more utterances at the lowest level than those at the highest level. Therefore, we did not use hierarchical level for ranking discourse entities in the Cf. In this section, we examine this issue further. We modify the extended centering algorithm to account for the position of utterances in the hierarchical structure. Two versions of the extended centering were tested. One preferred entities in utterances at the higher level. The other preferred entities in utterances at the lower level. To see the effect of this factor alone, frequency and recency effects were excluded from the ranking of Cf in this test. The results in Table 17 indicate that the model that prefers entities in utterances at the higher level can resolve zero pronouns more efficiently than the other model.

	NoOfZero	First-try Success	Total Success	Attempt	Work-load
prefer higher level	1254	824	1216	565	1.46
prefer lower level	1254	809	1216	598	1.49

Table 17: Result of extended centering with hierarchy factor

Therefore, if position of utterances in the hierarchical structure were to be used for ranking the Cf, entities at the higher level should be more salient than entities at the lower level. But whether it is a significant factor for ranking the Cf is still subject to further research. In this study, when we add this factor for ranking the Cf, the results are the same as the one using only frequency and recency

In addition to the level of hierarchical structure, it might be possible to look at position of utterances in other aspects. For example, Webber (1997) argues that only the right frontier of the structure is accessible for the resolution of demonstrative pronouns that is used as a discourse deixis. Polanyi (1988) also argues that only the rightmost node of the hierarchical structure is accessible for pronominalization in English. If this is also true for Thai, the extended centering can consider only entities that are in the utterance at the right frontier of the structure. This is certainly another area of further research.

6.6 Conclusion

In this study, we use discourse structure in our simulation, assuming that it is available. But in an actual NLP system, recognizing a discourse structure is a problem of its own. It is still an active area in natural language understanding research. Some researchers find clue phrases to be an important device to signal the beginning or ending of discourse segments (Grosz and Sidner 1986, Cohen 1987, Allen 1995). Others use coherent relation to determine the hierarchical structure of clauses in a discourse (Hobbs 1985, Polanyi 1988). It is obvious that further research on discourse structure is needed. In fact, it might be possible that the recognition of discourse structure is benefit from anaphora resolution. Whether the process of anaphora resolution comes after the process of discourse structure determination, or vice versa, is

an open question. In this study, we have to assume that the discourse structure is given, so that we can study the process of zero pronoun resolution. Our study focuses on investigating the contribution of discourse structure to zero pronoun resolution in Thai, and on extending the centering to work with the hierarchical structure of discourse. Although we did not find the hierarchical structure of clauses at the discourse level to be relevant for zero pronoun resolution in Thai, we found a few examples, in which hierarchical structure at the sentence level seems to be relevant. These examples suggest that the resolution could be done easier if the hierarchical structure of clauses and the distinction between nucleus and satellite parts are recognized. However, the number of examples found in this study are too small to confirm the conclusion. Further research should be pursued on a larger corpus to see whether the hierarchy structure of discourse is really relevant for the resolution.