Zero Pronoun Resolution in Thai: A Centering Approach

Wirote Aroonmanakun

Abstract

Since pronouns can be dropped in Thai, a natural language processing system for Thai must be able to resolve referents of the missing pronouns. One of several approaches that have been used for reference resolution is Centering Theory. Centering Theory is a focusing process in which salience of discourse entities is being kept track of. Referents of pronouns or zero pronouns are usually entities that are in focus. However, centering model can resolve only pronouns or zero pronouns whose antecedents are in the immediately preceding utterance. In this study, we indicate that antecedents of Thai zero pronouns are not always in the immediately preceding utterance. Discourse structure is hypothesized to be relevant for resolving zero pronouns, and centering model is extended to work with the hierarchical structure of discourse. To investigate whether hierarchical structure of discourse is relevant for zero pronoun resolution in Thai, the extended centering and the existing centering algorithms were tested on the same corpus. The results indicate that the extended model did not perform better than the existing model because most of antecedents are in the immediately preceding utterance. A few are in a distant utterance. Coreferences of these zeroes could be explained in terms of hierarchical structure of clauses, which seems to operate at the sentence level rather than at the discourse level. However, the number of examples found in this study are too small to make a strong conclusion. Further research should be pursued on a larger corpus to see whether the hierarchical structure of discourse is relevant for the resolution.

1. Background

Focusing is a process during discourse interpretation in which participants center their attention on particular discourse entities[1]. As a result, some entities are considered more focused than others at a given time. Since a pronoun or a zero pronoun is normally used to refer to an entity that is 'in focus' (Gundel, Hedberg, and Zacharski 1993), the process of focusing, when implemented in a natural language processing (NLP) system, should limit the number of possible discourse referents for a pronoun or a zero pronoun, or even provide preferred referents for them. Inference mechanisms may then be used to confirm or reject the antecedent. A system using focusing, in general, is considered more cost-efficient than a system using only inference mechanisms for anaphora resolution (Carter 1987:118)[2].

In many NLP systems such as PAL (see Hirst 1981), SPAR (Carter 1987), and PUNDIT (Dahl and Ball 1989), focused entities are primarily selected from entities in the immediately preceding sentence. This approach works well in languages where an antecedent of a pronoun or a zero pronoun is usually found in the immediately preceding sentence. Nevertheless, in the Thai language, an antecedent of a zero pronoun sometimes may not be found in the immediately preceding sentence. It may be, in fact, found in a distant sentence. According to Grima (1986), a zero pronoun in Thai can be separated from its antecedent by many sentences (over one hundred words). He argues that the relationship between zero pronouns and their antecedents should be explained on the basis of discourse structure.

In this paper, we provide evidences to support Grima's argument that reference between zero pronouns in Thai and their antecedents should be described on the basis of hierarchical structure of discourse. We then extend the Centering Theory to work with hierarchical structure of discourse.

2. Centering Theory

Centering Theory (Grosz et al. 1983, 1995, Walker, Iida, and Cote 1990, 1994, Kameyama 1985, 1986, Iida 1997, Eugenio 1990, 1996, 1997, and Strube and Hahn 1996) is a computational model that accounts for local coherence in a discourse segment. It is claimed to work in a discourse segment as defined in Grosz and Sidner's (1986) discourse structure theory.

Centering Theory explains 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). Cb is regarded as the center of attention of the utterance while Cf is an ordered list of discourse entities realized in the utterance. Elements in Cf are partially ordered according to discourse salience (Grosz et al. 1995:209). One entity in the Cf would become the Cb of the utterance. It is defined as the highest ranked entity of the immediately preceding utterance's Cf 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 (Cb) of the next utterance. Thus, the highest-ranked entity in the Cf is regarded as the preferred center for the next utterance (Cp). Constraints and rules in Centering Theory are stated below:

(1)
Constraints:
For each utterance U_i in a discourse segment $U_1,...,U_n$:
1. There is precisely one backward-looking center Cb.
2. Every element of the forward centers list, $Cf(U_i)$, must be realised 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. (Walker et al. 1994:198)

(2)
RULE 1: If any element of $Cf(U_n)$ is realized by a pronoun in U_{n+1}, then the $Cb(U_{n+1})$ must be realized by a 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.

(Grosz et al. 1995:214-215)

'Continuation', 'retaining', and 'shifting' are transition states between a pair of utterances. They are determined from the realization of Cb and Cp as below:

(3)

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

A pair of utterances, Ui and Ui-1, is continuous when both utterances have the same Cb and Cb of Ui is the same as Cp of Ui. Continuation represents a transition state in which the center of attention is the same in both utterances. In example (4), Cb in each utterance is the first element of the previous Cf, that is realized in the current utterance. Thus, the Cb of (4b), (4c), and (4d) is 'John', while the Cb of (4e) is 'Mike'. The transition state between (4b) and (4c) is continuation since the Cbs of both utterances are the same and the Cb of (4c) is the same as the Cp of (4c). When Cb of Ui is different from Cp of Ui while both Ui-1 and Ui have the same Cb, the transition state is called 'retaining'. Retaining represents a transition state in which the center of attention is retained in the current utterance (Ui) but it is likely to be changed in the next utterance. In example (4), the transition state between (4c) and (4d) is retaining since the Cb of both utterances are the same but the Cb of (4d), 'John', is different from the Cp of (4d), 'Mike'. When Cb of Ui is different from Cb of Ui-1, shifting of attention occurs. In (4e), shifting occurs because the Cb of (4e) is not the same as the Cb of (4d). Although Grosz et al. (1983, 1995) does not distinguish between 'smooth-shift' and 'rough-shift', the difference between these two shifting has been discussed in Brennan et al. (1987) and in Walker et al. (1994). We will follow Brennan et al. (1987) and Walker et al. (1994) in using two kinds of shifting in this study.

(4) a. John has been having a lot of trouble arranging his vacation.
b. He cannot find anyone to take over his responsibilities.
Cb = 'John'; Cf = (John)
c. He called up Mike yesterday to work out a plan.
Cb = 'John'; Cf = (John, Mike); Continue
d. Mike has annoyed him a lot recently.
Cb = 'John'; Cf = (Mike, John); Retain
e. He called John at 5 am on Friday last week.
Cb = 'Mike'; Cf = (Mike, John); Shift

(Grosz et al. 1995:217)

Centering Theory can be used for pronoun resolution because it keeps track of discourse salience by using some registers like Cb and Cf. In addition, the first centering rule states that if an entity from the immediately preceding utterance is realized as a pronoun, the Cb of the current utterance must be realized as a pronoun too. And the Cb of an utterance is the highest ranked entity of the previous Cf, that is realized in the current utterance. Thus, if we assume that antecedents of pronouns are found in the immediately preceding utterance, we can use the Centering Theory for pronoun resolution.

When an utterance (Ui) has one zero pronoun, if we assume that its referent could be found in the immediately preceding utterance (Ui-1), according to Rule 1, it is not possible for the Cb to be any other entity else besides the referent of the zero. When an utterance (Ui) has more than one zero, if we assume that all the referents could be found in the immediately preceding utterance (Ui-1), one of the referents must be the Cb. Therefore, algorithms for pronoun resolution can suggest a referent of a pronoun from the Cf(Ui-1) with respect to the ranking of entities in the Cf(Ui-1) and the preference of transition states. The centering 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. For example, to observe Constraint 3, the referent of one zero must be the Cb which is the highest ranked entity from the previous Cf. To observe Rule 2, sequences of continuation is preferred to retaining to shifting.

Centering Theory has been used not only with English but with other languages like Japanese (Walker, Iida, and Cote 1990, 1994, Kameyama 1985, 1986, Iida 1997), Italian (Eugenio 1990, 1996, 1997), and German (Strube and Hahn 1996). In addition, Centering Theory has been adapted for languages which have zero pronouns like Japanese and Italian. The same constraints and rules are proposed and applied for these languages.

Therefore, we will use Centering Theory as the basis of the focusing model to be adopted in this study. But Centering Theory is still not sufficient for handling naturally occurring data. Centering Theory cannot be used for pronoun resolution when the antecedents of these pronouns are not in the immediately preceding utterance. Centering Theory can be used for pronoun resolution if the antecedents of the pronouns can be found in the immediately preceding utterance. But when no pronoun has an antecedent in the immediately preceding utterance (Un-1), Rule 1 is not applicable. This means that the referent of a pronoun does not have to be the Cb of the current utterance (Un). In addition, when no entity from Un-1 is realized in Un, the Cb(Un) is undefined. In these cases, we cannot use constraints and rules in the theory for pronoun resolution. If the referent of a pronoun is referred to in the immediately preceding utterance, we can use Rule 1 to resolve the pronoun as the Cb of the utterance. But when the referent of a pronoun is not referred to in the immediately preceding utterance, the theory gives us no clue where the antecedent of the pronoun should be.

A question to be concerned with here is whether an antecedent of a pronoun always has to be found in the immediately preceding utterance. In Thai language, antecedents of zero pronouns do not always occur in the immediately preceding utterance (see

Section 3). Thus, Centering Theory has to be extended to handle resolution of zero pronouns with distant antecedents as in the case in the Thai language.

3. Discourse structure and zero pronouns in Thai

Although the discourse analyzed by Grima is a text written in the early part of the twentieth century, yet the long distance anaphora may also be found in contemporary Thai texts. Example (5), taken from a news report, suggests that the structure of discourse can be an important factor for zero pronoun resolution in Thai.

(5)

#1 phútthásàmaakhom 30 ʔoŋkɔɔn mii mátì hɛ̌nphóoŋ kan
Buddhist-Society 30 organization have decision agree RECP
'Thirty of the Bhuddhist Societies agreed'

#2thîi ø cà kràapbaŋkhomthuunthàwăaj năŋsŭu pèètphànùk tɔ̀ɔ sŏm dɛ̀tphrásăŋkhárâat thîi mii cajkhwaam doojjɔ̂ɔ wâa
COMP ø=Buddhist-society will give letter open to the-Supreme-Patriarch COMP have content in-brief that
'that they will send an open letter to the Supreme Patriach, which has the content in brief as follow:

#3 ø khɔ̌ɔ hâj ø
thoŋphícaaránaa damnəənkaan sàsăaŋ kɔɔránii wátthammákaaj
ø=Buddhist-society ask let ø= the-Supreme-Patriach consider do clear case Dhammakaya
'They would ask the Supreme Patriach to clear the Dhammakaya case'

#4phró mátì máhăarésàmaakhom thîi ʔɔ̀ɔk maa jaŋ mâj khrɔ̂ɔpkhlum
because decision the-Sangha-Council COMP out ASP still not cover
'because the decision of the Sangha Council that was out is not enough'

#5 ø khàat mâattàkaan thîi pɛn rûuppàtham
ø=the-decision lack measurement COMP be concrete
'It lacks a measurement that is concrete.'

#6 phɯâa hâj kèèt pràsìtthíphâap taam mátì
so-that let occur efficiency follow decision
'To make the decision work,'

#7 thaaŋ phútthásàmaakhom
cɯŋ dâj sànɛ̌ɛ nɛɛwthaaŋ daŋtɔ̀ɔpajníi khɯɯ
for Buddhist-Society thus ASP suggest solution as-follow that-is
'the Buddhist Societies then propose the following methods:'

#8 phráwínítchăj khɔ̌ɔŋ sŏmdɛ̀tphrásăŋkhàrâat
decision of the-Supreme-Patriach
'As for the decision of the Supreme Patriach,'

#9 sûŋ ø mii kaan rábùthŭŋ kaan bìtbɯaan phútthátham khamsɔ̌ɔn
CONJ ø=the-decision have NOM indicate NOM distort Buddha instruction
'which has mentioned about the distort of Buddha's instructions.'

#10 ø thamhâj sŏŋ tèɛkjêɛk
ø=the-distort cause monk disruption
'which caused the disruption among monks'
#11lé kaan môɔp sŏmbàt tháŋmòt thîi kèət khûn najkhànàthîi ø p ɛn phrá hâj kèɛ wát nán
and NOM return property all COMP occur ASP while
ø=Dhammachaiyo be monk give to temple DEM
'and the returning of all properties, that have been processed since (he) has been a monk, to the temple'
#12máhăarésàmaakhom khuaancà damnəənkaan taam phráwínítchăj khŏɔŋ sŏmdɛ̀tphrásăŋkhárâat
the-Sangha-Council should do follow decision of the-Supreme-Patriach
'The-Sangha-Council should follow the decision of the Supreme Patriach'
#13 ø khŏɔ hâj tâŋ kammákaan ruâam ráwàaŋ phrá kàp khàraawâat
ø= Buddhist-Society ask let set committee join between monk and layman
'The Buddhist Societies ask for the setting of joint commettees between monks and laymen'
#14 phɯâa ø tìttaam duulɛɛ kìtcàkam wátthammákaaj
so-that ø=the committee follow look activity Wat-Dhammakaya
'So the committe can follow Wat Dhammakaya's activities.'
(Thairath, April 6, 1999, p.14)

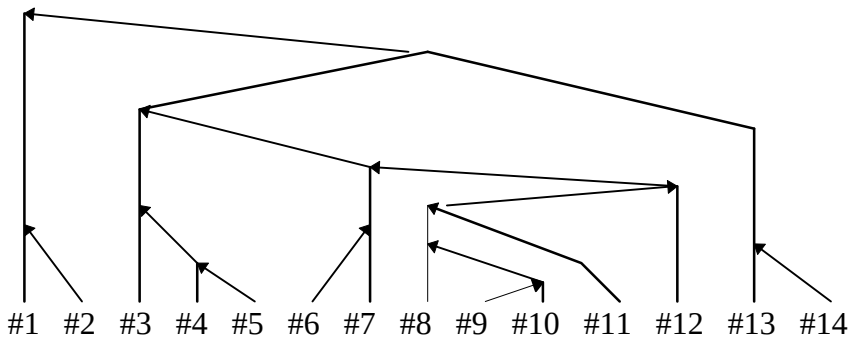


Figure 1: discourse structure of example (5)

The discourse structure of example (5) could be analyzed as shown in Figure1. The structure represents rhetorical relations between rhetorical units, which can be either a clause or a group of clauses. The arrow line represents a relation between satellite and nucleus units while a group of straight lines represents a multi-nucleus relation. The names of rhetorical relations are ignored here since they are not relevant for the discussion in this study. (for further information on Rhetorical Structure Theory see Mann and Thompson (1987)).

In example (5), the antecedent of the zero pronoun in (#13) is not found in the closest utterance (#12), but it is found in a distant utterance, i.e. (#1), (#2), (#3), or (#7). But if we consider the structure of discourse, the antecedent is found in (#3), which is on the same level as (#13). Without considering the structure of discourse, any focusing

algorithms that are primarily based on linear order of sentences would suggest an incorrect preferred referent. For example, Centering Theory (Grosz et al. 1983, 1995) would fail to resolve this zero pronoun because it keeps track of entities only in the previous sentence.

4. Extending Centering Theory

In this section, we propose a model of extended centering that incorporates the structure of discourse. The extension is based on Fox's analysis of English written texts (Fox 1987), in which the 'active', 'controlling', and 'return pop' patterns as conditions for using a pronoun are proposed.

4.1 RST and pronoun resolution

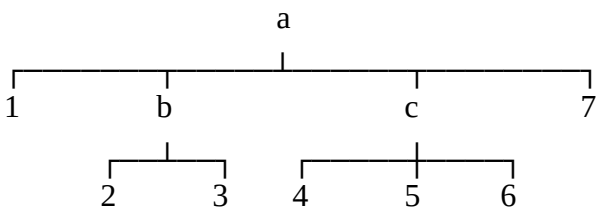
Fox uses Rhetorical Structure Theory (Mann and Thompson 1987) to identify discourse structures of four American-English texts. She finds that the use of pronouns is constrainted by three patterns: 'active', 'controlling', and 'return pop' patterns. Although she identifies both rhetorical structure and rhetorical relations on the texts, the condition of these patterns does not involve rhetorical relation names. Therefore, we will simplify the structure of discourse by ignoring rhetorical relation names in the structure. A discourse structure, then, will look like a hierarchical tree structure. The smallest span (or a clause) would be represented as a terminal node, while a larger span would be represented as a non-terminal node. For example, the discourse (6) below will be represented as a hierarchical structure in (7).

(6)

u1. Bob "Smitty" Smith will be installed as the 1984 president of the Monrovia Chamber of Commerce at the annual January dinner.
u2. He has been a partner in the Monrovia Travel Agency with Bob Bennett since 1974,
u3. but after the first year, when Bennett retires, Smith will become the sole power.
u4. An 11-year member of the Chamber, Smith serves on the Ambassadors committee, a group which systematically visits the over 600 members of the Chamber in a series of two- or three-day "blitzes".
u5. He has served on the Chamber's Board of Directors for three years,
u6. and he is a member of the public relation committee.
u7. His most recent community involvement has been appointment to the Centennial Committee .....(*Monrovia Today*, January 1984)

(Fox 1987:103)

(7)



To summarize Fox's findings of her expository prose data, the following patterns are identified by Fox:

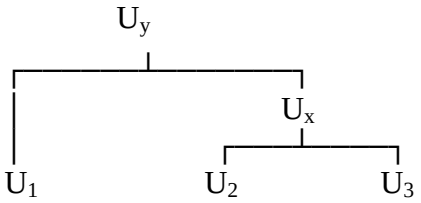
(8)

1. Active pattern: A proposition is active when its rhetorical structure (R-structure) partner is being produced.
2. Controlling pattern: A proposition is controlling when its R-structure partner is active.
3. Return pop pattern: Return pop is a case where a proposition is tied to a proposition which is not the immediately previous proposition.

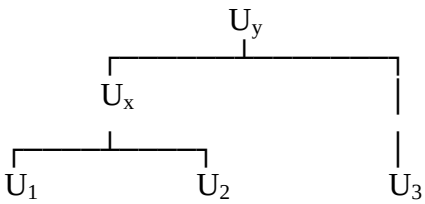
According to Fox, antecedents of pronouns will be found in an active, a controlling, or a return pop pattern. Thus, we can use these patterns to identify the scope where focusing algorithms should look for referents of pronouns. In an active pattern, a pronoun can be used in U_i when its referent is referred in an active unit, which is also the immediately preceding unit (U_{i-1}) of the current utterance (U_i). Thus, the immediately preceding unit will be the scope in which focusing algorithms should look for antecedents. And the immediately preceding unit can be either a clause or a rhetorical structure.

In a controlling pattern, antecedents of pronouns are in a controlling proposition. A controlling pattern presented in Fox (1987:100-101) can be represented as two types of a hierarchical structure, as shown in (9) and (10). In these examples, a pronoun is in U_3 and its antecedent is in U_1. U_2 is the active proposition and U_1 is the controlling proposition. In (9), a controlling proposition is an immediately preceding unit of U_x, which is composed of U_2 and U_3. The active proposition (U_2) is the immediately preceding unit of the current utterance (U_3). Thus, to handle this case, focusing algorithms have to be implemented to look back two steps. When the referent cannot be found in the immediately preceding unit (U_2), the next preceding unit (U_1) should be the scope in which the focusing algorithm looks for the referent. In (10), a controlling proposition (U_1) is combined to the active proposition (U_2) as a larger unit (U_x), and this unit (U_x) is the immediately preceding unit of the current utterance (U_3). Thus, focusing algorithms can find the referent of pronoun in one step within this immediately preceding unit (U_x).

(9)



(10)



In a return pop pattern, though antecedents of pronouns are found in a distant unit, Fox notices that a pronoun can be used in U_i when its referent is referred to not only in the nucleus of U_i, but also in an adjunct unit between U_i and the nucleus. However, since the presence of the referent in the previous adjunct is necessary for a pronoun to be used, we prefer to view the use of a pronoun in this case as a continuation of the focus or the center (Cb) between units rather than as a direct connection between a nucleus and a return-pop as implied by Fox. For example, in example (6), we view the use of pronoun in (u7) as a continuation of the Cb from unit (uc) to (u7) rather than a reference between the nucleus (u1) and the return pop (u7). In this example, (uc) is the previous adjacent unit of (u7). (uc) composes of three utterances, (u4)-(u6). In (u6), a pronoun *he* can be used because (u5) contains mentions of the referent 'Smith'. In (u5), a pronoun *he* can be used because (u4) contains mentions of the referent 'Smith'. As a result of the continuation of the Cb ('Smith') from (u4) to (u6), it can be assumed here that 'Smith' will be the Cb of the unit (uc). Then, the use of a pronoun in (u7) to refer to 'Smith' is just a continuation of the Cb from (uc) to (u7).

It appears that both 'active' and 'return-pop' patterns require the presence of a referent in the preceding unit. Thus, the immediately preceding unit should be the first place the focusing algorithms look for an antecedent. If antecedents of pronouns are not in the immediately preceding unit, the focusing algorithms then will look for antecedents in the next preceding unit. The latter operation is the implementation of the first type of controlling pattern. However, according to Fox (1987:102), the controlling pattern is found only 2 percents of her data. Thus, if Fox's finding is also true in other written proses, the focusing algorithms are likely to find antecedents of pronouns in the immediately preceding unit most of the times.

4.2 An extended centering model

By taking discourse structure into account, the centering model can be extended to work beyond the scope of discourse segment along the following lines. Given that a discourse is analyzed as a hierarchical structure of discourse units (U_i), we propose that each discourse unit has one backward-looking center (Cb) and a set of forward-looking centers (Cf). A discourse unit can be either a single utterance or multiple utterances. Cf is an ordered list of discourse entities realized in that unit. We now modify constraints and rules of the Centering Theory as shown in (1) and (2) earlier as follows:

(11)
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.

(12)
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: Sequences of continuation are preferred over sequences of retaining; and sequences of retaining are to be preferred over sequences of shifting.

The above constraints and rules require a different interpretation of subscriptions. Given that U_i is a current discourse unit, U_{i-1} is an immediately preceding discourse unit and U_{i+1} is a following discourse unit. The definition of precedence is defined as follow:

(13)
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 has the same parent, U_k, and there is no other unit between U_i and U_j, and U_i occurs before U_j.

The constraints and rules above are variants of those defined in the Centering Theory. In this new extended model, the immediately preceding unit of U_i, is not necessarily the immediately preceding utterance of U_i, but the immediately preceding discourse unit. U_i is the immediately preceding unit of U_j when both U_i and U_j have the same parent and U_i is the left adjacent unit of U_j; or when U_j is the left most unit under U_k and U_k is the right adjacent unit of U_i. In example (6), the immediately preceding unit of (u4) is (ub) not (u3); and the immediately preceding unit of (u7) is (uc). Thus, the pronoun *he* in (u2) will be resolved with 'Smith' referred in the immediately preceding unit, (u1). The pronoun *he* in (u5) and (u6) will be resolved with 'Smith' referred to in the units (u4) and (u5) respectively. *His* in (u7) will be resolved with 'Smith' referred in the unit (uc).

4.3 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 (Cb_i) should remain the same as the previous one (Cb_{i-1}) and whether the preferred center (Cp_i) should be the same as the center (Cb_i). 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):

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

It is generally assumed that thie 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.

Continuation $Cb_{i-1} = Cb_{i-2}$ $Cb_{i-1} = Cp_{i-1}$	Continuation >> $Cb_i = Cb_{i-1}$ $Cb_i = Cp_i$	Retaining >> $Cb_i = Cb_{i-1}$ $Cb_i \neq Cp_i$	Smooth-shift >> $Cb_i \neq Cb_{i-1}$ $Cb_i = Cp_i$	Rough-shift $Cb_i \neq Cb_{i-1}$ $Cb_i \neq Cp_i$
	$Cb_i = Cp_{i-1}$	$Cb_i = Cp_{i-1}$	$Cb_i \neq Cp_{i-1}$	$Cb_i \neq Cp_{i-1}$
Retaining $Cb_{i-1} = Cb_{i-2}$ $Cb_{i-1} \neq Cp_{i-1}$	Smooth-shift >> $Cb_i \neq Cb_{i-1}$ $Cb_i = Cp_i$	Rough-shift >> $Cb_i \neq Cb_{i-1}$ $Cb_i \neq Cp_i$	Continuation >> $Cb_i = Cb_{i-1}$ $Cb_i = Cp_i$	Retaining >> $Cb_i = Cb_{i-1}$ $Cb_i \neq Cp_i$
	Cb_i ... Cp_{i-1}	Cb_i ... Cp_{i-1}	$Cb_i \neq Cp_{i-1}$	$Cb_i \neq Cp_{i-1}$
Smooth-shift $Cb_{i-1} \neq Cb_{i-2}$ $Cb_{i-1} = Cp_{i-1}$	Continuation >> $Cb_i = Cb_{i-1}$ $Cb_i = Cp_i$	Retaining >> $Cb_i = Cb_{i-1}$ $Cb_i \neq Cp_i$	Smooth-shift >> $Cb_i \neq Cb_{i-1}$ $Cb_i = Cp_i$	Rough-shift $Cb_i \neq Cb_{i-1}$ $Cb_i \neq Cp_i$
	$Cb_i = Cp_{i-1}$	$Cb_i = Cp_{i-1}$	$Cb_i \neq Cp_{i-1}$	$Cb_i \neq Cp_{i-1}$
Rough-shift $Cb_{i-1} \neq Cb_{i-2}$ $Cb_{i-1} \neq Cp_{i-1}$	Smooth-shift >> $Cb_i \neq Cb_{i-1}$ $Cb_i = Cp_i$	Rough-shift >> $Cb_i \neq Cb_{i-1}$ $Cb_i \neq Cp_i$	Continuation >> $Cb_i = Cb_{i-1}$ $Cb_i = Cp_i$	Retaining >> $Cb_i = Cb_{i-1}$ $Cb_i \neq Cp_i$
	Cb_i ... Cp_{i-1}	Cb_i ... Cp_{i-1}	$Cb_i \neq Cp_{i-1}$	$Cb_i \neq Cp_{i-1}$

Figure 2: Preferences of transition states

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 .

4.4 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:

(15)

Grosz et al. (1995:214) English
SUBJECT >> OBJECT(S) >> OTHER

Kameyama (1985:115) Japanese
TOPIC >> SUBJECT >> OBJECT(2) >> others
ADJUNCT)

Walker et al. (1990:3) Japanese
TOPIC >> EMPATHY >> SUBJ >> OBJ2 >> OBJ

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.[3] The Cf ranking for Thai used in this study is in (16).

(16) TOPIC >> SUBJ >> OBJ(S) >> OTHERS

5. Testing on Thai corpus[4]

This section 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. The corpus is limited to expository discourse because we are interested in discourse anaphora and a study by Maneeroje (1985) showed a good rate of zero pronouns in this type of discourse.[5] Twenty expository prose texts were selected. Altogether 719 zero pronouns are marked in these twenty texts (15,949 total number of words). The corpus is analyzed by twelve native speakers of Thai. They were asked to perform two tasks: one, to identify referents of zero pronouns and two, to identify the structure of discourse. The referents agreed by the majority of the subjects are regarded as 'correct' referents.

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.

5.1 Scope of the test

Not all zeroes marked on the corpus will be handled by the centering algorithms. First to be excluded are zeroes which the majority of subjects fail to agree on their referents. Next are deictic zeroes, zeroes with abstract referents, and zeroes whose referents are analyzed as 'unidentified' referents. In addition, zeroes in embedded clauses will be excluded from the test because the centering algorithms are not designed to resolve zeroes whose referents are in the same utterance. Of the 719 zeroes marked on the corpus, referents of 40 cannot be agreed on by the majority of subjects. 82 zeroes are analyzed as having 'unidentified' referents; 133 are used as deictic zeroes; 34 are in embedded clauses; and 12 are used to refer to abstract referents. Therefore, only 418 zeroes will be tested by the centering algorithms.

5.2 Results

The results results of the test indicate 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. Table 1 below is the summary of the result of the test.

	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 1: Results of extended centering and existing centering

Since the centering algorithms only used the immediately preceding unit for resolving zeroes, it did not account for zero pronouns in a controlling pattern as discussed in section 4.1. In a pattern like example (9), the antecedent can be in (u1), which is not covered in the scope of referent resolution. Therefore, we modified the extended and the existing algorithms so that the algorithms can search the next preceding unit for the referent when the 'correct' referent was not in the immediately preceding unit. Then we tested the two algorithms again. The result of the new testing is shown in Table 2 below:

	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 2: Results of extended and existing centerings with two-step look back

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 3 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 3: 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 3, 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 4 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 4: Attempts counted in existing and extended centerings

The extended centering did not perform better than the existing cenetring 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 discuss whether hierarchical structure of discourse is an important factor for the resolution of these zero pronouns.

5.3 Discussions

There are 22 instances of zeroes that were resolved by the extended centering but could not be resolved by the existing centering. 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.[6] Consider examples (17)-(19) below.

(17) Text from Health1

#23 daŋnán thâa suàan tàaŋtàaŋ khŏoŋ hŭu dâjráp kaan k ràthópkràthɯaan

therefore if part any of ear$_{Z11}$ receive NOM impact$_{Z17}$

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

#24 rŭu [Z11] dâjráp chɯáarôok

or [Z11]$_{Z12}$ receive germ

‘or (it) receives germ’
#25 chêen [Z12] doon kràthêɛk rɛɛŋrɛɛŋ
such-as [Z12]$_{Z13}$ get strike strong
‘For example, (it) is struck’
#26 [Z13] dâjjin siăaŋ daŋ mâak
[Z13] hear voice loud very
‘(It) is imposed to very loud noise’
#27 mii nám khâw hŭu
there-is water enter ear$_{Z14}$
‘Water enters ears’
#28 [Z14] dâjráp chuáarôok
[Z14] receive germ
‘(Ears) receive germ’
#29 phró [Z15] khɛ́ hŭu duâaj khruâaŋmuu sòkkàpròk
because [Z15] pick ear with instrument dirty
‘because (we) pick ears with dirty instrument’
#30 [Z16] khɛ́ hŭu rɛɛŋ kəən paj
[Z16] pick ear strong over ASP
‘(we) pick ears too hard’
#31 [Z17] ʔàat thamhâj kèət rôok hŭu 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.

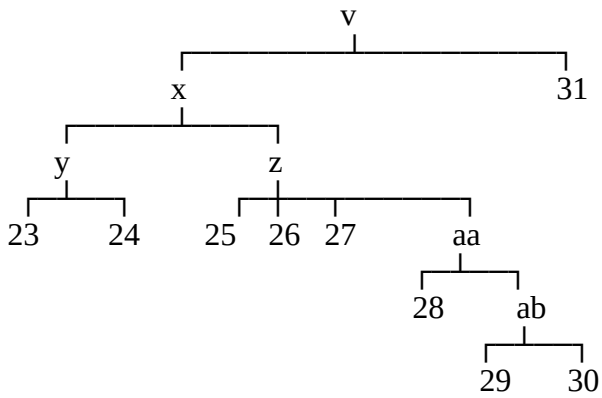


Figure 2: Hierarchical structure of example (17)

In examples (17), 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).

(18) Text from Health1

#80 cà kèət kaan?àksèep naj chôɔŋ hŭu chánklaaŋ
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 siăaŋ nɔ́ɔj loŋ
[Z51] will hear sound decrease ASP
'(Ear) will receive less sound'

#82 rŭu kèət kêɛwhŭu thálú
or occur eardrum torn
'Or, eardrum is torn'

#83 mii námnɔ̆ɔŋ lăj
there-is lymph out
'Lymph comes out'

#84 hàak [Z52] pɛn 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, ...

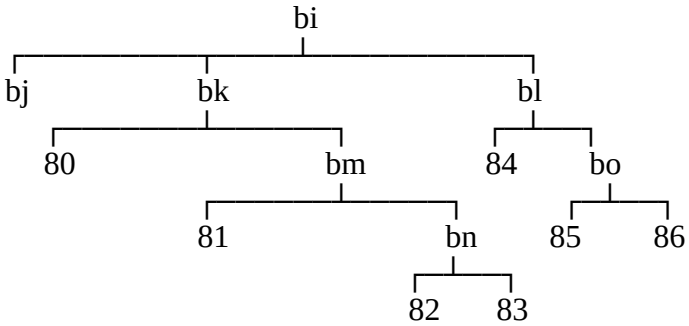


Figure 3: Hierarchical structure of example (18)

In examples (18), 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 concisting 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.

(19) Text from Editor2.txt

#5
kaanluâaktâŋ thuâapaj naj kamphuuchaa khráŋ níi càt khûn m aa dooj khwaamphájaajaam khɔ̆ɔŋ 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 khàměensìifàaj loŋnaam naj khɔ̂ɔtòkloŋ săntìphâap thîi kruŋpaarîit pràthêetfàrâŋsèet mɯ̂a pii phɔɔsɔ̌ɔ 2534

COMP [Z1] ASP make group <u>four-party</u>$_{Z2}$ sign in agreement peace at Paris France in year 1991

'who made the four parties sign the peace agreement in Paris in 1991'

#7 lɛ́ɛw [Z2] damnəənkaan taam nɛɛwthaaŋ thîi kamnòt wâj

then [Z2] practice along guideline COMP outline ASP

'and (the four party) to follow the guideline that is outlined'

#8 phɯ̂a hâj kèət săntìphâap jàaŋ thăawəən

for give occur peace Adv-Mrk permanent

'so that peace could occur permanently'

#9 dooj [Z3] cháj kràbuaankaan thaaŋ kaanmɯaaŋ naj ráb̀ɔ̀ppràchaathípp̀ataj khɯu kaanluâaktâŋ 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] mă̌aj hâj khàměen tɛ̀ɛlá klùm sòŋ tuaathɛɛn loŋsàmàkrápluâaktâŋ

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

'(The UN) wants each group of Khamer to send representatives in the election'

#11 lɛ́ɛw [Z5] hâj pràchaachon chaawkhàměen pɛn 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.'

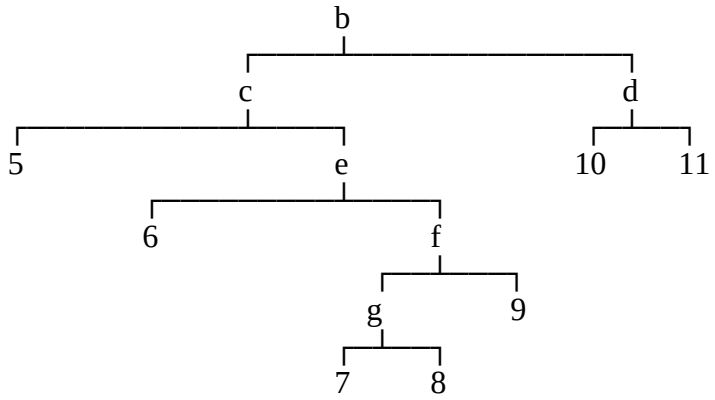


Figure 4: Hierarchical structure of example (19)

In example (19), 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.

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.

[1]A 'discourse entity' is an entity that is evoked from the discourse context (Webber 1981). Sometimes, the term 'discourse referent' is used. In this paper, these two terms are interchangeable. A 'discourse referent' is used when reference is involved. The term ' referent' used in this paper refers to a discourse referent (Karttunen 1976).

[2] It does not necessary mean that focusing algorithms will always suggest correct antecedent for a zero pronoun or a pronoun. But we expect that a good focusing algorithm should be able to suggest correct antecedent as a preferred referent as much as possible.

[3] Thai is a topic prominent language.

[4]I am indebt to NECTEC who provides the corpus for this study.

[5] Maneeroje studies the use of four NP forms (zeros, repeated NPs, demonstrative NPs, and pronouns) as a cohesive device in ten written Thai texts. Zeroes are found to be the most frequently used form (49.88%) while pronouns are the least frequently used (5.90%)

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

References

Allen, James. (1995). *Natural Language Understanding.* Redwood City: Benjamin/Cummings.

Brennan, Susan E., M.W. Friedman, and C.J. Pollard. (1987). A Centering Approach to Pronouns. In *25th Annual Meeting of the Association of Computational Linguistics*, (pp. 155-162). Stanford.

Carter, David. (1987). *Interpreting Anaphors in Natural Language Texts.* New York: Ellis Horwood Limited.

Cohen, Robin. (1987). Analyzing the Structure of Argumentative Discourse. *Computational Linguistics 13:1-2.* 11-24.

Dahl, Deborah A., and Catherine N. Ball. (1989). Reference Resolution in PUNDIT. In *Logic and Logic Grammar for Language Processing,* Chichester, ed. by P. Saint-Dizier and S. Szpakowicz, 168-184. Chichester: Ellis Horwood.

Eugenio, Barbara Di. (1997). Centering in Italian. In *Centering in Discourse,* ed. by Walker, Joshi and Prince. Oxford University Press.

Eugenio, Barbara Di. (1996). The Discourse Functions of Italian Subjects: A Centering Approach. In *Proceedings COLING96.*

Eugenio, Barbara Di. (1990). Centering Theory and the Italian Pronominal System. In *COLING-90, vol. II.* 270-275.

Fox, Barbara A. (1987). *Discourse Structure and Anaphora: Written and Conversational English.* Cambridge: Cambridge University Press.

Gordon P.C., B.J. Grosz, and L.A. Gilliom. (1993). Pronouns, Names, and the Centering of Attention in Discourse. *Cognitive Science 17.* 311-347.

Grima, John A. (1986). Discourse Factors Contributing to the Understanding of a Zero Pronoun in a Passage from the phra raat cha wi caan. In *Papers from a Conference on Thai Studies in Honor of William J . Gedney,* ed. by Robert J.

Bickner et al., 159-169. Ann Arbor: Center for South and Southeast Asian Studies, University of Michigan.

Grosz, Barbara J., A.K. Joshi, and S. Weinstein. (1995). Centering: A Framework for Modelling the Local Coherence of Discourse. *Computational Linguistics, 21:2,* 203-225.

Grosz, Barbara J., A.K. Joshi, and S. Weinsteins. (1983). Providing a Unified Account of Definite Noun Phrases in Discourse. In *Proceedings of ACL,* 44-50.

Grosz, Barbara J. and Candance L. Sidner. (1986). Attention, Intentions, and the Structure of Discourse. *Computational Linguistics, 12:3,* 175-204.

Gundel J.K., N. Hedberg, and R. Zacharski. (1993). Cognitive Status and the Form of Referring Expressions in Discourse. *Language 69:2,* 274-307.

Hirst, Graeme. (1981). Discourse-Oriented Anaphora Resolution in natural Language Understanding: A Review. *American Journal of Computational Linguistics, 7:2,* 85-98.

Hobbs, Jerry R. (1985). *On the Coherence and Structure of Discourse.* Standford: CSLI.

Iida, Masayo. (1997). Discourse Coherence and Shifting Centers in Japanese Texts. In *Centering in Discourse,* ed. by Walker, Joshi and Prince. Oxford University Press.

Kameyama, Megumi. (1986). Japanese Zero pronominal Binding: Where Syntax and Discourse Meet. In *Papers from the Second International Workshop on Japanese Syntax,* ed. by William J. Poser, 47-73. CSLI.

Kameyama, Megumi. (1985). Zero Anaphora: The Case of Japanese. Ph.D. Dissertation, Stanford University.

Karttunen, Lauri. (1976). Discourse Referents. in Syntax and Semantics 7: *Notes from the Linguistic Underground,* ed. by James D. McCawley, 363-383. New York: Academic Press.

Maneeroje, Prapaporn. (1985). The use of Noun-Substitues as Cohesive Devices in Thai. Master Thesis, Chulalongkorn University. Bangkok, Thailand.

Mann, William C., and Sandra A. Thompson. (1987). Rhetorical Structure Theory: A Theory of Text Organization. *ISI/RS*-87-190.

Polanyi, Livia. (1988). A Formal Model of the Structure of Discourse. *Journal of Pragmatics, 12,* 601-638.

Strube, Michael and Udo Hahn. (1996). Functional Centering. In *ACL '96: Proc. of the 34th Annual Meeting of the Association for Computational Linguistics,* Santa Cruz, CA.

Walker, Marilyn, M. Iida, and S. Cote. (1994). Japanese Discourse and the Process of Centering. *Computational Linguistics, 20:2,* 193-231.

Walker, Marilyn, M. Iida, and S. Cote. (1990). Centering in Japanese Discourse. In *COLING-90, vol. I* 1-6.

Webber, Bonnie Lynn. (1981). Discourse Model Synthesis: Preliminaries to Reference. In Elements of Discourse Understanding, ed. by Aravind K. Joshi et al., 283-299.Cambridge: Cambridge University Press.