

EXTENDING FOCUSING
FOR ZERO PRONOUN RESOLUTION IN THAI

VOLUME ONE OF TWO

A Dissertation
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy
in Linguistics

By

Wirote Aroonmanakun, M.A.

Washington, D.C.
March 24, 1999

EXTENDING FOCUSING FOR ZERO PRONOUN RESOLUTION IN THAI

Wirote Aroonmanakun, M.A.

Mentor: Catherine N. Ball, Ph.D.

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.

ACKNOWLEDGEMENTS

I am indebted to many people who have contributed to this dissertation, both directly and indirectly.

First and foremost, I would like to express my gratitude to my mentor, Prof.Catherine Ball, for her valuable suggestions, and her endless patience and thorough supervision, without which this dissertation would never have been completed. I am indebted to all the committee members, Prof.Ralph Fasold and Prof.Paul Portner, for their helpful and valuable comments and suggestions. Discussions with them has led to many important developments of this dissertation. My thanks also go to Prof.Jeff Corner-Linton for his valuable comments on the statistics method used in the dissertation. I would like also thank to Prof.Abella Mason of the EFL department for her patience proof reading during the early draft of the dissertation. My appreciation goes also to all my friends and colleagues in the department of Linguistics at Chulalongkorn University for their encouragement.

Many thanks go to my friends, Hyouk-Keun Kim, Margaret Ronkin, Jindaporn Sangganjanavanish, and Manat Homtavil, for their generous helps in many ways. Thanks to Manela for always be kind and helpful for foreign students.

Lastly, I would like to express my gratitude to the Ministry of Science, Technology, and Environment for granting me a 5-year scholarship, and to Linguistics and Knowledge Science Laboratory (LINKS) for providing a Thai corpus of this study. And to my wife, my parents and everyone in my family, I offer my heartfelt thanks, especially to my wife for her endless patience translating the corpus, her understanding, and her thorough supportive.

TABLE OF CONTENTS

I. INTRODUCTION	page 1
1.1 The purpose of this study	1
1.2 Background	2
1.3 Zero pronouns and pronoun system in Thai	13
1.3.1 Pronoun system in Thai	13
1.3.1.1 Personal pronouns	13
1.3.1.2 Zero pronouns	15
1.3.2 Usage of zero pronouns	17
1.3.2.1 Deictic usage	17
1.3.2.2 Anaphoric usage	18
1.3.2.3 Discourse deictic usage	18
1.3.2.4 Generic usage	18
1.4 The scope of this study	19
1.5 Outline	20
 II. EMPTY CATEGORIES IN THAI	 23
2.1 Empty categories in Government and Binding Theory	23
2.1.1 Empty categories	23
2.1.2 Binding theory	25
2.1.3 Bounding theory	27
2.1.4 Control theory	28
2.1.5 Determining ECs and their antecedents	29
2.2 An overview of Thai syntax	31
2.3 Empty categories in Thai	33
2.3.1 Sentence level empty categories	34
2.3.1.1 Relative clauses	34
2.3.1.2 Topicalization and Left-dislocation	39
2.3.1.3 Serial verb constructions	41
2.3.1.3.1 Object sharing.	41
2.3.1.3.2 ECs in serial verb constructions	43
2.3.2 Discourse level empty categories	44
2.3.2.1 Coordination	44
2.3.2.2 Subordination	48
2.4 Conclusion	50

III. FOCUSING AND REFERENCE RESOLUTION	51
3.1 Focus	51
3.2 The role of focus in reference resolution	56
3.3 Focusing algorithms	57
3.3.1 Sidner's focusing algorithms	58
3.3.1.1 Problems	64
3.3.2 Centering Theory	66
3.3.2.1 Problems	74
3.4 Extending focusing algorithms	81
3.4.1 Focusing and Discourse Structure	81
3.4.2 A proposal for an extended centering model	89
3.5 Conclusion	98
IV. CORPUS DESIGN AND ANALYSIS	100
4.1 Corpus design	100
4.2 Corpus preparation	101
4.2.1 Marking of zero pronouns	102
4.2.2 Marking of discourse units	103
4.2.3 Questionnaires	105
4.2.4 The subjects	105
4.3 Corpus analysis	105
4.3.1 Identifying referents	106
4.3.2 Identifying discourse structure	108
4.4 Results of the analysis	108
4.4.1 Results of referent analysis	109
4.4.2 Results of structural analysis	110
4.4.2.1 Agreement on segmentation	111
4.5 Conclusion	117
V. AN EXTENDED CENTERING MODEL FOR THAI	120
5.1 Centering in Thai	120
5.1.1 Constraints and rules	121
5.1.2 Preferences of transition states	123
5.1.3 Cb establishment	127
5.1.4 Ranking of Cf	128
5.1.5 Zero pronoun resolution	131

5.2 An extended centering model for Thai	134
5.2.1 Constraints and rules	134
5.2.2 Cf of multiple utterances	137
5.2.3 Zero pronoun resolution	142
5.3 Conclusion	147
VI: COMPARISON OF CENTERING ALGORITHMS	149
6.1 Testing centering algorithms	149
6.1.1 Input	150
6.1.1.1 Entity list file	150
6.1.1.2 Discourse structure file	152
6.1.1.3 Referent list file	153
6.1.2 Scope of the test	154
6.1.3 Simulation of centering algorithms	155
6.2 The two centering algorithms	157
6.2.1 Existing centering algorithm	157
6.2.2 Extended centering algorithm	159
6.2.3 Algorithm for determining the Cb	160
6.2.4 Algorithm for determining a transition state	161
6.2.5 Algorithm for generating and ranking all possible interpretations	161
6.2.6 Algorithm for constructing the Cf of multiple utterances	163
6.3 Results	164
6.4 Discussions	168
6.4.1 Zero pronouns group III	168
6.4.2 Zero pronouns group IV	175
6.5 Further research	181
6.5.1 Open issues on Centering Theoy	182
6.5.2 Functional centering	184
6.5.3 Nucleus and satellite units	190
6.5.4 Positions on hierarchical structure	195
6.6 Conclusion	197
APPENDIX	199
A. CORPUS	199
A.1 Text Article1	199
A.2 Text Article2	212
A.3 Text Comp1	220
A.4 Text Comp2	232

A.5 Text Comp3	239
A.6 Text Comp4	248
A.7 Text Editor1	260
A.8 Text Editor2	267
A.9 Text Health1	274
A.10 Text Health2	284
A.11 Text Mang1	291
A.12 Text Mang2	311
A.13 Text News1	330
A.14 Text News2	334
A.15 Text News3	338
A.16 Text Story1	347
A.17 Text Tnews1	357
A.18 Text Tnews2	359
A.19 Text Tnews3	361
A.20 Text Tnews4	363
B. QUESTIONNAIRES	367
BIBLIOGRAPHY	369

ABBREVIATION TERMS

Adv-Mrk	Adverb marker
ASP	Aspect
CL	Classifier
COMP	Complementiser
CONJ	Conjunction
DEM	Demonstrative
EMPH	Emphasis
NOM	Nominaliser
Pass-Mrk	Passive marker
PRT	Particle
PROG	Progressive
Q	Question word
QUAN	Quantity
RECP	Reciprocal
REFX	Reflexive

LIST OF TABLES AND FIGURES

Table 1: Inventory of empty categories	24
Table 2: Results of referent analysis	109
Table 3: Degree agreement and identification of referents	110
Table 4: Percent agreement on segmentation	113
Table 5: Guideline for interpreting kappa coefficient	115
Table 6: Segmentation analysis of text Tnews4	116
Table 7: Degree of agreements on segmentation	117
Table 8: Probability of referents found with respect to three factors	138
Table 9: Zero pronouns in the corpus	155
Table 10: Results of existing centering and extended centering	164
Table 11: Results of existing and extended centerings with two-step lookback	165
Table 12: First-try success and attempts in existing and extended centerings	166
Table 13: Attempts counted in existing and extended centerings	167
Table 14: Costs of transition pairs	185
Table 15: Results of non-functional and functional centerings	186
Table 16: Number of continuation and transition cost	189
Table 17: Result of extended centering with hierarchy factor	196
Figure 1: The structure of discourse (1)	6
Figure 2: The structure of discourse in example (2)	10
Figure 3: Rhetorical structure of example (82)	83
Figure 4: Rhetorical structure of example (84)	85
Figure 5: Rhetorical structure of example (85)	85
Figure 6: Rhetorical structure of example (86)	86
Figure 7: Rhetorical structure of example (87)	87
Figure 8: Rhetorical structure of example (88)	88
Figure 9: Discourse structure of example (89)	90
Figure 10: Mapping of the first type of controlling pattern	92
Figure 11: Mapping of the second type of controlling pattern	93
Figure 12: Mapping of the third type of controlling pattern	93
Figure 13: Discourse structure of example (91)	95
Figure 14: Segment boundary	111
Figure 15: Preferences of transition states	125
Figure 16: An example of discourse structure	136
Figure 17: A discourse structure of example (119)	140
Figure 18: Cfs of discourse units in Example (119)	141
Figure 19: A discourse structure of example (121)	144
Figure 20: An example of the discourse structure	153

Figure 21: Hierarchical structure of example (136)	170
Figure 22: Hierarchical structure of example (137)	171
Figure 23: Hierarchical structure of example (138)	172
Figure 24: Hierarchical structure of example (139)	174
Figure 25: Hierarchical structures of example (140)	178
Figure 26: Hierarchical structures of example (141)	179
Figure 27: Hierarchical structure of example (142)	180
Figure 28: Hierarchical structure of example (146)	192
Figure 29: Hierarchical structure of example (147)	194
Figure 30: Hierarchical structure of example (148)	195

Chapter 1

Introduction

1.1 The purpose of this study

Focusing is a process that has been used for anaphora resolution (Grosz 1977, 1981, Grosz and Sidner 1986). It is used to keep track of salience of discourse entities and set up preferred referents for anaphors. One of the focusing models that has been used for anaphora resolution is Centering Theory (Grosz, Joshi, and Weinstein 1983, 1995). Centering Theory has been used for pronoun or zero pronoun resolution in many languages, such as English (Grosz et al. 1983, 1995, Brennan, Friedman, and Pollard 1987), 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). Most of the studies except Iida (1998) and Eugenio (1996, 1998) are based on constructed discourses, and the centering model can resolve only pronouns or zero pronouns whose antecedents are in the immediately preceding utterance. However, current research on Centering Theory is now aiming at extending the theory to work with naturally occurring text (Walker, Joshi, and Prince 1998). This study is another extension of the Centering Theory to account for zero pronoun resolution in naturally-occurring Thai texts. To do this, we first demonstrate the problem that antecedents of zero pronouns in Thai are not always in the immediately

preceding utterance. Rather, some examples suggest that the hierarchical structure of discourse might be relevant for the resolution of zero pronouns in Thai. To investigate whether the hierarchical structure of discourse is relevant for zero pronoun resolution in Thai, we propose an extended model of Centering Theory which can account for the hierarchical structure of discourse, and test it on Thai corpus.

1.2 Background

As discussed by Hirst (1981a) and Sidner (1983), different factors for anaphora resolution have been studied in previous research, including general heuristics (Winograd 1972), syntactic and semantic constraints (Woods et al. 1972, Reinhart 1976), and inference (Hobbs 1976, Wilks 1975). However, according to Hirst (1981a, 1981b) and Sidner (1983), these factors are not effective for anaphora resolution. They suggest that algorithms for anaphora resolution would be more efficient if they include discourse constraints like focusing (Sidner 1983, Grosz et al. 1983, 1995, Walker et al. 1990, 1994, Kameyama 1985, 1986, 1998) and discourse structure (Grosz 1977, 1981, Grosz and Sidner 1986).

Focusing is a process during discourse interpretation in which participants center their attention on particular discourse entities¹. As a result, some entities are

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).

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 1981b), 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

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

words). His analysis of zero pronouns in the phra'raat^cha-wi'caan- text is provided in example (1) below.

- (1) The following is the analysis in Grima (1986:159-163). The discourse is taken from the phra'raat^cha-wi'caan- of Rama V, King Chulalongkorn (1973:54-55). It was written in the early part of the twentieth century. The author, King Chulalongkorn, described the memoir written for Lord of Thonburi.

Abbreviations

ES	Empty subject (independent ø)
CP	Completive particle
IR	Irrealis
KS	Khun Luang sua (a late Ayudhaya king)
LT	Lord of Thonburi
RC	Relative conjunction
Q	Quotative
SP	Sequence-marking particle (also marks predicates)
W	The person who wrote the memoirs
Rama-V	King Chulalongkorn, the author of this text
//	Signals a syntactic boundary more or less equivalent to a sentence boundary

- I. khôokhwaam thîi ø hên chên ní
material RC ø=Rama-V opine like this
- II. phró
because
- II.A phûukhiăan nâpthûu câwkrunthonbùrii //
- W respect LT //
- II.A1a ø riâak ø wâa phêendin tôn
ø=W call ø=LT Q reign first
- II.A1b ø cháj thôjkhām klàaw thŭŋ ø duâaj khwaamkhawróp muăan
jàaŋ lûuklăan câwkrunthonbùrii phûut taamthîi ø dâj khœy
faŋ penʔanmâak //
- ø=W use idiom speak arrive ø=LT with respectfulness same kind
descendants LT speak according-to ø=Rama-V CP ever listen-to a-lot //

II.A2a muâa ø klàaw thŭŋ sǎnjaawípàlâat ø ø kò klàaw duâaj
khwaamhěncaj wâa ø pen kaanbaŋʔəən ø pen paj chēen nán
duâaj ø pen weelaa ø khró kam lé pen weelaa ø cà sîn bun sîn
waasànǎa //

time ø=W speak arrive mental-aberration ø=LT ø=W SP speak with
sympathetic-understanding Q ø=ES to-be accident ø=ES to-be go like that
because ø=ES to-be time ø=LT bad karma and to-be time ø=LT IR out-of
merit out-of merit //

II.A2b muâa ø klàaw thŭŋ kaandùráaj ø ø kò khôonkhâaŋ cà
pen kham jùu khâaŋʔuàatʔuàat wâa ø kèenkaat rŭu ø cajkhoo
dètdiàaw jàaŋ diaaw kan kàp lûuklǎan khŭnluǎaŋsuǎa klàaw
thŭŋ khŭnluǎaŋsuǎa jókjôŋ ø naj kaanthîi ø mii
khwaamhěnluaŋnâa chēen ø rúusùktuua wâa ø sîn bun léew //

muâa khǎw chēen ø hāj ø buàat ø kò jindii priidǎa thîi ø cà
ʔòokbuàat // khrán muâa câw bunmiiraammárâat paj chuaan ø hāj
ø sùk ø kò maʔj joom sùk // ø wâa ø sîn bun léew // ø jàa paj
sûu khǎw ləəy //

time ø=W speak arrive fierceness ø=LT ø=what-she-says SP
almost/appears IR to-be word to-be-located side brag Q ø=LT bold or
ø=LT character decisive kind same reciprocal with descendants KS speak
arrive KS praise ø=KS in clausal-nominalizer ø=KS have foresight an-
instance KS be-aware Q ø=KS out-of merit CP // time 'they' invite ø=KS
give ø=KS become-a-monk ø=KS SP happy happy conjunction ø=KS IR
leave become-a-monk // time time lord Bunmiiraammalak go persuade
ø=KS give ø=KS leave-the-monkshood ø=KS SP not agree leave-mh. //

ø=KS Q ø=KS out-of merit already // ø=KS&Bunmiiraammalak
negative-imperative go fight 'them' at all //

II.A3 daŋnípentôn //

'like-this-for-openners' //

II.B lé ø pen phûu rúu kîríjaa ʔàtchaasǎj câwkrunthonbùrii sŭŋ
lûuklǎan khǎw lâw kan jùu wâa muâa ø cà rápsàŋ kàp khraj
khraj ø kò jôm riâak phráʔon ʔeeŋ wâa phôo //

and ø=W to-be person know habits character LT RC descendants 3rd-
person relate reciprocal be-located Q time ø=LT IR speak with who who-
(anybody) ø=LT SP usually call body reflexive Q father //

II.B1 dagníi //
like this //

Translation:

I (King Chulalongkorn) conclude thus because the writer respects the Lord of Thonburi. She calls him "The First Reign"; she uses respectful idiom when referring to him, which is the same as the descendants of the Lord of Thonburi speak, as I have heard a great deal. When she refers to his mental aberration, she speaks with sympathetic understanding, saying that it was an accident that things happened like that because it was a time of bad Karma for him or a time when his merit had been used up. When she speaks of his fierceness, what she says almost appears to be bragging, saying that he was bold or decisive. This is exactly the same as how the descendants of Khun Luang sua refer to Khun Luang sua, praising him for having foresight in that he Bunmiiraammalak went to persuade him to leave the monastery [and attempt to regain the throne], he refused, saying that his merit was used up, that they should not fight "them". It is like this for one thing. And she was a person who knew the habits of the Lord of Thonburi, of which was aware that his merit was used up. When "they" invited him [to leave the throne] to become a monk, he was happy to do so. The time when Lord his descendants relate that when he would speak with anybody, he usually called himself "father". It is like this.

Grima (1986:160-163).

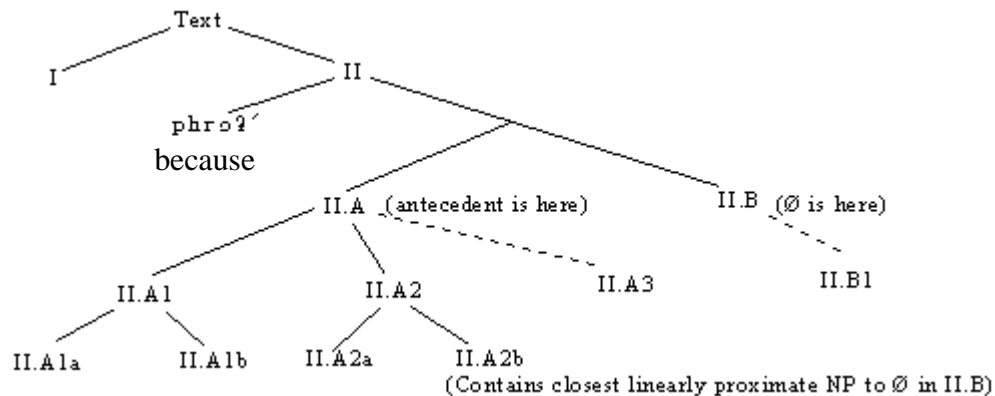


Figure 1: The structure of discourse (1)

According to Grima (1986), the antecedent of the zero pronoun in (II.B) is found in (II.A) rather than in the immediately preceding sentence in II.A2b. The antecedent and the zero pronoun are separated by many sentences and by over one hundred words (Grima 1986:159). The coreference of this zero pronoun could be explained in terms of hierarchical structure of the discourse. Since both (II.B) and (II.A) are assertions to support (I), they are on the same level of the structure. The antecedent of a zero pronoun in (II.B) should, then, be resolved from (II.A) rather than from a lower level structure, either (II.A2a) or (II.A2b). (II.A2b) is the part that contains the closest NP. In (II.A2b), there are four semantically compatible referents (+human) which are ‘Khun Luang sua’, ‘Lord Bunmiiraammalak’, ‘Lord of Thonburi’, and ‘the writer’. If the structure of discourse is not considered, a world knowledge inference (historical knowledge) is necessary for resolving the zero pronoun in this example. The referents ‘Khun Luang sua’ and ‘Lord Bunmiiraammalak’ are rejected because they did not live in the same time the text was written. The referent ‘Lord of Thonburi’ is rejected because of the binding principle.³ The correct referent (‘the writer’) is not referred to in the closest sentence in (II.A2b), but it is referred to in (II.A) and in the first sentence of (II.A2b). Thus, any focusing algorithm that suggests

³ Principle-B of the binding theory (Chomsky 1982, 1986) prohibits a pronoun to corefer to another noun phrase in the same clause.

an entity in the previous sentence as a preferred referent will fail to do its task of reducing the cost of inference by suggesting a correct antecedent.

Although the discourse analyzed by Grima is a text written in the early part of the twentieth century, the long distance anaphora discussed above may be found in contemporary Thai texts. Example (2), taken from a newspaper, suggests that the hierarchical structure of clauses in a discourse can be an important factor for zero pronoun resolution in Thai.

(2) Text from a newspaper

- #1 phútthásàmaakhom 30 ʔonkoon mii mátì hěnp hóŋ kan
Buddhist-Society 30 organization have decision agree RECP
 ‘Thirty of the Bhuddhist Societies agreed’
- #2 thîi ø cà kràapbaŋkhomthuunthàwăaj năŋsǔu pèetphànək tòo
 sǝmdètphrásăŋkhârâat thîi mii cajkhwaam doojjôo 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 Patriarch, which has the
 content in brief as follow:
- #3 ø khǝo hâj ø thonphícaarânaa damnœenkaan sàsăaŋ koorâni
 wátthammákaj
 ø=Buddhist-society ask let ø= the-Supreme-Patriarch consider do clear
 case Dhammakaya
 ‘They would ask the Supreme Patriarch to clear the Dhammakaya case’
- #4 phró mátì máhăarésàmaakhom thîi ʔòok maa jaŋ mâj khrôpkhlum
 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 pên rûuppàtham
 ø=the-decision lack measurement COMP be concrete
 ‘It lacks a measurement that is concrete.’
- #6 phuâa hâj kœet pràsithíphâap taam mátì
 so-that let occur efficiency follow decision
 ‘To make the decision work,’

- #7 thaang phútthásàmaakhom cung dâj sàněe nɛɛwthaang daŋtòopajníi khuu
for Buddhist-Society thus ASP suggest solution as-follow that-is
'the Buddhist Societies then propose the following methods.'
- #8 phráwínítchǎj khǒɔŋ sǒmdètphrásǎŋkhàràat
decision of the-Supreme-Patriach
'As for the decision of the Supreme Patriach,'
- #9 sũŋ ø mii kaan rábùthũŋ kaan bìtbuaan phúttháatham khamsoon
CONJ ø=the-decision have NOM indicate NOM distort Buddha
instruction
'which has mentioned about the distort of Buddha's instructions.'
- #10 ø thamhâj sǒŋ tɛɛkjɛɛk
ø=the-distort cause monk disruption
'which caused the disruption among monks'
- #11 lé kaan môop sǒmbàt thǎŋmòt thîi kèet khũn najkhànàthîi ø
pɛn phrá hâj kɛɛ 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'
- #12 máhǎarésàmaakhom khuaancà damnəenkaan taam phráwínítchǎj
khǒɔŋ sǒmdètphrásǎŋkhàràat
the-Sangha-Council should do follow decision of the-Supreme-Patriach
'The-Sangha-Council should follow the decision of the Supreme
Patriach'
- #13 ø khǒɔ 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 committees between
monks and laymen'
- #14 phuâa ø tittam duuleɛ kítçàkam wátthammákaaŋ
so-that ø=the committee follow look activity Wat-Dhammakaya
'So the committee can follow Wat Dhammakaya's activities.'
(Thairath, April 6, 1999, p.14)

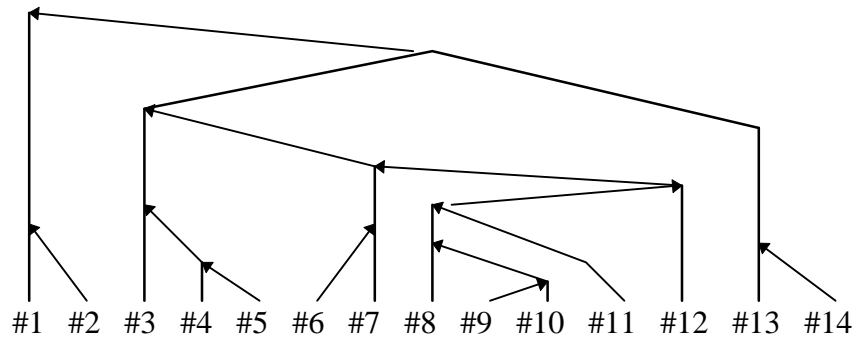


Figure 2: The structure of discourse in example (2)

The structure of discourse in example (2) could be analyzed as a hierarchical structure of clauses by using Rhetorical Structure Theory (Mann and Thompson 1987) as shown in Figure 2. 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. (See Chapter 3 for further information on Rhetorical Structure Theory).

In example (2), 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. Sidner's focusing algorithms would select the focused entity in (#12), i.e. 'The Sangha Council', as the preferred referent of (Z2). While Sidner notes the possibility of focus-popping based on structural ones, she did not include structural effects on focusing in her study (Sidner 1983:300-302)⁴.

The relationship between discourse structure and anaphora resolution has been studied in Grosz (1977, 1981) and Grosz and Sidner (1986). Grosz (1977, 1981) used the structure of discourse (which she found to correspond to the structure of tasks being discussed) to identify an active focus space which contains highly focused discourse entities at a certain time. One of these entities is likely to be a preferred referent for an anaphor. While the term 'focus' used in Sidner's focusing is called a 'local focus', it is called a 'global focus' in Grosz's focusing. Grosz's global focus is organized into a

⁴ Focus popping is the process in which the focus shifts back to a previous focus. In example below, the focus in (4) shifts back to the focus in (1).

1. I want to schedule a meeting_j with Harry, Willie and Edwina.
2. We can use my office_i.
3. It's_i kind of small
4. but the meeting_j won't last long anyway. (Sidner 1983:299)

However, her algorithms could eventually resolve this zero pronoun because the antecedent will be found in the focus stack. Nevertheless the performance of her algorithms in this case is not better than that of a simple matching algorithm because the focus stack is the last place at which her focusing algorithms look for an antecedent.

network structure of focus spaces with respect to the task structure. Grosz's research on global focusing and discourse structure was further examined in Grosz and Sidner (1986) in which a model of discourse was proposed composing of three interrelated components: linguistic structure, intentional structure, and attentional structure. The intentional structure is a mirror of the linguistic structure, which is a hierarchical structure of discourse segments. The attention structure is a stack of focus spaces which are pushed or popped with respect to the intentional structure. An antecedent of an anaphor is likely to be found in the most current focus space. These works suggest that the structure of discourse contributes to anaphora resolution. Moreover, it has been accepted that the structure of discourse could affect local focusing and reference resolution (Sidner 1983:300-302, Suri 1993:252, Carter 1987:118, Allen 1995). If a local focusing algorithm could recognize the shifting of global focus, the accuracy should be increased.

Based on this view and examples (1-2) discussed above, we suspect that the hierarchical structure of discourse could be an important factor for zero pronoun resolution in Thai. We investigate this issue by proposing a focusing algorithm that can work with hierarchical structure of discourse and by testing the algorithm on naturally-occurring Thai texts. The results will reveal whether the hierarchical structure of discourse is an important factor for zero pronoun resolution in Thai.

1.3 Zero pronouns and pronoun system in Thai

Before identifying the scope of zero pronouns in this study, we will first give an overview of the pronoun system and zero pronouns in Thai. The Thai language in this study refers to standard Thai (the central dialect), which is the official language of Thailand. An overview of the Thai pronoun system will be discussed in section 1.3.1. The usage of zero pronouns will be briefly discussed in section 1.3.2. We will be in a position to define the scope of zero pronouns in this study in section 1.4

1.3.1 Pronoun system in Thai

In Thai, the pronoun system is not restricted to only personal pronouns. Other forms such as personal names, status terms, and kinship terms can be used pronominally. The use of these forms as pronouns is quite common in Thai and in other languages like Burmese and Vietnamese (Cooke 1968:2). But in this study only personal pronouns and zero pronouns will be discussed. The following description is based primarily on Cooke (1968) and Hoonchamlong (1991:13-20).

1.3.1.1 Personal pronouns

In Thai, there are three distinct sets of pronouns: one for royalty, one for the Buddhist clergy, and one for ordinary people. Only the pronoun system used for ordinary people will be presented here. There are many first, second, and third pronouns. Their usage is governed by the social identities of the speakers and hearers.

Some are polysemous. For example, /raw/ can be either a first or second person pronoun. The following are some examples of personal pronouns commonly used in Thai.

First person pronouns

- /dìchǎn/ : female speaking to superiors, or equals.
- /chǎn/ : male speaking to female intimate, or to inferior.
- /chǎn/ : female speaking to intimate equal, or to inferior.
- /khâa/ : speaking to inferior.
- /khâaphácâw/ : formal, used in public address.
- /kràphǒm/ : male, very formal.
- /kuu/ : informal, considered crude if not used to intimate friends, mostly male.
- /nǔu/ : female, informal, used by a girl or a younger woman.
- /phǒm / : male, polite.
- /raw/ : general plural form of most singular first person forms.
- /raw/ : informal, intimate, to friends.

Second person pronouns

- /khun/ : polite, to equals or superiors.
- /kɛɛ/ : speaking to inferior, impolite if used to nonintimates.

- /maŋ/ : mostly male speaking to intimates, impolite if used to nonintimates.
- /nũu/ : informal, female, addressee is much younger.
- /raw/ : addressee is inferior in ages and social status.
- /thâan/ : formal, to superiors.
- /thəə/ : speaking to inferiors, especially female teachers to pupil.
- /thəə/ : speaking to intimate equal, by or to female.

Third person pronouns

- /man/ : third person for animals, things, or abstract ideas.
- /man/ : informally, to intimates or inferiors.
- /kɛɛ/ : informal, male or female.
- /kháw/, /khǎw/ : singular or plural, intimates or nonintimates.
- /thâan/ : formal, addressee is superior in social status.

1.3.1.2 Zero pronouns

Zero pronouns are prevalently used in actual discourse. The study of Maneeroje (1985) indicates that zeroes are used eight times more than pronouns in her sample of written texts.⁵ Usually, their referents could be inferred from the context. Cooke

⁵ Maneeroje (1985) studies the use of four NP forms (zeros, repeated NPs, demonstrative NPs, and pronouns) as cohesive devices in ten written Thai texts.

(1968:10) notes that zero pronouns may be adopted as a strategy of politeness. Zero pronouns are used when a speaker is not certain if he or she can make an appropriate choice of pronoun (Hoonchamlong 1991:21, Cooke 1968:63). In a situation where an appropriate choice of pronoun is problematic, zero pronouns seem to be the best alternative, as described in Cooke's example below (1968:63).

- (3) A young man who has been exposed to democratic ideals addresses a peddler. The pair /khâa/ (I) and /ʔeŋ/ (You) is out of date and too lordly; /chăn/ (I) and /kεε/ (You) is likewise undemocratic and is especially inappropriate for use by a young man to an older person; /phôm/ (I) and /khun/ (You) is absurdly respectful. The only alternative is to avoid pronouns altogether.

In terms of distribution, zero pronouns are not different from overt pronouns. Zero pronouns can occur in any position that an overt pronoun can except in the position after a preposition. The following examples indicate the distribution of zero pronouns and overt pronouns in different grammatical roles.

- (4) a. phôm/∅ kliàat tuaaʔeŋ luăakəən
 I_i/∅_i hate oneself_i so-much
 '(I) hate myself very much'
 b. dεεŋ bòɔk wâa khăw/∅ maa léεw
 Daeng_i say that he_{i,j}/∅_{i,j} come ASP
 'Daeng_i said that (he_{i,j}) came'
 c. khun chôɔp man/∅ mǎj
 you_i like it_j/∅_j Q
 'Do you like it?'

According to her study, zeroes are the most frequently used form (49.88%) while pronouns are the least frequently used (5.90%).

- d. dɛɛŋ bɔ̀ɔk wâa dɛɛŋ chɔ̀ɔp man/∅
 Daeng_i say that Daeng_i like it_j/∅_j
 ‘Daeng said that he likes it.’
- e. phrûŋníi khun maa kɛ̀p ɲəən càak phǎm/*∅
 tomorrow you_i come collect money from me_j/*∅
 ‘Tomorrow, you come to collect money from me.’

1.3.2 Usage of zero pronouns

Zero pronouns in Thai can be used as deictic pronouns, anaphoric pronouns, discourse deictic pronouns, or generic pronouns. Below are examples of the usage of zero pronouns.

1.3.2.1 Deictic usage

A zero pronoun is used deictically when its referent is ‘situationally evoked’ (Prince 1981). Not only are zero pronouns used to refer to the speaker or the addressee in a conversation, they can also be used to refer to a third person if the referent is present or can be inferred from the situation. For example, in a situation that two persons are talking behind their boss’s back. One person could say a sentence like (5) if their boss walks into the room at that time. The zero pronoun in this sentence refers to their boss.

- (5) ∅ maa lɛ̀ɛw
 ∅ come ASP
 ‘(Here, he) came.’

1.3.2.2 Anaphoric usage

Zero pronouns can be used as an anaphor like other pronouns. In example (6), the antecedent of the zero pronoun is found in the main clause. However, it should be noted that antecedent of a zero pronoun may be in a distant sentence, as already seen in examples (1-2).

- (6) dɛɛŋ bɔ̀ɔk wâa ø chɔ̀ɔp năŋ ruâaŋ níi
 Daeng_i say that ø_i like movie CL this
 'Daeng said that (he) likes this movie.'

1.3.2.3 Discourse deictic usage

Zero pronouns can be used as discourse deixis (see Webber 1988) if they refer to a chunk of linguistic expression instead of the referent of that expression, or the interpretation of one or more clauses. In example (7), ø₂ is a case of discourse deixis because it refers to the linguistic expression 'Krailaj'.

- (7) A: phǒm chûu krajlâat
 I name Krailaj₂
 'My name is Krailaj'
 B: ø chuâaj sàkòt ø hâj phǒm thi
 ø₁ help spell ø₂ for me PRT
 'Spell it for me, please'

1.3.2.4 Generic usage

Some zero pronouns are used because their referents are not significant in speakers' views. In fact, their referents often cannot be uniquely identified. Rather, they

only signal the type of things described, or ‘type identifiable’ (see Gundel et al. 1993).

Below is an example of a zero pronoun whose referent could not be identified from the discourse context.

(8)

ø₁ pràmaan wâa râtthàbaan fàrànsèet cà sǎamâat ruâapruaam
 ɲəən càak kaan pɛərûp râtɰísǎakít naj pii níi dâj mâj tà̃m
 kwàa 11 phan láan doonlâa ləejthiidiaaw
 ø₁ estimate that government French will can gather money from
 NOM transform state-enterprise in year this ASP not less than 11
 thousand million dollar ASP
 ‘(It) is estimated that the government would earn at least 11 billion
 dollars from transforming state enterprises this year.’

In example (8), the referent of ø₁ cannot be identified from the discourse. We only know that its referent type should be a human because of the verb /pràmaan/.

1.4 The scope of this study

Zero pronouns that are used anaphorically (1.3.2.2) are the main concern of this study. Antecedents of these zero pronouns are often but not always found within the same sentence or in the immediately preceding sentence. We hypothesize that the coreference of zero pronouns and their antecedents can be better explained in terms of discourse structure, which is analyzed as a hierarchical structure of organized clauses. An analysis of a corpus of twenty Thai texts (15,949 words) will be used for this purpose. The structure of these discourses will be analyzed by native speakers of Thai. Then, we will find out whether antecedents of zero pronouns could be identified with

respect to the hierarchical structure of discourse. We assume that zero pronoun resolution can be done at two levels: the sentence level and the discourse level. Zero pronouns that can be resolved at the sentence level are not included in this study and will not be marked in the corpus. Only zero pronouns which cannot be resolved at the sentence level and, thus, have to be handled at the discourse level will be marked and analyzed.

1.5 Outline

Chapter 2 focuses on the treatment of Thai zero pronouns in the Government and Binding Theory. Since zero pronouns are analyzed as empty categories in the Government and Binding Theory, the inventory of empty categories and principles related to coindexation of empty categories will be discussed first. Next, an overview of Thai syntax and empty categories in different constructions in Thai will be discussed. The chapter indicates zero pronouns which can be resolved by some principles in the Government and Binding Theory. These zero pronouns will not be marked in the corpus and will be excluded from this study.

Chapter 3 is a proposal of an extended centering model. The relationship between focusing and reference resolution will be discussed first. Next, previous research on focusing, Sidner's focusing algorithms, Centering Theory, and Suri's extended focusing algorithms, will be reviewed with certain aspects found deficient for

zero pronoun resolution in Thai. Fox's research on rhetorical structure will then be reviewed to indicate the relationship between discourse structure and anaphora. The chapter ends with a proposed model of extended centering, which extends the model of centering theory incorporating some of Fox's conclusion.

Chapter 4 discusses the design of corpus for comparing the performance of two focusing algorithms: an existing centering and an extended centering. Native speakers of Thai are asked to analyze the corpus specifically to identify referents of zero pronouns and hierarchical structures of discourses. When applying the two focusing algorithms to this corpus, the result will reveal whether the hierarchical structure of discourse contributes to zero pronoun resolution in Thai.

Chapter 5 concerns centering in Thai. The model of centering and how it could be used to resolve zero pronouns in Thai will be discussed. Since the original model is not sufficient for resolving zero pronouns whose antecedents are in a distant utterance, an extended model of centering proposed in chapter 3 will be adopted as the basis for zero pronoun resolution in Thai.

Chapter 6 reports the performance of the two focusing algorithms. The results indicate that hierarchical structure does not have a strong effect on the centering model's performance. This is because most of zero pronouns have their antecedents in the immediately preceding utterance. However, coreference of some zero pronouns can

be explained better in terms of hierarchical structure of clauses. And the hierarchical structure seems to be at the sentence level rather than at the discourse level.