

Chapter 4

Corpus Design and Analysis

In this chapter we discuss the design of a corpus for comparing the performance of two focusing algorithms: the existing Centering Theory and Centering Theory extended to take into account the hierarchical structure of discourse. Since the comparison will be made with respect to resolving the referents of zero pronouns in Thai, the corpus must be designed in such a way as to yield sufficient tokens of zero pronouns for analysis. To provide a basis for the evaluation of the focusing algorithms, it is necessary first to identify both the referents of the zero pronouns in the corpus, and to identify the hierarchical structure of discourse. Although this analysis is often performed by the researcher, we decided to increase the reliability of the judgments by asking a group of native speakers to identify both the referents and the hierarchical structure of discourse. In this chapter, therefore, we discuss the design of the corpus, the methodology, and the results of the referent and discourse-structure identification task.

4.1 Corpus design

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.³⁵ In this study, twenty expository prose texts were selected for the corpus. They are categorized as follows:

(96)

- Three news reports from newspapers: News1, News2, News3
- Four news reports from the Thai News Agency: Tnews1, Tnews2, Tnews3, Tnews4
- Two articles from a newspaper: Article1, Article2
- Two editorial sections from a newspaper: Editor1, Editor2
- Four computer articles from computer magazines: Comp1, Comp2, Comp3, Comp4
- Two sections from a health care book: Health1, Health2
- Two sections from a management book: Mang1, Mang2
- One section from a short story: Story1

Altogether 719 zero pronouns were identified in these twenty texts (15,949 total number of words). The corpus was 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. In the next section, we will explain the steps taken in preparation of the corpus.

4.2 Corpus preparation

The corpus was first pre-analyzed, with zero pronouns and utterance boundaries marked. This pre-analysis is necessary because Thai orthography does not consistently mark boundaries of a word, a clause or a sentence. Thus, if subjects are asked to

³⁵ 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%)

identify the number of clauses or sentences in Thai, it is likely that they would come up with different numbers of sentences or clauses. Furthermore, if subjects are asked to identify the presence of zero pronouns, it is likely that they may not agree on their presence in all cases, or they may miss the presence of zero pronouns in certain positions. Since we want to study the relationship between zero pronoun resolution and discourse structure, we need to limit the number of possible variations. By identifying in advance the presence of zero pronouns and the smallest unit of discourse structure, subjects would concentrate only on identifying referents and identifying discourse structure. Also, they would work on the same data and their judgment could be compared.

4.2.1 Marking of zero pronouns

A zero pronoun was added to the text markup where a noun phrase was missing from a position required by subcategorization or the extended projection principle in Government and Binding Theory (Chomsky 1982:10). However, purely syntactic anaphora were excluded and left unmarked in the corpus (see Chapter 2). We marked only zero pronouns that cannot be resolved by principles in Government and Binding Theory. These zero pronouns are to be resolved by focusing algorithms.

We expect that some of the zeroes marked in the corpus are used as diectic, discourse diectic, or generic pronouns (see section 1.3.2.4). The results of analysis will

reveal these zero pronouns. Since these zeroes are not in the scope of this study, they are excluded from the comparison of focusing algorithms.

4.2.2 Marking of discourse units

We follow Rhetorical Structure Theory (RST) in identifying the smallest discourse unit as a clause. However, according to RST (Mann and Thompson 1987:6, Fox 1987:78), clauses that do not have any discourse function as defined by RST will not be identified as a separate unit. For example, restrictive relative clauses and subject/object complement clauses are not analyzed as separate units. This analysis is based on the difference between ‘embedding’ and ‘clause combining’ discussed in Matthiessen and Thompson (1988:279-282). Embedded clauses are clauses that function as a constituent or a complement of another clause. For example, the clause ‘before the Magistrate had been invited’ in (97) is an embedded clause because it is a part of the noun phrase ‘the happy days before the Magistrate had been invited’. With this line of analysis, restrictive relative clauses and subject/object complement clauses are also embedded clauses.

(97)

Those were the days when every single poem had bristled with good qualities like a hedgehog and had glutted itself with praise like a jackal, the happy days before the Magistrate had been invited.

(Matthiessen and Thompson, 1988:279)

‘Clause combining’ is a case where clauses are combined together in the same manner as rhetorical units are connected in a discourse. Matthiessen and Thompson (1988) show that clause combining is not necessarily a combination of two simple clauses but can be a combination of a clause and a clause combination. Clause combining includes not only coordination but also certain kinds of subordination. In example (98), the clause ‘while Ed was coming downstairs’ is combined with a unit which is a combination of three coordinate clauses (‘Mary slipped out the front door, went around the clause, and came in the back door’). In example (99), (c) is the condition of (d), and (a) and (b) are combined as the conditions of a unit composed of (c) and (d).

(98)

While Ed was coming downstairs, Mary slipped out the front door, went around the clause, and came in the back door.

(Matthiessen and Thompson, 1988:281)

(99)

Our teacher says that

a. if your neighbour has a new baby and

b. you don’t know whether it’s a he or a she,

c. if you call it ‘it’

d. well then the neighbour will be very offended.

(Matthiessen and Thompson, 1988:303)

Following this line of analysis, restrictive relative clauses and subject/object nominalized clauses in Thai will be analyzed as a part of an utterance rather than as a separate utterance, while a subordinate clause or a coordinate clause is counted as a separate utterance.

4.2.3 Questionnaires

Two sets of questionnaire were given to twelve subjects (see Appendix B). The first set deals with the identification of referents of zero pronouns. The texts used for the first questionnaire were not marked with boundaries of utterances. Only zero pronouns were marked in the texts, using a running number from $Z1$ to Zn . Subjects were asked to fill in the referent of each zero pronoun on the questionnaire. The second set of questionnaires dealt with the identification of discourse structure. Each questionnaire consists of a discourse divided into lines running from 1 to m , where each line represents the smallest discourse unit, which we will call an utterance here. Zero pronouns were not marked, but paragraph units were. Subjects were asked to group coherent units together to form a hierarchical structure.

4.2.4 The subjects

While the analysis of zero pronoun referents could be performed by any native speaker of Thai, the analysis of discourse structure, requires basic knowledge of hierarchical structure analysis in linguistics. Therefore, we chose twelve Thai graduate students in their second year of linguistics study as our subjects.

4.3 Corpus analysis

Subjects were asked to identify both the zero pronoun referents and the discourse structure of the text. 100 questionnaires were administered for the referents

of zero pronouns and 60 questionnaires for the identification of discourse structure. An example of each type of analysis was provided to familiarize the subjects with the tasks. For both sets of questionnaire, the original texts precede the questionnaires. Subjects were asked to read the original text before doing the analysis. Having read the original texts, they worked on identifying referents of zero pronouns or on identifying discourse structure depending on the questionnaire they received. (Questionnaires were distributed randomly to the subjects.)

4.3.1 Identifying referents

Since we expected that subjects might not agree on the referent of every zero pronoun marked in the corpus, we asked five subjects to identify referents of zero pronouns on the same text. Ten texts were analyzed by five subjects each while the other ten texts were analyzed by four subjects and by the author. In other words, 90 questionnaires were analyzed by the subjects while 10 questionnaires were analyzed by the author. The referents of zero pronouns which were indicated by the majority of the responses are chosen as the correct response. Thus, a referent that was agreed to by at least three out of five subjects was regarded as the ‘correct’ referent of the zero pronoun. This ‘correct’ referent will be used to compare with the suggested referent from the focusing algorithms. Where the majority of subjects failed to agree on a

referent for a zero pronoun, it was not used in the comparison with the focusing algorithms.

With regard to identifying referents from the discourse context, subjects were instructed to use the word ‘unidentified’ if they thought that there was no specific referent of the zero pronoun in the discourse. We hypothesized that there might be a number of such cases because it is possible for a zero pronoun in Thai to be used with generic reference, as discussed in section 1.3.2.4. In fact, 82 out of 719 cases (11.40%) were analyzed as ‘unidentified’ referents. These zero pronouns were ignored in the testing of focusing algorithms. Examples of zeroes analyzed as having an ‘unidentified’ referent are shown in (100) and (101) below.

(100) Text from Article2

#54 [Z31] pràmaan wâa râtthàbaan fàrànsèet cà sǎamâat ruâapruaam
 nǝen càak kaan pǝerûup râtwsǎakít naj pii níi dâj mâj tám
 kwâa 11 phan láan doonlâa lǝejthiidiaaw
 [Z31=unidentified] 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.’

(101) Text from Comp3

#16 kaan sǝon hâj khoomphiwtǝe khâwcaj phaasǎathammáchâat pǝn
 khwaamfǎn khǝonj mánút maa pǝn weelaa naan lǎaj pii léew
 NOM teach to computer understand natural-language be dream of
 human ASP be time long many year ASP
 ‘Making a computer to understand human languages has been a human
 dream for a long time’
 #17 daŋ [Z10] cà hǝn dâj taam níjaaj rǝu phâapphàjon
 wítthájaasàat tàaŋtàaŋ

as [Z10=unidentified] will see able in fiction or movie scientific
 any
 ‘as (we) can see in many science fictions or in sci-fi movies’

4.3.2 Identifying discourse structure

While we can use majority agreement for identifying referents of zeroes, we cannot use the same method for the analysis of discourse structure. Therefore, we used each analysis of discourse structure as an individual structure for testing focusing algorithms. Twenty texts in the corpus were analyzed by the author and by two other subjects. In other words, 40 questionnaires were responded by the subjects while 20 questionnaires were answered by the author. Therefore, sixty discourse structures were used for testing focusing algorithms.

Subjects were asked to draw a line to combine units that were coherent. They could work either in the bottom-up or top-down fashion. We asked subjects to identify the structure within a paragraph rather than in the whole text because we anticipated that it is unlikely for a zero pronoun to have a referent across a paragraph. Thus, each paragraph of the text would have only one hierarchical structure.

4.4 Results of the analysis

This section reports the results of our corpus analysis. Section 4.4.1 deals with the result of referent analysis. Section 4.4.2 deals with the results of discourse structure analysis.

4.4.1 Results of referent analysis

Of the 719 zero pronouns marked in the corpus, 679 (94.44%) were agreed upon by the majority of the subjects. The majority of the subjects failed to agree on the referents of 40 pronouns (5.56%). Of the 679 zeroes agreed on by the majority of subjects, 597 zero pronouns (83.03%) were analyzed as having identifiable referents while 82 zero pronouns (11.41%) were marked as ‘unidentified’. The set of 597 zeroes is the basis of our study.

	NoOfZero	Agreement			Conflict
		Identified	Unidentified	Total	
Total	719	597	82	679	40
Percent	100%	83.03%	11.41%	94.44%	5.56%

Table 2: Results of referent analysis

Of the 40 zeroes where the majority of the subjects failed to agree on their referents, 8 zeroes (20%) are identified as one entity by two subjects and as another entity by the other two subjects. For the other 32 zeroes (80%), they are analyzed as ‘unidentified’ referents by at least one subject. We suspect that the lack of majority agreement may be related to the status of ‘unidentified’ referents. Therefore, we shall briefly examine the relationship between referent identification and degree of agreement. Table 3 shows the status of referent identification and the degree of agreement by the majority of subjects. We categorize referent identification by the majority of subjects into two types: identified referents and unidentified referents. The

second column (60%) indicates the number of referents agreed by three out of five subjects; the third column (80%) indicates the number of referents agreed by four out of five subjects; and the fifth column (100%) is the number of referents agreed by all subjects. The distribution in Table 3 confirms that the status of referent identification is related to degree of agreement ($\chi^2 = 46.57$, $p < 0.001$). Referents with less agreement among respondents are likely to be analyzed as ‘unidentified’ referents than those with higher agreement.

	60%	80%	100%	Total
Unidentified	39	23	20	82
Identified	106	138	353	597
Total	145	161	373	679

Table 3: Degree agreement and identification of referents

Unidentified zeroes were excluded from the testing of focusing algorithms. In addition, zeroes used as deictic pronouns and zeroes in embedded clauses were excluded from the testing because centering algorithms are not designed to resolve zeroes whose referents are in the same utterance.

4.4.2 Results of structural analysis

Even if we cannot derive a single discourse structure for each text with majority agreement among respondents, we can discuss the similarity of their structure analyses. Passonneau and Litman (1993) discussed the results of discourse structure analysis in their study in terms of segmentation agreement. They used ‘percent agreement’ as

defined in Gale et al. (1992) to measure the agreement of analysis. Percent agreement is defined as ‘the ratio of observed agreement with the majority opinion to possible agreements with the majority’ (Passonneau and Litman 1993:149).

Although the analysis of discourse structure in this study differs from that of Passonneau and Litman’s study, we may look at the agreement of analysis in a similar way if we view the hierarchical structure of discourse in terms of segmentation. While Passonneau and Litman asked subjects to identify segment boundaries and not the hierarchical structure of discourse, we asked subjects to identify the hierarchical structure of discourse. For example, in the structural analysis of five utterances below, there will be four utterance boundaries between U1 and U2, U2 and U3, U3 and U4, and U4 and U5. In this example, only the second boundary is a segment boundary.

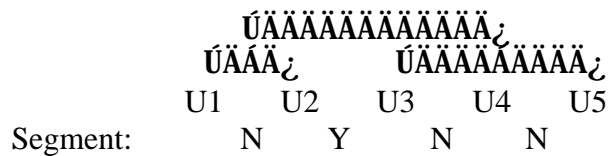


Figure 14: Segment boundary

4.4.2.1 Agreement on segmentation

Agreement on segmentation could be measured by using ‘percent agreement with the majority opinion’ as used in Passonneau and Litman (1993). However, we do not use this measurement to measure the agreement among our subjects because it tends to give a high value of agreement. Since percent agreement with the majority is

defined as the ratio of observed agreement with the majority to possible agreements with the majority, it will give us more than 50% agreement in any situation. For example, if eleven utterances were analyzed by seven subjects, the total possible agreements with the majority would be 10×7 or 70. The observed agreement with the majority would be 70 minus the number of disagreements with the majority. In the worst case when three subjects do not agree with the other four subjects in every utterance boundary, this method will give the number of agreement as $(10 \times 7 - 10 \times 3) \times 100 / (10 \times 7) = 57.14\%$.

In this study, we use percent agreement and the kappa coefficient (Fleiss 1971) to indicate the agreement regarding segmentation among subjects. Percent agreement is a basis measurement for agreement of analysis. It is calculated from the absolute number of agreement. In this analysis, where there are three subjects and two choices of analysis, the total number of possible agreement is $3 \times G$ (G is the number of utterance boundaries in the text) and the number of agreement on each text ranges from $2 \times G$ to $3 \times G$. Table 4 shows the percent agreement for each text.

The mean of percent agreement in this study is 84.33%. Since there are two choices of analysis on each gap (whether there is a segment boundary or not) and there are three subjects, at least two subjects must agree on the analysis of each boundary. Therefore, percent agreement will range from 66.67% to 100%. As we can see, degree of agreement based on this measurement is affected by the number of choices and the

number of subjects. Therefore, we would like to use another measurement which is not affected by the number of choices and the number of subjects. That measurement is ‘kappa coefficient’.

Text	% agreement	Text	% agreement
Article1	86.23%	Mang1	88.21%
Article2	77.44%	Mang2	82.69%
Comp1	85.71%	News1	77.24%
Comp2	78.21%	News2	81.82%
Comp3	81.64%	News3	83.12%
Comp4	91.23%	Story1	88.89%
Editor1	87.42%	Tnews1	100.00%
Editor2	74.01%	Tnews2	100.00%
Health1	88.16%	Tnews3	72.22%
Health2	80.10%	Tnews4	82.22%
Mean		84.33%	

Table 4: Percent agreement on segmentation

The kappa coefficient is an index that measures agreement of analysis among subjects. Kappa coefficient is designed to account for agreement that might occur by chance. For example, in an analysis of three subjects and two categories, there is a 25% chance $((1/2 \times 1/2 \times 1/2) + (1/2 \times 1/2 \times 1/2))$ ³⁶ that all subjects would choose the same category at a time. In an experiment where category A is chosen m times and category B is chosen n times, the possibility that w subjects may agree by chance is $(m/(m+n))^w$

³⁶ In this case, the chance that three subjects would choose ‘Y’ for each boundary is $1/2 \times 1/2 \times 1/2$ and the chance that three subjects would choose ‘N’ for each boundary is $1/2 \times 1/2 \times 1/2$. Thus, the chance that three subjects would choose the same category is $1/8 + 1/8 = 1/4 = 25\%$.

+ $(n/(m+n))^w$).³⁷ This indicates that the chance expected agreement in different research settings will yield different results. Therefore, as Carletta (1996) argues, it is not useful to compare or interpret results of raw agreement in different research settings. But if the chance expected agreement is included in the calculation of the total agreement results, comparison of these results from different research settings will be possible. Thus, use of the kappa coefficient will provide us with this necessary information.

In the present study, Fleiss' kappa is used to measure the agreement of segmentation because it is designed for use with nominal data when there are more than two subjects or judges. The formula of Fleiss's kappa is presented in (102).

(102)

$$k = \frac{Po - Pe}{1 - Pe}$$

with: Po = proportion of agreeing pairs of judgments

Pe = proportion of agreeing pairs on the basis of chance

These proportions are calculated as follows:

$$Po = \frac{\sum n_{ij}^2 - Nk}{Nk(k-1)}$$

$$Pe = \sum_{j=1}^v P_j^2$$

$$P_j = \frac{\sum_{i=1}^N n_{ij}}{Nk}$$

³⁷ n^w equals to n times itself w times. For example, 2^3 is equal to $2 \times 2 \times 2$.

The symbols used are:

N = number of judged objects

k = number of judgments per object or number of judges

v = number of categories

n_{ij} = number of judges who assign object i to category j

(Rietveld and van Hout 1993:221-222)

Kappa shows the degree of agreement ranging from $-Pe/(1-Pe)$ to 1. The value '1' indicates perfect agreement, while '0' indicates that the agreement is no better than chance. Landis and Koch (1977:165) provide the following table as the guideline for interpreting the degree of agreement:

< 0.00	Poor
0.00 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 1.00	Almost Perfect

Table 5: Guideline for interpreting kappa coefficient

In this study, the number of subjects (k) equals 3³⁸ and the number of categories (v) equals 2. The number of judged objects (N) depends on the number of utterance boundaries in each text. Table 6 represents an example of the analysis of text Tnews4. The value of N , then, is equal to 15.

³⁸ Discourse structure of each text was identified by three subjects.

Boundary No	Subj1	Subj2	Subj3	Cat0	Cat1	Cat0 ²	Cat1 ²
1-2	0	0	0	3	0	9	0
2-3	1	1	0	1	2	1	4
4-5	1	1	1	0	3	0	9
5-6	1	1	0	1	2	1	4
6-7	0	1	1	1	2	1	4
7-8	1	1	1	0	3	0	9
8-9	0	0	0	3	0	9	0
10-11	1	1	1	0	3	0	9
11-12	0	1	0	2	1	4	1
12-13	1	1	1	0	3	0	9
13-14	1	1	1	0	3	0	9
14-15	0	0	0	3	0	9	0
16-17	1	1	1	0	3	0	9
17-18	1	1	1	0	3	0	9
18-19	0	0	0	3	0	9	0
Total				17	28	43	76

Table 6: Segmentation analysis of text Tnews4

Segmentation analyses of three subjects are presented in the second, the third, and the fourth columns. The value '1' is marked when there is a segment boundary at particular location while the value '0' is marked when there is no segment boundary at that point. The number of value '0' or '1' for each gap is shown in the fifth and the sixth columns. Following the formula above, we can calculate the degree of agreements among subjects as follows:

(103)

$$Pe = (17/45)^2 + (28/45)^2 = 0.5299$$

$$Po = ((43+76) - (15 \times 3)) / 15 \times 3 \times 2 = 0.8222$$

$$K = (0.8222 - 0.529) / (1 - 0.529) = 0.6219$$

The kappa of 0.629 indicates that subjects agreed on segmentation at the level of 62% above chance, showing that agreement among subjects in this text is substantial but not perfect.

Text	Kappa	Text	Kappa
Article1	0.6983	Mang1	0.7513
Article2	0.5337	Mang2	0.6280
Comp1	0.7001	News1	0.4994
Comp2	0.5226	News2	0.5844
Comp3	0.6193	News3	0.6226
Comp4	0.8043	Story1	0.7662
Editor1	0.7371	Tnews1	1.0000
Editor2	0.4508	Tnews2	1.0000
Health1	0.7581	Tnews3	0.4286
Health2	0.5698	Tnews4	0.6218
Mean		0.6648	

Table 7: Degree of agreements on segmentation

Table 7 shows the agreement in terms of kappa coefficient for each text. Agreement on segmentation in these texts ranged from 0.4286 (Tnews3) to 1.0 (Tnews1 and Tnews2) with a mean of 0.6648. Thus, there is a substantial degree of agreement in the analysis.

4.5 Conclusion

In this chapter, we have discussed the design and preparation of the corpus and the analysis of the results. Twenty Thai texts were selected as the corpus. These texts were analyzed by twelve native speakers of Thai. Subjects were asked to identify referents of zero pronouns and the structure of discourse in the corpus. The results of

their analyses provide a basis for the evaluation of the focusing algorithms. Referents identified by the majority of subjects will be regarded as the ‘correct’ referents of zero pronouns. These ‘correct’ referents will be compared next with the suggested referents from our focusing algorithms. The structure of discourse analyzed by each subject will be regarded as a possible discourse structure for the extended centering.

With regard to the identification of zero pronoun referents, our analysis showed 94% majority agreement, which yields a higher agreement than that of the discourse structure analysis. This was not unexpected, since identifying referents of zero pronouns from the previous context is an easier task than identifying structure.

With regard to the identification of discourse structure, it should be noted that both percent agreement and kappa coefficient only captures agreement on a certain aspect of discourse structure. They do not give us the absolute agreement of hierarchical structure analysis. Until an appropriate statistical method for analyzing the agreement with respect to hierarchical structure can be found, the segmentation will be used for describing the agreement of structure analysis.

Kappa coefficients indicates a substantial level of agreement regarding structure analysis. However, the agreement is not perfect. The mean of kappa for the twenty texts is 0.66. This is not unexpected from a subjective analysis like this one. Nonetheless, the level of agreement here seems low when compared with the level of segmentation agreement reported by Passonneau and Litman (1993), where agreement

ranged from 82%-92%, with a mean of 89.1%. However, their number does not take into account the chance expected agreement. If we were using the same measurement ('percent agreement with the majority opinion') for segmentation agreement, our results would range from 77.78% to 100%, with the mean of 89.4%. But we believe that by factoring in the chance expected agreement, our results will have a greater degree of accuracy and will also be more comparable with analyses of other researchers.