Basic Principles for Segmenting Thai EDUs

Nalinee Intasaw
Department of Linguistics,
Faculty of Arts, Chulalongkorn University,
Bangkok 10330, THAILAND
nalinee.int@gmail.com

Wirote Aroonmanakun
Department of Linguistics,
Faculty of Arts, Chulalongkorn University,
Bangkok 10330, THAILAND
awirote@chula.ac.th

Abstract

This paper proposes a guideline to determine Thai elementary discourse units (EDUs) based on rhetorical structure theory. Carson and Marcu's (2001) guideline for segmenting English EDUs is modified to propose a suitable guideline for segmenting EDUs in Thai. The proposed principles are used in tagging EDUs for constructing a corpus of discourse tree structures. It can also be used as the basis for implementing automatic Thai EDU segmentation. The problems of determining Thai EDUs both manually and automatically are also explored and discussed in this paper.

1 Introduction

Elementary discourse unit or EDU is a building block that can combine together to form a larger unit or structure in discourse. It is significant to applications that process discourse such as text summarization, machine translation, text generation, and discourse parsing. In some applications e.g. text summarization and machine translation, an EDU is suitable to be used as an input than a sentence or a paragraph since it is smaller and contains a single piece of information, In addition, in languages in which sentence boundaries are not clearly marked like Thai, determining an EDU would be more practical and more useful since an EDU serves as a building block for constructing the discourse structure. However, little study has been devoted to Thai elementary discourse unit. Previous research on Thai discourse structure (Charoensuk, 2005; Sinthupoun, 2009; Katui et al., 2012) did not clearly discussed how to determine an EDU in Thai. Determining an EDU is not an easy task. As a result, Carson and Marcu (2001) had developed a guideline for determining an EDU in English, which is used for tagging discourse tree structure. In this paper, our objective is to propose a guideline to determine Thai EDUs. The proposal is grounded on the framework of rhetorical structure theory by Mann and Thomson (1988). The background knowledge related to our paper will be discussed in section 2. Data used in this work is described in section 3. In section 4, principles for segmenting Thai EDU are proposed. Problems arisen with Thai EDU segmentation are explored in section 5. The last section will be the conclusion.

2 Background knowledge

To analyze the structure of text, the text has to be segmented into pieces of information and linked together to reflect the coherence of text. Rhetorical structure theory (RST), one of the most widely used in both linguistics and computational linguistics, was proposed by Mann and Thomson (1988) to explain discourse structure of written texts. Briefly, RST explains the discourse relation of two spans of texts. It explains how parts of text are organized and formed into a larger structure of text which can be represented as a tree structure. For any two spans of text, one of them will have a specific relation to the other. The one that is more essential is the nucleus while the other one functioning as a supporting text is a satellite unit. The discourse tree is described on the basis of successive rhetorical relation between these discourse units. The terminal node of the tree structure represents the minimal unit of the discourse called elementary discourse unit or EDU. Relation that holds between two EDUs can be mononuclear or multinuclear. Mononuclear relation holds between two units which are a nucleus and a satellite. Multinuclear relation holds between two units which are both nucleus. An example of RST analysis of an English text is shown in Figure 1. In this example, the structure is composed of six discourse unit. Units 2-6 are hold together with the relation LIST. All of these units then have a relation PURPOSE with the first discourse unit.

27th Pacific Asia Conference on Language, Information, and Computation pages 491－498

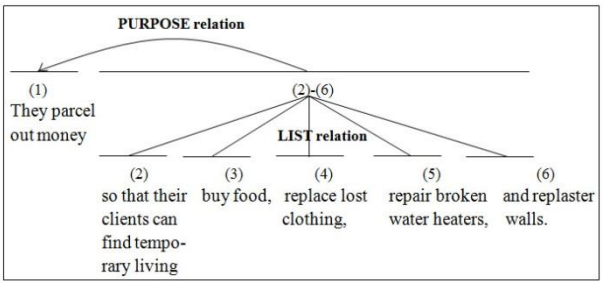

Figure 1. RST tree structure of English text "They parcel out money so that their clients can find temporary living, buy food, replace lost clothing, repair broken water heaters, and replaster walls." (Carson and Marcu, 2001)

However, RST does not specify what minimal discourse unit should look like. It only provides general explanation of the relation among those units in discourse. Later, Carson and Marcu (2001) who were trying to interpret and make use of the theory, proposed a guideline to determine an EDU in English in his discourse tagging reference manual for building an annotated RST corpus (Carson et al., 2001; Carson and Marcu, 2001). Their EDUs were based on the balance between granularity of tagging and ability to identify units consistently. It is well-recognized that their EDU determination is widely accepted, and thus, has been adapted in other research concerning the use of EDU. Carson and Marcu's EDU is not always a clause or sentence. Phrases can be EDU too but with restricted conditions. Coordinated verb phrases are not treated as separate EDUs if they are transitive verbs sharing the same direct object or intransitive verbs sharing a modifier.

There are a few studies on Thai discourse structure in computational aspect. Those studies determined an EDU differently. That is, Sinthupoun (2009) and Katui et al. (2012) took only clauses as EDUs while Charoensuk (2005) took clauses and phrases with strong discourse markers as EDUs. Charoensuk and Katui et al.'s works are based on RST. However, they did not provide a clear explanation of what should be considered an EDU in Thai. In this paper, our focus is proposing a guideline for determining Thai EDUs boundaries and exploring problems in segmenting Thai EDUs.

3 Data collection

The data used in this paper are collected from the Thai National Corpus (TNC). We choose only written academic texts because written and spoken languages differ on the structure of discourse. Our written data are randomly selected from 3 domains which are liberal arts, social sciences, and sciences, about 2,000 EDUs in total. Carson and Marcu's principles for English EDU determination are adapted and adjusted to suit the Thai data. At the end, the basic principles for Thai EDU segmentation are listed as the guideline for segmentation. The guideline will be discussed in the next section following by problems of Thai EDU segmentation.

4 Basic principles for Thai EDU segmentation

In this section, we present a guideline for segmenting Thai EDUs on the basis of the data as described in the previous section. Our goal is to determine the minimal units in every possible structure in discourse. The proposed principles to segment Thai EDU must be clear enough to be used consistently. After segmenting, those EDUs should be able to combine together to reflect the rhetorical relation holding between them.

In this study, the conventions used in our examples are as follows. The EDUs are marked in square brackets. Boldface and italic are used to highlight items being mentioned. Subscripts indicate the number of unit. We follow Carlson and Marcu's (2001) basic idea that clauses and noun phrases with strong markers are treated as EDUs. The proposed principles to determine what is or is not a Thai EDU are listed below.

4.1 Finite clauses

Finite verb is a verb form that can function as a root of a clause. In some languages, finite verb can be inflected for gender, person, number, tense, aspect, mood, and/or voice. However, Thai is an isolating language, its verbs do not inflect to show whether they are finite or non-finite. The criterion to test whether the verb in question is finite or non-finite is to insert an auxiliary such as ต้อง-'must', ควร-'should', จะ-'irrealis marker', เคย-'perfective marker' or กำลัง-'progressive marker'. Only finite verbs can co-occur with those words (Hoonchamlong, 1991; Yaowapat and Prasithrathsint, 2008). A clause can be classified into a finite clause and a non-finite clause according to finiteness of verb. Like Carson and Marcu's basic principle, we treat a finite clause as EDU but with some exception which will be discussed later. If it starts with a discourse

marker, that marker is treated as a part of EDU. In contrast, non-finite clause is not treated as EDU.

A finite clause can be independent or dependent clause. Independent or main clause is a clause that can stand alone while dependent or subordinate clause cannot stand alone and always depends on the main clause. Independent and dependent clauses can link together with a subordinate conjunction and hold mononuclear relation whereas two independent clauses can be combined with a coordinate conjunction and hold multinuclear relation between them.

On the basis of function, a dependent clause can be divided into **subject/object clause**, **finite relative clause,** and **adverbial clause.** According to Carson and Marcu's EDU determination, relative clause and adverbial clause are marked as EDUs while subject or object clause is not EDU. Furthermore, **coordinate clauses** are also treated as EDUs while **coordinate verb phrases** are not. We will discuss more about these types of dependent clause below.

4.1.1 Subject and object clause

A clause functioning as subject or object of predicate is not treated as separate EDU because it is not a modifying part of any text portion and cannot be omitted or separated into a stand-alone unit. Moreover, subject or object clause does not hold any relation to the matrix clause. Example of subject clause is shown below in boldface.

[**ผู้จบปริญญาเอกด้านวิทยาศาสตร์ต้องมีคุณสมบัติอย่างไรบ้าง**]$_1$

[What qualification should **one who receives a doctorate of science** have?]$_1$

4.1.2 Finite relative clause

A finite relative clause is a type of dependent clause and also a noun modifying clause. It will be treated as an EDU. In Thai, relative clause can be formed by either gap strategy or pronoun retention strategy. A relative clause formed by gap strategy does not contain any overt coreference to the head noun while a relative clause formed by pronoun retention contains a pronoun realizing the head noun in the relative clause. The clause may or may not be introduced by a relativizer. Thai relativizers include ที่-'that', ซึ่ง-'that', and อัน-'that'. (Yaowapat and Prasithrathsint, 2008). Example is shown below with relative clause in boldface.

[เนื่องจากขาดการศึกษาและวางนโยบาย]$_1$[**ที่ชัดเจน**]$_2$[**อันจะทำให้ประชาชนสามารถตัดสินใจได้**]$_3$

[since Ø lack of studying and planning policy]$_1$[**that is clear**]$_2$[**which will make people be able to decide**]$_3$

4.1.3 Adverbial clauses

Adverbial clause is a clause that combines with the other clause to give additional information through some rhetorical relation of time, manner, condition, reason, etc. Generally, this type of dependent clause is marked by a subordinate conjunction. Each type of subordinate conjunction is an important clue for identifying rhetorical relation because its grammatical meaning can tell what kind of rhetorical relation two clauses are holding. For instance, purposive conjunctions เพื่อ-'for' and เพื่อว่า-'for that' show purpose relation while contrastive conjunctions แต่-'but' and ส่วน-'whereas' show contrast relation between two clauses. (Chanawangsa, 1986; Matthiessen, 2002) The following example shows how adverbial clause in boldface is segmented into EDU.

[ให้ความสำคัญแก่การวางโครงเรื่อง]$_1$[ที่สลับซับซ้อน]$_2$[**เพื่อหลอกล่อให้คนอ่านเดาเรื่องไม่ออก**]$_3$

[Ø emphasize on plot planning]$_1$[which is complicated]$_2$[**in order to make the readers unable to predict the story**]$_3$

4.1.4 Coordinate clauses

Coordinate clauses are composed of two independent clauses with or without a coordinate conjunction. Note that coordinate clauses are different from coordinate verb phrases in the way that verbs in coordinate clauses do not share the same object or modifier while verbs in coordinate verb phrases always share the same object and modifier. We treat coordinate clauses as EDUs since they hold elaboration relation. On the other hand, we do not separate coordinate verb phrases because they do not have any rhetorical relation to one another. The following examples show how coordinate clauses and coordinate verb phrases are segmented respectively.

[ความยากจนเป็นปัจจัยนำไปสู่การเกิดพยาธิสภาพแก่ปัจเจกบุคคล]$_1$ [และมีผลกระทบต่อส่วนรวม]$_2$

[Poverty is a cause of individual pathology]$_1$[and affects the community at large]$_2$ [แต่หลายส่วนลอกและเพิ่มเติมมาจากกฎหมายตราสามดวง]$_1$

[But many parts were copied and inserted from the Three Emblems Law]$_1$

4.2 Non-finite relative clauses

We do not treat non-finite relative clause as a separate EDU because of its non-finite status of verb. Non-finite relative clause or reduced relative clause is a type of noun modifier without a relativizer. The verb in this type of relative clause is non-finite, therefore, cannot co-occur with auxiliaries or tense-aspect markers. For instance, "ดี "in "คนดี"-'nice people' is a non-finite relative clause used for modifying the head noun "คน" (Yaowapat and Prasithrathsint, 2006). Example of non-finite clause is show below. Text in boldface is considered a non-finite clause.

[โดยได้แสดงวิธีการวิเคราะห์สาร**สำคัญ**จากตำนานเรื่องอีดีพุส]$_1$

[By demonstrating an analysis of **important** contents from the Oedipus myth]$_1$

4.3 Clausal complements

A complement is a constituent of a clause and an obligatory element that completes the meaning of its head (Dowty, 2003). It can be in the form of phrase or clause. In case of clausal complement, its verb can be either finite or non-finite. Finite causal complements are found in complements of attributive verbs. Attributive verbs include verbs of reporting speech and verbs of cognition. Examples of attributive verb in Thai are ยอมรับ-'accept', คิด-'think', เชื่อ-'believe', แสดง-'show', สันนิษฐาน-'assume', เสนอ-'propose', รู้-'know', อธิบาย-'explain', แนะนำ-'suggest', ตัดสินใจ-'decide', สมมติ-'suppose', ถาม-'ask', สงสัย-'doubt', etc. The clausal complements may be introduced by a complementizer ว่า or ที่. We treat clausal complement of attributive verb as a separate EDU since it shows attributive relation to its verb head. The following example shows EDUs with attributive verb in italic and its clausal complement in boldface.

[ชี้ให้เห็น*ชัดเจน*]$_1$[**ว่ามีการละเมิดสิทธิขั้นพื้นฐานของประชาชน**]$_2$

[Ø point out *clearly*]$_1$[**that there is violation of citizens' fundamental rights**]$_2$

In contrast, non-finite clausal complement is not treated as EDU. According to Jenks (2006), the complements in Thai are usually introduced by infinitival complementizer. The complementizer ที่จะ-'that+irrealis marker' and ที่ว่า-'that+say' is used to introduce the clausal complement of noun while ที่จะ and จะ is found in the clausal complement of verbs, except for that of attributive verbs mentioned above. The following examples show an EDU containing clausal complement of noun and of verb in boldface and their heads in italic.

[บทความนี้มี*วัตถุประสงค์***ที่จะศึกษาเปรียบเทียบลักษณะเด่นและลักษณะร่วมระหว่างเรือนพื้นถิ่นของกลุ่มไทลาว**]$_1$

[This article has *an objective* **to compare outstanding characteristics and common characteristics among local houses of Lao Tai people**]$_1$

[หาก*พร้อม***ที่จะปลูกสร้างเรือนใหม่**]$_1$[จึงแยกเรือน]$_2$

[if Ø *(be) ready* **to build a new house**]$_1$[then Ø separate the house ('move to the new house')]$_2$

4.4 Serial verb construction

Serial verb construction (SVC) is one of the common characteristics of the Thai language. Thai SVC can be classified into basic and non-basic types. The basic SVC consists of two contiguous verb phrases with no overt linker while non-basic type consists of two or more verbs interrupted by markers or objects of verb (Thepkanjana, 1986; Takahashi, 2009). According to Foley and Olson (1985), each verb in SVC has the same status as predicate and they are all finite. Moreover, there are some studies about negation in Thai SVC. It is found that negation word ไม่-'not' can occur in front of the first verb and also in the middle of serial verbs (Takahashi, 1996). Though this evidence proves the finiteness status of Thai verbs in serial, SVC expresses only one unite single event and represents one piece of information. This comes to our decision that SVC should be treated as a single clause and segmented into a single EDU. The following example is Thai SVC with direct object in the middle of two verbs. The whole construction is treated as one EDU with serial verbs in boldface.

[ขณะเดียวกันก็**รอคอย**โชคชะตา**มาพลิกผัน**ชีวิต**ให้แปรเปลี่ยนไป**]$_1$

[Meanwhile, Ø **waiting** for the destiny to **come** and **change** the life]$_1$

However, if there is an attributive verb within SVC, that SVC should be broken into separate EDUs. This only occurs with grammaticalized SVC which contains a grammaticalized verb ว่า-'say/complementizer' The following example shows SVC that is broken into two EDUs because it contains an attributive verb คิด-'think'. The grammaticalized verb ว่า-'say/comp.' plays a role of a complementizer rather than a

verb since it cannot co-occur with negation word.

[เพราะเขา**คิด**]$_1$[**ว่า**เขา**ขาด**โอกาสทางธุรกิจ]$_2$

(Literally) because + he + **thinks** + **say/that** + he + **lacks** + opportunity + business

(Translation) Because he thinks that he lacks business opportunity.

*เพราะเขาคิดไม่ว่าเขาขาดโอกาสทางธุรกิจ

(Literally) because + he + think + not + say/that + he + lacks + opportunity + business

4.5 Cleft

Cleft is one of focusing devices used to emphasize a particular element. Although it appears as a complex clause consisting of one independent and one dependent clause, there is no rhetorical relation between them. Like Carson and Marcu's criteria for English, Thai cleft is not treated as a separate EDU. Thai cleft construction can be noticed by the copula เป็น-'be' or คือ-'be', which is the main verb of the whole cleft construction followed by a cleft clause, which is usually a relative clause (Taladngoen, 2012). Thai cleft is treated as a part of one EDU as in the following example.

[เขาเป็นคนที่ทอดทิ้งภรรยาให้เดียวดาย]$_1$

[He is the one who abandons the wife]$_1$

[คนไหนคือคนที่นิดแอบชอบ] $_1$

[Which one is the man whom Nid like]$_1$

4.6 Phrases with strong markers

Phrases can be EDUs if they are preceded by strong discourse markers. The strong discourse markers are markers that not only function as connectors but also have strong meaning to show relation to other units in discourse. These markers are important clues to identify EDU boundaries and discourse relation between discourse units. In Thai, we found two kinds of strong markers. One shows example relation and the other shows purpose relation. Examples of Thai strong discourse markers are เช่น...ฯลฯ-'for example…etc', ได้แก่...เป็นต้น-'for example…etc', ยกตัวอย่างเช่น-'for example', อย่างเช่น-'for example', เพื่อ-'for', etc. The markers are not strong makers and do not make the following phrases an EDU if they do not show neither example nor purpose relation. The boldface in the following examples show strong discourse markers followed by noun phrases.

[ตำนานปรัมปราเป็นการอธิบายถึงกำเนิดของจักรวาล โครงสร้างและระบบของจักรวาล มนุษย์ สัตว์ ปรากฏการณ์ทางธรรมชาติ]^1[**เช่น** ลม ฝน กลางวัน กลางคืน ฟ้าร้อง ฟ้าผ่า]^2

[Legend is the explanation about the creation, structure, and system of the universe, human beings, animals, natural phenomenon]$_1$[**such as** wind, rain, day, night, thunder, and lightning]$_2$

[กู้หนี้ยืมสินมา]^1[**เพื่อ**การต่อสู้คดี]^2

[Ø borrow money]$_1$[**for** fighting the case]$_2$

In addition, noun phrases in the form of parentheticals, name of the title and author, and other nominal units linking with the body of the text are possible to be EDUs. The following example shows how noun phrase in parenthesis is marked as an EDU.

[อพยพมาจากเวียงจันทร์]$_1$ [**(ลาวเวียง)**]$_2$

[Ø migrated from Vientiane]$_1$[**(Lao Vieng)**]$_2$

4.7 Same unit construction

Sometimes, a clause is split up by an insertion of another clause. Carson and Marcu (2001) proposed a multinuclear pseudo-relation called "same-unit" which is the relation holding between two parts of the clause that is being interrupted. Though being separated part, the same-unit construction is treated as one single EDU. Same unit constructions can be found in construction with relative clauses, appositives, and parentheticals. The following example shows an embedded unit in normal font and a split EDU in boldface. The units subscripted as 1 and 3 are same unit constructions and are treated as one EDU.

[**ต่อมาในสมัยหลังสมัยใหม่**]$_1$ [(Post-modern)]$_2$ [**ได้เกิดวรรณกรรมแนวทดลอง**] $_3$

[**Later in post-modern period**]$_1$ [Post-modern]$_2$ [**there comes an experimental literature**]$_3$

4.8 Punctuation

Punctuation is treated as a part of EDU. In Thai, some punctuation can be used to identify EDU boundary. From the data, we observed that punctuation that is always at the end of EDU is question mark (?). Punctuation that is in pairs and usually found at the beginning and the end of EDU is parenthesis ((…)) and quotation marks ("…"). Other punctuation such as dash (-), separator in numbered lists, comma (,), period (.), colon (:), semi colon (;), Thai punctuation

used to abbreviate certain words (ฯ), Thai punctuation used to indicate more of a like kind (ฯลฯ), and Thai punctuation used to indicate repetition (ๆ) usually appear inside EDU and do not play a role in EDU boundary identification.

5 Implementation

To ensure that the proposed principles above are suitable for automatic segmentation, we did a pilot on automatic EDU segmentation. A set of training data (90%) and testing data (10%) are prepared using this guideline. The system relies on Thai word segmentation and POS tagging as preprocessing. We used support vector machine training algorithm to build a model that assigns EDU boundaries of strings of texts. In a preliminary experiment in which 240 EDUs are used in the testing, the precision is 95% and the recall is 70%. This indicates that the proposed principles are practical for automatic EDU segmentation.

6 Problem with Thai EDUs segmentation

Based on the use of the proposed principles on the test data, we found some characteristics of the Thai language that pose difficulties in identifying EDU boundaries. The problems we encountered are as follows.

First, Thai verbs have only one form and are not inflected for any grammatical information. Therefore, they are difficult to be determined whether they are finite or non-finite. But since finiteness of verb is the main criterion for EDU determination, this topic becomes an issue for both manual and automatic EDU segmentation. For manual EDU segmentation, it can be solved because there are criteria to test whether the verb is finite or non-finite. Since a finite verb is the locus of grammatical information such as tense, aspect, and mood, we can test finiteness of verb by observing whether the verb in question co-occur with time adverbs and aspect/mood markers such as จะ-'irrealis marker', เคย-'perfective marker', กำลัง-'progressive marker', and แล้ว-'perfective marker'. In the case that there is no overt marker, we can try inserting some of those markers to verify its finiteness. In a similar way, we can test whether the verb is non-finite by inserting infinitival markers จะ‘irrealis marker’, ที่จะ-‘that+irrealis marker’ and ที่ว่า-‘that+say’ since a non-finite clause is usually introduced by these markers

However, testing finiteness of verb by inserting tense, aspect, mood markers, and infinitival markers requires Thai native speaker to judge whether the sentence is valid or not. This method is not suitable for automatic segmentation. How to determine finiteness automatically is a challenging task.

The second problem is about Thai compound noun. In Thai, a new word can be created by forming a compound noun. One pattern of noun compound is a noun + a transitive verb (+ a noun). For example, หม้อกรองอากาศ-‘air filter’ is composed of a noun หม้อ-‘pot’, a transitive verb กรอง-‘filter’, and a noun อากาศ-‘air’. This compound noun may be incorrectly detected as a sentence by a machine because its pattern is the same as a sentence (Kriengket et al., 2007). Thus, to avoid this kind of mistake in EDU segmentation, compound noun boundary must be disambiguated first.

The third problem is about syntactic ambiguity of a relative clause as in this example ลูกหลานของ*คน***ยากจน**-‘descendants of *people* **that are poor**’. The verb ยากจน-‘poor’ in boldface can be analyzed as a relative clause without a relativizer or a non-finite relative clause with คน-‘people’ in italic as its head noun. The difficulty in EDU segmentation is that this type of relative clause does not have any clue to show that it is a non-finite clause and should not be marked as EDU. Moreover, the word ยากจน-‘poor’ can be seen as a modifying verb and the string คนยากจน-‘people + poor’ can be analyzed as a compound noun-‘poor people’. In the latter case, it will be the problem of compound noun identification discussed earlier.

The fourth problem is concerned with Thai SVC. This is not quite a problem when segmenting EDUs manually. But when it comes to segmenting EDUs by a machine, Thai SVC identification can be a difficult task. Since Thai SVC is a complex predicate structure consisting of two or more finite verbs in which each verb can have its object, it can be very confusing whether each verb phrase should be segmented into separate EDU or not. For example from the previous section, the verb รอคอย-‘wait’ takes โชคชะตา-‘destiny’ as its direct object, serialized verbs มาพลิกผัน-‘come + change’ takes ชีวิต-‘life’ as its direct object, and serialized verbs ให้แปรเปลี่ยนไป-

'give + alter + go' has no object. The correct EDU segmentation is that the whole SVC should be marked as one single EDU. This is why automatic segmenting SVC is a challenging task.

[ขณะเดียวกันก็**รอคอย**โชคชะตา**มาพลิกผัน**ชีวิต**ให้แปรเปลี่ยนไป**]

(Literally) meanwhile + discourse marker + **wait** + destiny + **come** + **change** + life + give + alter + go

(Translation) Meanwhile, Ø wait for the destiny to come and change the life.

The fifth problem is about clauses with no overt discourse marker. Discourse marker is not only an important clue to help identify the EDU boundary but it also signals the type of rhetorical relation holding between clauses. Normally, a subordinate clause and coordinate clause are linked to the other clause by a discourse marker. However, it is possible for two clauses to have rhetorical relation to each other without a discourse marker between them. As in the example below, two clauses are holding consequence relation between them without a consequence marker. Without the presence of overt marker, a machine may find it difficult to identify EDU boundary.

[นโยบายพลังงานเป็นเรื่องใหญ่]$_1$ [สามารถกระทบชีวิตคนทุกคนทั้งโดยตรงและโดยอ้อม]$_2$

(Literally) [policy + energy + be + big deal] [Ø + can + affect + every life + both + directly + and + indirectly]

(Translation) Energy policy is a big deal (because it) can affect everyone both directly and indirectly.

The ambiguity of spaces can also cause a big problem for EDU segmentation. In Thai, text is written without a space between words. Instead, a space is used in Thai text to segment parts of discourse. However, the use of space in Thai text can be ambiguous because not every space functions as a sentence or clause separator. This is because Thai does not have strict and precise convention of using a space. Thus, we cannot rely on every space to determine EDU boundaries. To illustrate, the following sentences are all correct and the meanings are the same, even though the spaces are placed in different positions. Still, their EDU boundaries are all the same.

[คนเล่านิทานไม่ได้เล่า]^1[ว่านางเอื้อยในนิทานเรื่อง "ปลาบู่ทอง" มีหน้าตารูปร่างหรือมีนิสัยใจคออย่างไร]^2

[คนเล่านิทานไม่ได้เล่า]^1[ว่านางเอื้อยในนิทานเรื่อง "ปลาบู่ทอง" มีหน้าตา รูปร่าง หรือมีนิสัยใจคออย่างไร]^2

[คนเล่านิทานไม่ได้เล่า]^1[ว่า นางเอื้อยในนิทานเรื่อง "ปลาบู่ทอง" มีหน้าตา รูปร่าง หรือมีนิสัยใจคออย่างไร]^2

(Translation) The story teller did not tell how Nang Uay in "Pla Boo Thong" looks like or what personality she has.

In addition, Thai words can have multiple meanings. For instance, a discourse marker ส่วน- 'whereas' can also be a noun meaning 'part'. Yet, a discourse marker of one form may have several functions. For example, the word แล้ว can be a sequential marker meaning 'then' and also a perfective marker meaning 'already'. Therefore, POS tagging has to be applied correctly before doing automatic EDU segmentation.

7 Conclusion

The principles of Thai EDU determination proposed in this paper can be used as a guideline to segment Thai EDUs in written text. The creation of EDU segmented corpus is the first step in building a resource for the study of Thai discourse structure and automatically EDU segmentation. We believe that EDU is a suitable unit to be an input for Thai text processing because Thai writing system does not use any explicit marker for sentence boundary. Thai discourse is a continuation of text chunks holding together with or without a connection or a discourse marker. However, we found that determining EDUs in Thai text is not clear and easy especially for a machine. Further studies on automatic EDU segmentation using machine learning algorithms should be explored. But in order to do this, a corpus which is EDU segmented using the principles proposed in this study has to be built first. Therefore, the guideline proposed in this paper is the essential first step for this line of study.

Acknowledgments

This research is a part of the first author's thesis. It is partially supported by the Chulalongkorn University Centenary Academic Development Project and by the Ratchadaphiseksomphot Endowment Fund of Chulalongkorn University (RES560530179-HS).

References

Carlson, L. and Marcu, D. 2001. Discourse Tagging Manual. ISI Tech Report ISI-TR-545. July 2001.

Carlson, L., Marcu, D., and Okurowski, M.E. 2001. Building a DiscourseTagged Corpus in the

Framework of Rhetorical Structure Theory. In Proceedings of the 2nd SIGdial Workshop on Discourse and Dialog, Aalborg, Denmark.

Chanawangsa, W. 1986. Cohesion in Thai. Ph.D. dissertation, Georgetown University.

Charoensuk, J. 2005. Thai Elementary Discourse Unit Segmentation by Discourse Segmentation Cues and Syntactic Information. Master's thesis, Kasetsart University, Bangkok.

Dowty, D. 2003. The Dual Analysis of Adjuncts/Complements in Categorial Grammar. In Ewald Lang, et. al. (eds.) Modifying Adjuncts. NewYork: MoutondeGruyter.

Foley, W.A. and Mike, O.1985. Clausehood and verb serialization. In Nichols, Johanna and Anthony C. Woodbury (eds.) Grammar Inside and Outside the Clause: Some Approaches to Theory from the Field, 17-60. Cambridge, Cambridge University Press.

Hoonchamlong, Y. 1991. Some issues in Thai anaphora: A government and binding approach. Ph.D.dissertation, University of Wisconson-Madison.

Jenks, P. 2006. Control in Thai. In Variation in control structures, ed. M. Polinksy and E. Potsdam.

Ketui, N., Theeramunkong, T., and Onsuwan, C. 2012. A rule-based method for Thai Elementary Discourse Unit Segmentation (TED-Seg). Proceedings of the 7th International Conference on Knowledge Information and Creativity Support Systems (KICSS) 2012, Melbourne, Australia.

Kriengket, K., Kosawat, K., and Anchaleenukul, S. 2007. A Computational Linguistics Study of Compound Nouns in Thai, In Proceedings of the Seventh International Symposium on Natural Language Processing (SNLP 2007), pp. 31–36.

Mann, W. and Thompson, S.1988. Rhetorical structure theory: Toward a functional theory of text organization. Text, 8(3).

Mann, W., Matthiessen, C., and Thompson, S.1992. Rhetorical structure theory and text analysis. In Mann and Thompson (eds.) Discourse Description: Diverse Linguistc Analyses of a Fundraising Text. Amsterdam: John Benjamins.

Matthiessen, Christian M.I.M. 2002. Combining clauses into clause complexes: a multifaceted view. In Joan Bybee & Michael Noonan (eds.), Complex sentences in grammar and discourse: essays in honor of Sandra A. Thompson. Amsterdam:Benjamins. 237-322.

Taladngoen. U. 2012. The Semantics-Syntax Analysis of 'pen' in Thai. M.A.thesis, Srinakharinwirot University.

Takahashi, K. 1996. Negation in Thai Basic Serial Verb Constructions. M.A. thesis, Chulalongkorn University.

Takahashi, K. 2009. Basic Serial Verb Constructions in Thai. Journal of the Southeast Asian Linguistics Society 1.

Thepkanjana, K.1986. Serial Verb Constructions in Thai. Ph.D. dissertation, University of Michigan.

Sinthupoun, S. 2009. Thai Rhetorical Structure Analysis. (Ph.D dissertation). National Institute of Development Administration, Bangkok.

Stowell, T. 2005. Appositive and Parenthetical Relative Clauses. In Hans Broekhuis, Norbert Corver, Jan Koster, Riny Huybregts and Ursula Kleinhenz (eds.) Organizing Grammar: Linguistic Studies in Honor of Henk van Riemsdijk, Berlin/New York, Mouton de Gruyter.

Yaowapat, N. and Prasithrathsint, A. 2006. Reduced relative clauses in Thai and Vietnamese. In Sidwell, Paul, and Uri Tadmor (eds.) SEALS XVI: Papers from the Sixteenth Annual Meeting of the Southeast Asian Linguistics Society. Canberra: Pacific Linguistics.

Yaowapat, N. and Prasithrathsint, A. 2008. A typology of relative clauses in mainland Southeast Asian languages, in The Mon-Khmer Studies Journal, vol. 38, pp. 1-23.