A Computational Linguistics Study of Compound Nouns in Thai[1]

Kanyanut Kriengket[2] Krit Kosawat[3] Sunant Anchaleenukul[4]

[2,4] Department of Thai Language, Faculty of Arts, Chulalongkorn University,
Pathumwan, Bangkok, 10330, Thailand
Tel. +6686-558-0735, +662-218-4686
e-mail : luckysep13@hotmail.com, asunant@pioneer.chula.ac.th

[3] Human Language Technology Laboratory,
National Electronics and Computer Technology Center,
National Science and Technology Development Agency,
Ministry of Science and Technology,
Klong Luang, Pathum Thani, 12120, Thailand
Tel. +66-2-564-6900 Ext. 2212
e-mail : krit.kosawat@nectec.or.th

Abstract

This paper presents the boundary of Thai compound nouns, the internal relation, and the properties of the predicates and arguments by the approach of computational linguistics. Nouns referring to scientific instruments are selected as the representation of Thai compound nouns. The logical structure is used for finding the boundary of the compounds, the relation and the properties of any elements, as well. The result is that the predicate functions, as the head of the compound nouns, mostly expresses [INSTRUMENT], [SHAPE], or [CONTAINER]. For the boundary of the compounds, it can be told by calculating the number of the arguments which occur with their predicates. The compounds have at most 2 arguments. The first one expresses [PURPOSE], [CHARACTERISTIC], [STATE], [METHOD], or [PROCESS]. The other expresses [PURPOSE] or [CHARACTERISTIC].

1 Introduction

A compound word is an interesting topic that is studied by many researchers, for example, Vichin Panupong (1970), Nisa Udomphol (1964), Anong Iangubol (1982), Sunant Anchaleenukul (2004),Anchalee Singnoi (2005), Jensen (1990), Lieber (1983, 1992), etc. The compound words in any researches are mostly short and consist of 2 – 3 elements (words). However, compound words nowadays are longer and complex especially in Thai, for example, กระจกแบบโซนเทมเปอร์ /kràcòk bè:p so:n thempɤ̂:/ (zone tempered glass) เครื่องถ่วงล้อ /khrɯ̂aŋ thùaŋ lɔ́:/ (wheel balance) ตู้อบความร้อนฆ่าเชื้อโรค /tû: ʔòp khwamrɔ́:n khâ: chɯ́arô:k/ (hot-air sterilizer). Their structures are like a phrase and sentence, as well. When they appear in texts, it is quite difficult to tell the end of the words. Besides, the long and complex structure of compounds is an important problem for the automatic segmentation and evaluation of their boundaries because a computer does not have a cognition system like humans to understand any words or structures clearly. So, to study and analyze these complex structures for finding the boundary of the compound nouns, the internal relation, and the properties of the predicates and arguments can solve this

[1] This paper was previously presented as an academic poster in NAC2007, March 2007.

problem and program a computer to automatically segment words.

In general, studying the surface and deep structures is the method for telling the relation between any elements of the compounds. For example, the word "เครื่องวัด"/khrûaŋ wát/ (gauge) is a compound noun made of 2 words: "เครื่อง"/khrûaŋ/(instrument) and "วัด"/wát/ (measure). It has 2 elements which are a head "เครื่อง"/khrûaŋ/ and a modifier "วัด"/wát/. The head is the left member that is a noun expressing [INSTRUMENT], while the modifier is the right member that is a transitive verb expressing [ACT]. The structure is then "N+VT ⇨ N (compound noun)". The word "หม้อกรองอากาศ" /mɔ̂: krɔŋ ʔa:kà:t/(air filter; air cleaner) is made of 3 words: "หม้อ"/mɔ̂:/(pot), "กรอง"/krɔŋ/(filt), and "อากาศ"/ʔa:kà:t/(air). The head "หม้อ"/mɔ̂:/ (pot) is the left member that is a noun expressing [CONTAINER]. The two words following on the right are the modifiers. The word "กรอง"/krɔŋ/(filt) is a transitive verb expressing [ACT], and the word "อากาศ"/ʔa:kà:t/ (air) is a noun expressing [SUBSTANCE]. The structure is then "N+VT+N ⇨ N (compound noun)", which looks like a sentence. So, if the compound nouns are studied only on their structures, it will be hard to divide whether this structure is a sentence or a word.

Although there are some researches about semantic relations and automatic interpretation of compound nouns such as Vanderwende, L. (1994), Rosario, B. and Hearst, M. (2001), Girju, R., Badulescu, A., and Moldovan, D. (2003), and Nastase, V. and Szpakowicz, S. (2003), they mostly depends on word definitions in dictionaries and semantic relations or semantic networks in WordNet. It is hard for Thai because word definitions in Thai dictionaries are still too vague to be used as a method for finding semantic relations. So, the evaluation of the boundary of the compound nouns by using the number of the arguments occurring with the predicates, the internal relation, and the properties of any elements (a predicate and an argument) should be recognized and help the automatic segmentation.

2 Definition and Source

For this paper, a compound word is the word made of 2 or more free-morpheme words. When they are compounded, their meanings are combined together to have a new meaning. For Thai compounds, there are 2 types which are an endocentric compound and an exocentric compound. If the meaning of the compound word can be guessed from its elements, it is called "endocentric compound". But if not, it is called "exocentric compound". For instance, the word "mailman", an endocentric compound, gives the basic meaning of the compound word from its head "man", so it means "a man who delivers the mail. The word "redcap", an exocentric compound, has only heads. It means "a porter in a railway", not a kind of caps. Also, the compound word can't be separated or added any words.

Besides, nouns referring to scientific instruments are used as the case study of Thai compound nouns because the compound nouns are productive, which means that they can continuously be produced to name things or inventions. Also, there are many new scientific inventions such as callipers, water-cooled engine, and hot-air sterilizer. Only some of them have had Thai names created from single words, so the compound nouns have the important roles in this way. Hence, the nouns referring to scientific instruments are the good examples for this study to find the boundary, the relations between members, and the properties of each member. This group of nouns should be defined as the nouns referring to instruments, tools and objects used in scientific fields to give information, do some special functions, or measure, repair, examine, and build things.

3 Logic Structure and the Argument Linking Principle

In order to do so, the main method is to find the logic structure of the compound nouns and their internal relation. The logic structure is the universal structure for any languages. It consists of a predicate and an argument. The predicate (P) is the word that represents the essential function and the meaning of the construction, while the argument is the word that must occur with its predicate. For sentences, there are 2

arguments: (1) an external argument (A_0) to function as the subject, and (2) an internal argument (A_{1-n}) to do other functions.

For compounds, the predicate functions as the head, and the argument functions as the modifier. Like English compounds, Thai compounds only have internal predicate. Although many compounds have a sentence-like structure: subject (noun)+verb(+object (noun)), the compound and the sentence are different. In the sentence, the verb is the predicate to express the meaning of the sentence, the subject is the external argument, and the object is the internal argument. Unlike the sentence, the compound has only the predicate and the internal argument. Though there is a verb in the structure, the verb in the compound needn't be the predicate because the main element to represent the meaning of the word is not the verb, but the noun before it. So, the noun in that position becomes the predicate of the compound, and the other elements are the internal arguments.

For example, in the word "เครื่องวัด"/khrûaŋ wát/(gauge), "เครื่อง"/khrûaŋ/(instrument) is the predicate and "วัด"/wát/(measure) is the arguments. It can be written in relation of P(_,A_1) as:
เครื่อง/khrûaŋ/(instrument)(_,วัด/wát/(measure)).

Moreover, "the Argument-linking Principle" of Lieber (1983 Quoted in Jensen 1990) is also used in this study because it presents 2 rules of member relations in compound words, which are:

(1) An element of a compound must be able to link all its obligatory internal arguments.

(2) A compound stem not linked by an argument-taking stem compounded with it must be interpretable as a restrictive modifier of that stem, i.e. as a locative, manner, instrumental, or benefactive.

According to this principle, it is possible to detect the relation between any elements, the properties of the predicates, and the following arguments. Also, the boundary of the compounds can be evaluated by calculating the number of the argument.

4 The boundary of the compound nouns

Any compound nouns have 2 essential elements: the predicate (P) and the argument (A). The predicate is the important element of the compound. In Thai compound nouns, the predicate can be the right or left member. The argument is the necessary word that occurs with the predicate. Its position will be opposite to the predicate it occurs with. Both elements are in every structure of compound nouns. Thai compound nouns have 3 main structures which are P_1+P_2, A_1+P, and $P+A_1(+A_2)$. The structure P_1+P_2 is found in exocentric compounds. The structures "A_1+P" and "$P+A_1(+A_2)$" are found in endocentric compounds.

For the endocentric compounds, they consist of the predicate and the argument. The predicates have 2 positions: on the right or the left.

Firstly, there are only 2 compound nouns that consist of one predicate as the right member, and one argument as the left member. They can be represented as P(_,A_1). For example, the word "กลอุปกรณ์"/kon ʔùppàʔkɔn/(mechanical device) has the logic structure as:

อุปกรณ์/ʔùppàʔkɔn/(device)(_,กล/kon/(mechanical)).

The word "สามขา"/sǎ:m khǎ:/(tripod) has the logic structure as:

ขา/khǎ:/(leg)(_,สาม/sǎ:m/(three)).

The second position of the predicate is on the left. It is the most found in compound nouns. It can be either a word or a group of words. The arguments following the predicate are the right members, and can occur 2 arguments at most, which are represented as A_1 and A_2 respectively. Like the predicate, they can be either a word or a group of words. The symbol for the relation is that P(_,A_1,A_2). For example,

The word "ก้ามวัด"/kâ:m wát/(callipers) has the logic structure as:

ก้าม/kâ:m/(pincers)(_,วัด/wát/(measure)).

The word "ตู้อบความร้อนฆ่าเชื้อโรค" /tû: ʔòp khwamrɔ́:n khâ: chɯ́arô:k/(hot-air sterilizer) has the logic structure as:
ตู้(_,อบความร้อน,ฆ่าเชื้อโรค).

/tû:/ (_,/ʔòp khwamrɔ́:n/,/khâ: chɯ́arô:k/).
(cabinet)(_,(fumigate with heat),(destroy infection)).

For the exocentric compounds, they only have the predicates with no argument because both predicates play important roles in compounds. They can be written as a logic structure (for example $P_1(Ø)P_2(Ø)$).

For example,

แม่แรง	⇨	แม่(Ø)แรง(Ø)
/mêː rɛːŋ/ (jack)		/mêː/(Ø)/rɛːŋ/(Ø) (mother)(Ø)(power)(Ø)
ไขควง	⇨	ไข(Ø)ควง(Ø)
/khăj khwuaŋ/ (screwdriver)		/khăj/(Ø)/khwuaŋ/(Ø) (drive)(Ø)(screw)(Ø)

In conclusion, Thai compound nouns are made of predicates and of related arguments. Most of the case, two arguments occur with the predicate in the compounds. So, the boundary of the compound nouns can be calculated from the number of the arguments.

5 The internal relation and the properties of the predicates and the arguments

The compound nouns include predicate and related argument(s). Each element has its own property and when several elements compound together, the internal relation between them appears.

For the endocentric compounds, they should be divided into 2 groups according to the positions of the predicates: on the right or the left. The first group is the predicate on the right. It does the function as the head (H) of the compounds to express the essential meaning of the compound nouns. The argument on the left does the function as the modifier (M) to show specific properties. For example, the word "กล อุปกรณ์"/kon ʔùppàʔkɔn/(mechanical device), both predicate and argument are nouns (N), and their semantic relation is [SYSTEM-INSTRUMENT] ([SYS-INS]). In the word "สามขา" /săːm khăː/(tripod), the predicate is a classifier (CLAS), but the preceding argument is a quantifier (QUAN). The semantic relation is [QUANTITY-PART] ([QNT-PART]). These are shown in the table below.

Structures				Elements	Words
Logic	Function	Semantic	POS		
A_1-P	M-H	[SYS-INS]	N-N	กล-อุปกรณ์ /kon/– /ʔùppàʔkɔn/ (mechanic-al)-(device)	กลอุปกรณ์ /kon ʔùppàʔkɔn/ (mechanic-al device)
		[QNT-PART]	QUAN-CLAS	สาม-ขา /săːm/– /khăː/ (three)-(leg)	สามขา /săːm khăː/ (tripod)

Table 1. The structure of endocentric compounds of which the predicates are on the right.

Secondly, the predicate on the left does the function as the head (H) to express the essential meaning of the compound nouns. The predicate's meanings are found in 5 kinds: [INSTRUMENT] ([INS]), [SHAPE] ([SHP]), [CONTAINER] ([CTN], [PURPOSE] ([PUR]), and [PATIENT] ([PAT]). It has 2 arguments following on the right way. The 1st and 2nd arguments modify the predicate and show specific properties. The first argument expresses [PURPOSE], [TYPE], [CHARACTERISTIC] ([CHA]), [STATE] ([STT]), [METHOD] ([MET]), [PROCESS] ([PROC]), [OBJECT] ([OBJ]) or [GOAL]. The other expresses [PURPOSE] or [CHARACTERISTIC]. Both arguments can be an isolated word or a phrase.

The predicate can be a noun or a transitive verb (VT). If the predicate is a noun, two arguments will be a noun, a transitive verb, an intransitive verb (VI), or a verb phrase (VP), which consists of a transitive verb and a noun. The predicate expresses the above meaning except [PURPOSE]. For the arguments, there are 2 ways to express their meanings. If there is one argument, it will express [PURPOSE], [CHARACTERISTIC], or [STATE]. If there are two arguments, the first one will express [METHOD], or [PROCESS], and the second one will be [PURPOSE] or [CHARACTERISTIC].

On the other hand, if the predicate is a transitive verb, it will have one argument. It can be a noun or a transitive verb. The predicate expresses [PURPOSE] and follows with the argument expressing [OBJECT] or [GOAL].

Structures				Elements	Words
Logic	Function	Semantic	POS		
$P\text{-}A_1$ $(\text{-}A_2)$	$H\text{-}M_1$ $(\text{-}M_2)$	[SHP-PUR]	N-V	ก้าม-วัด /kâ:m/–/wát/ (pincers)-(measure)	ก้ามวัด /kâ:m wát/ (callipers)
		[INS-PUR]	N-VP (VT-N)	เครื่อง-(ถ่วง-ล้อ) /khrɯ̂aŋ/–(/thùaŋ lɔ́:/) (instrument)-(balance-a wheel)	เครื่องถ่วงล้อ /khrɯ̂aŋ thùaŋ lɔ́:/ (wheel balance)
		[PROC-PUR]	N-VP (VT-N)-VP (VT-N)	ตู้-(อบ-ความร้อน)-(ฆ่า-เชื้อโรค) /tû:/–(/?òp khwamrɔ́:n/)–(/khâ: chɯ́arô:k/) (cabinet)-(fumigate with heat)-(destroy infection)	ตู้อบความร้อนฆ่าเชื้อโรค /tû: ?òp khwamrɔ́:n khâ: chɯ́arô:k/ (hot-air sterilizer)
		[PUR-OBJ]	VT-N	พัด-ลม /phát/-/lom/ (fan)-(wind)	พัดลม /phát lom/ (fan)
		[PUR-GOAL]	VT-VT	กัน-ชน /kan/-/chon/ (protect)- (crash)	กันชน /kan chon/ (bumper)

Table 2. The structure of endocentric compounds of which the predicates are on the left.

For the exocentric compounds, they have only the predicates with no argument. They both function as the heads (H) of the compounds to express the whole meaning. Also, they mostly pass the metaphorical process to become the compound nouns.

According to the data, the 1st predicate can be 3 parts-of-speech: a noun, a classifier, or a transitive verb. First, If the 1st predicate is a noun, the 2nd predicate will be a noun or a transitive verb. There are 6 semantic relations between both predicates: [PART-PURPOSE], [PART-WHOLE], [THING-PURPOSE], [MAIN-PRODUCT], [MAIN-POWER], and [FUNCTION-PURPOSE]. Secondly, the 1st predicate which is a classifier follows with the 2nd predicate which is a noun to express the relation [CHARACTERISTIC-THING]. The last one is that both predicates are transitive verbs. The semantic relation between any elements is [PURPOSE-METHOD].

Structures				Elements	Words
Logic	Function	Semantic	POS		
$P_1\text{-}P_2$	$H_1\text{-}H_2$	[PART-PUR]	N-VT	หู-ฟัง /hŭ:/-/faŋ/ (ear)- (listen)	หูฟัง /hŭ: faŋ/ (earphone)
		[MAIN-POW]	N-N	แม่-แรง /mɛ̂:/–/rɛ:ŋ/ (mother)- (power)	แม่แรง /mɛ̂: rɛ:ŋ/ (jack)
		[FUNC-PUR]	N-N	สะพาน-ไฟ /sàpha:n/-/faj/ (bridge)-(electric)	สะพานไฟ /sàpha:n faj/ (cut-out)
		[QLT-THN]	CL-N	พวง-มาลัย /phuaŋ/-/ma:laj/ (bunch)-(wreath)	พวงมาลัย /phuaŋ ma:laj/ (steering wheel)
		[PUR-MET]	VT-VT	ไข-ควง /khăj/–/khwuaŋ/ (drive)-(screw)	ไขควง /khăj khwuaŋ/ (screwdriver)

Table 3. The structure of exocentric compounds.

6 Conclusion

The study of Thai compound nouns uses nouns referring to scientific instruments as the case study. We propose that the boundary of the compounds can be calculated by the number of the arguments occurring with their predicates. The compound nouns have at most 2 arguments. Also, each element has the relation and their own properties. The predicates that function as the head of the compounds are mostly nouns. The arguments function as the modifiers to give some special information of the predicate.

For the exocentric compound, it has 2 predicates which can be a noun or a transitive verb to express the meaning as a whole, and mostly pass the metaphorical process before.

For the endocentric compound, it has the predicate and the argument(s). The predicate can be a noun or a transitive verb. If it is a noun, the argument(s) can be a noun, transitive verb, intransitive verb, or phrase. The predicate expresses [INSTRUMENT], [SHAPE], [CONTAINER], [PURPOSE], and [PATIENT]. The 1st argument expresses [PURPOSE], [TYPE], [CHARACTERISTIC], [STATE], [METHOD], [PROCESS], [OBJECT] or [GOAL], and the 2nd one expresses [PURPOSE] or [CHARACTERISTIC].

This study also shows that the structures of Thai compound nouns are very complex because the compounds are productive, and can be made of many POSs. Many of them have the sentence-like structure, but the differences are the predicate and its meaning. The predicate of the sentence is a verb, but of the compound is a noun preceding the verb which functions as the head. The predicates in the sentences represent [ACT], [STATE], or [FEEL], but they represent [INSTRUMENT] in the compound.

Acknowledgement

Thanks to Thailand Graduate Institute of Science and Technology (TGIST), National Science and Technology Development Agency (NSTDA) for supporting the scholarship.

References

Anong Iangubol. 1982. *An Analytical Study of Compound Words in Thai*. Master's dissertation. Chulalongkorn University.

Girju, Roxana, Badulescu, Adriana, and Moldovan, Dan. 2003. *Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations.* Proceedings of HLT-NAACL 2003, Edmonton, pp. 1-8.

Jensen, John T. 1990. *Morphology: Word Structure in Generative Grammar*. Amsterdam, Philadelphia: John Benjamins.

Lieber, Rochelle. 1983. "Argument linking and compounds in English," *Linguistic Inquiry* 14, 251-285. Quoted in Jensen. 1990. *Morphology: Word Structure in Generative Grammar*. Amsterdam, Philadelphia: John Benjamins.

_____. 1992. *Deconstructing Morphology: Word Formation in Syntactic Theory*. The University of Chicago Press, Chicago.

Nastase, Vivi and Szpakowicz, Stan. 2003. *Exploring Noun-Modifier Semantic Relations*. Fifth International Workshop on Computational Semantics (IWCS-5), Tilburg, The Netherlands, pp. 285-301.

Nisa Udomphol. 1964. *Compound Words in Thai.* Master's dissertation, Department of Foundation of Education, Graduate School, Chulalongkorn University.

Rosario, Barbara and Hearst, Marti. 2001. *Classifying the Semantic Relations in Noun Compounds via a Domain-Specific Lexical Hierarchy.* Proceedings of the 2001 Conference on Empirical Methods in Natural Language Proceeding (EMNLP-01), pp. 82-90.

Sunant Anchaleenukul. 2004. *Thai Morphology*. Faculty of Arts, Chulalongkorn University, Bangkok.

Unchalee Singnoi. 2005. *Compound Nouns: Science and Art for Thai Word formation.* Chulalongkorn University Press, Bangkok.

Vanderwende, Lucy. 1994. *Algorithm for Automatic Interpretation of Noun Sequences.* Proceedings of the Fifteenth International Conference on Computational Linguistics, Kyoto, Japan, pp. 782-788.

Vichin Panupong. 1970. *Inter-sentence Relations in Modern Conversational Thai*. The Siam Society, Bangkok.