Taiwan Journal of Linguistics
Vol. 17.1, 49-77, 2019
DOI: 10.6519/TJL.201901_17(1).0002

IS THERE A DICHOTOMY BETWEEN SYNTHETIC COMPOUNDS AND PHRASES IN THAI?*

Kamolchanok Hongthong
Kingkarn Thepkanjana
Wirote Aroonmanakun
Chulalongkorn University

ABSTRACT

This study discusses a structural ambiguity between synthetic compounds and other syntactic phrases in Thai, as they potentially appear identical. Productive Ns and Vs were extracted from the Thai National Corpus. The data were divided into two groups. Group A consists of seven N-V(P) strings with the strongest collocations, each of which exhibits a semantically different relation. Group B is a similar set of N-V(P)s to Group A, but they feature interventions, coordinations, and modifications or alterations within a 5-word span. To test the state of being a lexical or syntactic unit, the Lexical Integrity Hypothesis served as a template to build internal cohesion. The corpus frequency and native speakers' judgment were also taken into account. The comparison shows that Group A has a much higher frequency of occurrence than Group B. There are tendencies that native speakers consider tightly cohesive and frequently occurring strings in Group A as single-whole units, while the status of members of Group B is arguable depending on their Type/Token Ratio and cultural familiarity. As for Thai synthetic compounds, the division between the lexicon and syntax is yet a fuzzy boundary.

Key words: synthetic compounds, Thai, internal cohesion, corpus frequency, entrenchment

---

*This paper is a partial fulfilment of the dissertation titled "Noun-Modifying Clause Constructions in Thai." Financial support from the Thailand Research Fund (TRF) through the Royal Golden Jubilee Ph.D. program (Grant No. PHD/0079/2556) to Associate Professor Kingkarn Thepkanjana, PhD and Kamolchanok Hongthong is acknowledged. The authors' sincere gratitude goes to anonymous reviewers for their insightful comments.

1. INTRODUCTION

Thai has S-V-O word order and is an isolating language, so it does not have any overt linguistic encoding to indicate either the syntactic properties or semantic roles when constituents concatenate together to form complex words, phrases or clauses. Space is visible to indicate a pause or an information break at the end of phrases, clauses or sentences, but not between individual words. Moreover, in terms of modification, the position of the modifier is placed immediately right after the component that it modifies. Whether the modifier is a one-syllable word or a chunky clause, right-branching or head-initial structures apply to it, as this is a characteristic of Thai.

Compounding is a very productive word formation process in Thai (Diller 1992; Snyder 2016) even though affixation is scarce and exclusively found in loan words from Pali-Sanskrit or Khmer, which are agglutinating languages. Compounds can be made up of two (or more) independent words from the same or different syntactic category, as illustrated in Table 1.

Table 1. Examples of compound formation in Thai

Components' part of speech(es)	Example	Meaning
noun + noun	*khɔ:* + *hà:n* 'neck' + 'goose'	toilet syphon
noun + verb	*rɯ:a* + *bin* 'ship' + 'fly'	airplane
noun + modifier	*ɕhô:ŋ* + *khê:p* 'cavity' + 'narrow'	channel
verb + noun	*khǎ:j* + *nâ:* 'sell' + 'face'	humiliate
verb + verb	*kan* + *sà:t* 'prevent' + 'spill'	awning, overhang
noun + modifier + noun	*hŭaj* + *tâj* + *din* 'lottery' + 'under' + 'clothes'	illegal lottery

The current study focuses on synthetic compounds which, to some extent, resemble and differ from phrases at the same time. Syntactically speaking, synthetic compounds appear similar to the way linguistic components form a bigger chunk by the virtue of syntax (Selkirk 1982). Their linear structure seems identical, i.e. argument-predicate order (N-VP). However, synthetic compounds are not as productive as syntactic phrases or sentences that can be used to express brand new existences endlessly. There seem to be some constraints which govern the way compounds are formed (Matthews 1991 cited in Singnoi 2005). One cannot simply concatenate a random predicate and an argument to coin a novel compound. Semantically speaking, the meaning of a synthetic compound is relatively compositional. Just like when one encounters a phrase for the first time, s/he can predict its total meaning from its components (Singnoi 2000). This aspect makes synthetic compounds differ greatly from fully lexicalized compounds whose meaning is opaque, idiosyncratic and conventionalized.

Below are some examples of Thai synthetic compounds.

(1) *máːj* *khwɛ̌ːn* *sûːa*
stick hang clothes
'clothes hanger'

(2) *tam rûːat* *dàp* *phlɤːŋ*
police put off fire
'fire brigade'

(3) *hôːŋ* *ráp* *khɛ̀ːk*
room welcome guest
'living room'

Based on these similarities and differences, questions as to the status of Thai synthetic compounds arise: do they belong to a discrete category? If so, should they be classified as either a lexical or syntactic unit? But if not, are they evidence that the lexicon and syntax are actually intertwined? The arguments in the study are primarily based on internal cohesion, corpus frequency and native Thais' intuitive judgment.

The organization of the paper is as follows. Section 2 illustrates previous attempts to determine compounds and distinguish them from syntactic phrases. Section 3 reviews the basic concepts related to the study. Section 4 explains the data collection. Section 5 and 6 cover two studies, namely, analysis of corpus frequency and native speakers' intuitive judgment. Section 7 is devoted to conclusion and discussion.

2. PREVIOUS STUDIES

This section involves the problematic definition of compounding, the criteria that scholars use to determine compounds, the reasons why such proposed criteria do not work with Thai synthetic compounds, and the attempts Thai linguists have made.

The definition of compounding as "the formation of a new lexeme by adjoining two or more lexemes" (Bauer 2003:46) may seem straightforward and easy to understand. It is specific enough to distinguish this word formation process from affixation. Meanwhile, it is broad enough to cover roots, stems, free and bound morphemes that are involved in compounding across languages over the world. However, there are some intricate problems that abound with the given definition. Taking some examples from English, e.g. the word *over-* in *overstate, overload* or the word *out-* in *outnumber, outrun*. Are *over-* and *out-* understood as bound roots or affixes? Another issue worth pondering is, how "new" is a new lexeme in Bauer's sense? Does it really need to be included in the dictionary by lexicographers in order to be officially recognized as a "new" lexeme?

Fraught with these problems, scholars of the English language attempted to establish a set of criteria to distinguish compounds from phrases.

a) Orthographic criterion

Unlike German, spaces cannot account for the word boundary marker in English. It is also worth pointing out that orthographic words do not necessarily equal lexemes, and vice versa. *Pocket* and *knife* are two orthographic words, whereas *pocketbook* is one orthographic word.

However, native speakers of English prefer to treat both of them as one single unit in their mental lexicon (Jackson and Ze Amvela 2000).

Orthographic convention in English also seems arbitrary. There are known compounds written in a closed form, e.g. *afternoon, firefly, cockpit*, with a space, e.g. *high school, rocking chair, theme park*, and hyphenated, e.g. *runner-up, sister-in-law, commander-in-chief*. Additionally, some have more than one possible option, e.g. *icecream, ice cream* or *ice-cream*. Therefore, spelling cannot serve as a good criterion.

b) Phonological criterion

Kingdon (1966 cited in Dressler 2006) and Chomsky and Halle (1968 cited in Cinque 1993) mention the tendency that English compounds bear stress on the left-hand constituent, whereas syntactic phrases are stressed on the head, i.e. the right-hand constituent. It holds true for most compounds, e.g. '*bookstore, 'watchmaker, 'blackbird*, as well as participle-based compounds, e.g. '*easy-going, 'high-born, 'man-made* (Olsen 2000). However, the claim has been argued against by many (Pennanen 1980; Roach 1983; Bauer 1983; Štekauer, Valera and Diaz 2007) cited in Bauer (2009), because the position of stress when a word appears in isolation may differ from when it appears in a sentence context.

It is difficult to figure out a systematic explanation for a large variability of stress in English compounds. Here are a few pairs of compounds that indicate temporal/location relations, but they exhibit different stress patterns. For example,

(4) hotel 'kitchen 'restaurant kitchen
(5) summer 'night 'summer school
(6) summer 'dress 'winter coat

In short, phonological principles are not sufficient to distinguish compounds from phrases.

c) Morphological criteria

In inflectional languages, compounds can be recognized by inflections. Only the head can be inflected, but not the non-head components (Lieber and Štekauer 2009). For instance, *post offices* (not **posts office*), *skateboards* (not **skatesboard*) and *sisters-in-law* (not **sister-in-laws*). However, there are some exceptions in which the plural maker is found

inside compounds, such as *overseas investor, parks commissioner, programs coordinator* (Plag 2003).

Morphological principle also covers the issue of linking element, a meaningless extension which occurs in the middle of a compound's two elements. (Lieber and Štekauer 2009). For example, *stellenanzeige* 'job advertisement' [German], *huntsman and statesman* [English] and *rychlovlak* 'express train' [Slovak].

The presence of inflectional morphemes and linking elements can be a useful tool to indicate compoundhood only in morphologically rich languages, but not all.

d) Syntactic criteria

Believed to be the most reliable among other criteria, the inseparability criterion claims that it is impossible to insert any other element between the constituents of a compound, while syntactic phrases can be inserted into another word. It is seemingly workable though that phrasal verbs which are considered compounds of a sort can be penetrated without losing their meaning (Lieber and Štekauer 2009).

(7) look up — *look* a word *up* in a dictionary
(8) take off — *take* your hat *off*

Bauer (1998) suggested that a compound in English does not allow modification on either of its components. For example,

(9) mortal disease — * *very* mortal disease
(10) watchmaker — * watch *skilled* maker

Such claim holds true only to the extent that the adjective in question is an attributive (sometimes called qualitative) one. Thus, modification in other cases in (11) and (12) (Lieber and Štekauer 2009), as well as coordination in (13) (Spencer 2003 cited in Lieber and Štekauer 2009) are possible.

(11) noodle salad — *instant* noodle salad
(12) fraud investigation — *serious* fraud investigation
(13) windmills — wind *and water* mills

Bauer proposed that the second component of noun+noun compounds cannot be replaced by a pro-form, e.g. **a watchmaker and a clock one*. Yet Bauer himself accepted that his criterion is not foolproof. He cited a rare case where pro-form is possible: "He wanted a riding horse, as neither of the carriage *ones* would suffice" (1998:77).

e) Semantic criterion

Name-worthiness is also suggested to be one of the criteria in distinguishing compounds and phrases (Müller et al 2015). Language users tend to coin names for concepts, activities, or objects which are common in their culture, e.g. *babysit, highchair, upside-down fridge*. In fact, it is easy to fall in the trap of determining a word formed by other means. Specialized dictionaries of any field contain countless technical terms which signify one single concept, activity or object, e.g. *monomolecular, antimatter, quartet*, etc. These entries are obviously not formed by the process of compounding. Therefore, this criteria seems problematic

In Thai, it is impossible to rely on orthographic, phonological or morphological principles to tell compounds and phrases apart (Singnoi 2005; Aroonmanakun 2007). First, there is no space to explicitly mark the word boundary. Segmenting a minimal unit of words from a sentence is confusing enough, not to mention a more complex unit like compounds. Second, Thai is an isolating language, so it does not have any inflection (neither noun declension nor verb conjugation). Third, unlike English, words in Thai do not have a predictable stress pattern to distinguish compounds from phrases. To provide an illustrative explanation, look at *bâːn lék* in the ambiguous sentence below.

(14)	*khǎw*	*miː*	*bâːn*	*lék*
	2SG	have/ possess	house	small

'He has a small house.' or 'He has a mistress'

It is not clear whether *bâːn lék* is a result of syntactic operation (the noun *bâːn* 'house' being modified by the adjective *lék* 'small') or word formation process (*bâːn lék* 'mistress').

As for semantic criterion, different languages may have different ways to refer to a single concept, and an existing concept in one culture may not

exist in others. In English, the collection of academic papers published in the context of an academic conference is called *proceedings*. Even though the same kind of concept exists in Thailand, it is called *năŋsŭ: ru:am bòtkhwa:m wíɕha:ka:n naj ka:npraɕhum sămmana:*. The string comprises eight words, namely *năŋsŭ:* 'book' + *ru:am* 'collect' + *bòtkhwa:m* 'article' + *wíɕha:ka:n* 'academic' + *naj* 'in' + *ka:npra ɕhum* 'meeting' + *sămmana:* 'seminar'. Although it is name-worthy in both cultures and expresses a comparable meaning, in Thai it is certainly recognized as a noun phrase rather than a lexical item.

Thai linguists suggest that without context where a certain synthetic compound occurs, it is difficult to distinguish it from a phrase or sentence. (Singnoi 2000; Singnoi 2005; Prasithrathsint 2010; Aroonmanakun 2015). One of the useful contextual cues is classifiers because classifiers must agree with the head, but not the non-head component of compounds. Recall the example *bâ:n lék* again and see how different contexts play a significant role in disambiguating its meaning and status.

(15)	*khăw*	*mi:*	*bâ:n*	*lék*	*nùŋ*	*lăŋ*
	2SG	have	house	small1	1	classifier
	'He has a small house.'					

(16)	*khăw*	*mi:*	*bâ:n*	*lék*	*tháŋ thî:*
	2SG	have	house	small	CONJ
	tèŋ ŋa:n	*lé́:w*			
	marry	PERF			
	'Despite a marriage, he keeps a mistress.'				

Despite a few possibilities, there can be unsolvable ambiguous cases where parsing allows both possible meanings, such as

(17)	*khon*	*khàp*	*rót*	*paj*	*tâŋ tè:*	*ɕhá:w*
	person	drive	car	go	since	morning
	'A driver left early in the morning.' or					
	'A man drove his car away early in the morning.'					

(18) *khǎw* *mâj* *kin* *khâːw* *jen*
2SG NEG eat rice cold/evening
'He does not eat dinner.' or
'He does not eat cold rice'

3. FUNDAMENTAL CONCEPTS

3.1 Internal Cohesion

Lexicalists postulate a clear-cut division between the lexicon and the syntax. The former produces members of the lexical categories, while the latter produces members of phrasal categories. That is to say, N and NP are differentiated into two discrete modules, even though both involve the concatenation of morphemes into more complex linguistic units. The idea began to develop in the late 70s and early 80s, when syntax was about how components are ordered. Linguists then mainly discussed phrase structure rules, syntactic rules, and transformations. It was not until the mid-80s, when Government and Binding Theory, Transformational Grammar and Parameters were essentially developed by Chomsky. Yet, linguists still hold onto the assumption that words are "minimal, unanalyzable units" (Bresnan and Mchombo 1995:181 cited in Lieber and Scalise 2006).

Before this view became known under a more generalized term as the Lexical Integrity Hypothesis (LIH), there were a series of proposed theories. These include Generalized Lexicalist Hypothesis (Lapointe 1980), Word Structure Autonomy Condition (Selkirk 1982) and Atomicity Thesis (Di Sciullo and Williams 1987). They may differ slightly in details, but one thing they share in common is that syntactic rules do not access or operate the internal structure of words (Giegerich 2009). The LIH argues that "the syntax neither manipulates nor has access to the internal structure of words" (Anderson 1992:84). Therefore, when it comes to compounds, lexicalists predict that syntactic operations are unable to look into morphological components of complex lexical units. Syntactic operations can only act upon such lexical units as a whole because the internal structure is tightly encapsulated (Kari 2012).

The idea that complex words' internal components cannot be manipulated by syntactic rules and that syntactic processes do not have access to the semantic components of complex words is significant in defining the notion of "internal cohesion" in the data selection in this study.

3.2 Frequency Effects

Since the research procedure in this study heavily involves corpus, it is important to introduce two kinds of frequencies, namely token frequency and type frequency. The former refers to the actual number of occurrences of a certain unit in a text(s). Such a unit can be as small as an affix (e.g. *un-*) or as big as a phrase (e.g. '*I don't know'*). The latter is defined as the number of different lexical items a certain pattern is applicable to. For example, *-s* is a major type in marking plurality because it applies to thousands of nouns. While an irregular form exemplified by a vowel-change pattern, such as *foot-feet, goose-geese, tooth-teeth* is a much smaller type, it features very few nouns (Bybee and Thompson 1997).

Type and token frequency are said to have separable effects on language users. Token frequency promotes the entrenchment of a particular unit. Repetition in usage determines how strongly it is stored in language users' memory, and how fluently it can be accessed (Gries and Ellis 2015). The routinization of high token frequency also boosts the automation of processing. Language users process frequently-used sequences faster and tend to process them all together as single chunks. (Bybee and Thompson 1997). This point is also supported by Langacker (1987:59). He explained that "When a complex structure coalesces into a unit, its subparts do not thereby cease to exist or be identifiable as substructures…Its components do become less salient, however, precisely because the speaker no longer has to attend to them individually." On the other hand, type frequency facilitates productivity and abstraction, i.e. how many different items that can be applicable to a given slot in a construction. Bybee and Thompson (1997) elaborate that the substitution of lexical items in a certain position of a construction implies a less associative link between such construction and lexical items, as well as the likelihood that such construction will extend to new items. As for

Type/Token Ratio (TTR), it is calculated by dividing the types by the number of tokens occurring in a similar piece of text. TTR is believed to be an index of lexical diversity; the higher the TTR, the greater the diversity of words (Richards 1987).

4. DATA COLLECTION AND SELECTION

Data was collected from the Thai National Corpus (TNC) which is accessible online via http://www.arts.chula.ac.th/~ling/TNCII/corp.php. The corpus features approximately 32 million words of written texts in six genres, namely fiction, academic, non-academic, newspaper, law and miscellaneous. The web interface features a search bar for one single keyword. The search results show information regarding the distribution of such keywords in every available genre. When users click on the frequency of occurrence in any genre, the concordance window will be displayed. Optionally, users may specify their search results in terms of domains, years of publication, author's age ranges, and author's genders. In addition, the web interface offers the collocation search that shows collocations within a one to four-word span on the right and left contexts of the searched keyword. The frequency of co-occurrence (Mutual Information or MI) is shown here as well.

The current research features two sets of data, namely Group A and Group B. The members of each group and their properties are described in subsections 4.1 and 4.2, respectively.

4.1 Group A

Strings of adjacent and highly productive nouns (N) and verbs (V) were extracted from the TNC. The data were divided into two groups, namely Group A and Group B. Group A represents samples of data that neatly comply with the LIH. It contains seven strings of adjacent N-V(P) with the highest collocation strength, as shown in Table 2. The MI between each N-V(P) string was checked to assure that their juxtaposition is not just a coincidence. For example, both *khon* 'person' and *paj* 'go' are very productive and frequently found in the TNC. Their token frequencies

are 215,181 and 340,828, respectively. However, the extremely low Mutual Information (MI) of -1.27 suggests a considerably rare co-occurrence between the argument *khon* and the predicate *paj*. Therefore, strings like *khon paj* were excluded from the current study.

Moreover, each selected N-V(P) string also represents different thematic roles from one another. The researchers varied the semantic relations between the arguments (N) and the predicates (VP) in order to see whether or not their thematic roles interplay with their corpus frequency. Table 2 presents the strings in Group A and their components' relation.

Table 2. The seven N-V(P) strings in Group A

semantic relations	Noun	Verb phase	MI
Agent – action	*khon* 'person'	*khàp rót* 'drive'	7.40
Instrument - action	*khrûːaŋ* 'machine'	*sák (phâː)* 'wash (clothes)'	7.00
Theme –action	*phàk* 'vegetable'	*dɔːŋ* 'pickle'	10.28
Location – action	*ráːn* 'store'	*khăːj (khɔ̌ːŋ)* 'sell (goods)'	7.44
Time – action	*wan kɤ̀ːt* 'day'	*kɤ̀ːt* 'be born'	6.31
Result – action	*rɔːj* 'mark'	*sàk* 'tattoo'	6.65
Result – Content	*ɕòt măːj* 'letter'	*rák* 'love'	5.99

4.2 Group B

Like Group A, all the strings in Group B are naturally occurring data taken from the TNC. They differ from the former group as they represent the violation of the Lexical Integrity Hypothesis (LIH). Group B consists of the aforementioned 7 strings of N-V(P), but they are contaminated with intervention, modification, alteration or coordination within their 5-word span.

(19) and (20) illustrate the intervention of syntactic units which are found between or together with the word strings in question. They are mainly aspect and modality. Such as, the intentive marker *ɕàʔ* and progressive aspect *jùː*.

(19) *wan ɕàʔ kɤ̀ːt*
day FUT birth

(20) *khrûːaŋ sák phâː jùː*
machine wash clothes PROG

(21) and (22) exemplify the modification to one of the constituents in the word strings. They include modifiers or intensifiers. Such as, *jàj* that modifies the noun *ráːn*, or *rew* that can either modify the noun *rót* or intensify the predicate *khàp.*

(21) *ráːn jàj khǎːj khɔ̌ːŋ*
shop big sell things

(22) *khon khàp rót rew*
person drive car fast

(23) – (25) show the alteration of one of the constituents to another word from the same semantic domain. Such as, *khàp* ‘drive’ was altered to *khìː* and *khàp khìː*, *dɔːŋ* ‘ferment’ to *màk* and *màk dɔːŋ,* as well as *khǎːj* ‘sell’ to *ɕam nàːj.*

(23) a. *khon khìː rót*
person ride car
b. *khon khàp khìː rót*
person drive/ride car

(24) a. *phàk màk*
vegetable ferment
b. *phàk màk dɔːŋ*
vegetable ferment/ pickle

(25) *rá:n* *ɕam nà:j* *sĭn khá:*
store distribute goods

(26) and (27) represent the coordination. One of the constituents in the word strings is joined with another word of equivalent status. In Thai, *lɛ́ʔ* 'and' is the conjunction that occurs in between. Such as, the predicate *sák* coordinates with the predicate *ʔòp,* the noun *phàk* with the noun *khĭŋ.*

(26) *khrû:aŋ* *sák* *lɛ́ʔ* *ʔòp* *phâ:*
machine wash and dry clothes

(27) *phàk* *lɛ́ʔ* *khĭŋ* *dɔ:ŋ*
vegetable and ginger pickle

According to the LIH that assumes a dichotomy, the process of intervention, modification, alteration or coordination to one of its constituents may deprive the loosely tied strings in Group B of their lexical status.

5. ANALYSIS OF CORPUS FREQUENCY

5.1 Procedure

Every N-V(P) string in Group A and Group B that was taken from the Thai National Corpus (TNC) was measured for its type and token frequency. The researchers wrote code to electronically retrieve all the strings in question, as well as their frequency of occurrence in the corpus.

The corpus frequency of Group A members are presented individually, while Group B members are presented collectively in four subgroups, namely intervention, modification, alteration and coordination. The fact that Group B members' corpus frequency is clustered is because the number of each subgroup is not equal. It depends entirely on what is available in the corpus. Therefore, classifying and presenting them according to their syntactic or semantic operations seems to make more sense.

The overall token frequency of each N-V(P) string was equated to 100%, then the token frequencies of such a particular string in Group A and the subgroups in Group B were separately calculated in relation to the total percentage. To exemplify and elaborate the calculation method, there are 2,428 strings beginning with '*person*' making up 100%. 1,948 out of 2,428 strings (80.23%) fit into Group A; 36 out of 2,428 strings (1.48%) fit into the intervention subgroup; 216 out of 2,428 strings (8.90%) fit into the modification subgroup; and 228 out of 2,428 strings (9.39%) fit into the alteration subgroup.

The type frequency, on the other hand, requires the arithmetic count of classes. The researchers counted each time a different string appeared in the corpus. For example, *khon khàp rót*, *khon khìː rót*, and *khon khàpkhìː rót* count up to three types.

5.2 Hypotheses

Hypothesis #1: Group A, whose members' internal cohesion is tighter than members of Group B, should display higher frequency of occurrence in the TNC.

Hypothesis #2: Among all Group B members, modification should be the subgroup that outnumbers the others in terms of type and token frequency.

5.3 Findings

Table 3 sorts the data according to their percentages, token frequencies in parentheses, and type frequencies in square brackets. The percentages reflect how likely a group or a subgroup can be found, when compared to the total number of members with a similar N. The token frequencies are the actual count of a string or a subgroup on the TNC. The type frequencies indicate the number of different strings.

The analysis of corpus frequency shows that Group A apparently exhibits a much higher frequency of occurrence than Group B. In other words, Group A is more commonly found in naturally occurring language than its counterpart.

There are altogether 86 N-V(P) different types. Seven tightly integrated ones belong to Group A, and the rest belong to Group B. In general, the number of types in Group B's members are low

Another point worth consideration is the TTR. The results show that Group A's TTR is very low compared to Group B's. The lower the TTR, the higher associative each internal constituent is to one another.

Table 3. Group A and B's frequency of occurrence on the TNC

N-V(P) strings	Group A	Group B			
		Interven-tion	Modifica-tion	Altera-tion	Coor-dination
khon + khàp rót 'person' + 'drive'	80.23% (1948) [1]	1.48% (36) [7]	8.90% (216) [18]	9.39% (228) [2]	n/a
khrûːaŋ + sák phâː 'machine' + 'wash clothes'	96.39% (561) [1]	0.86% (5) [2]	2.06% (12) [1]	n/a	0.69% (4) [1]
phàk + dɔːŋ 'vegetable' + 'pickle'	74.50% (336) [1]	0.67% (3) [1]	23.28% (105) [9]	0.67% (3) [2]	0.89% (4) [2]
ráːn + khăːj khɔ̌ːŋ 'store' + 'sell things'	50.07% (1129) [1]	0.62% (14) [2]	45.32% (1022) [16]	3.99% (90) [2]	n/a
wan + kɤ̀ːt 'day' + 'be born'	96.50% (4109) [1]	0.28% (12) [1]	0.49% (17) [2]	n/a	2.72% (116) [2]
rɔːj + sàk 'mark' + 'tattoo'	92.59% (700) [1]	n/a	4.23% (32) [3]	3.17% (24) [1]	n/a
ɛ̀ot măːj + rák 'letter' + 'love'	83.93% (94) [1]	16.07% (18) [5]	n/a	n/a	n/a

Table 3 provides the information that strongly supports hypothesis #1; it is obvious that Group A, which strictly complies with the LIH and displays tight internal cohesion, outnumbers Group B on the TNC. Among the members of Group A, the string *wan + kɤ̀ːt* exhibits the highest percentage of occurrence (96.50%) and with token frequencies (4,109).

Next, hypothesis #2 focuses on the type and token frequency within Group B's subgroups. The researchers' prediction is virtually correct. The

modification subgroup's type frequency ranks the highest, but its overall token frequency nearly does. The reason that the researchers initially expected to see the highest token frequency in the modification subgroup is because this kind of linguistic operation involves both syntactic and semantic manipulation. That is to say, modification deals with combining elements (the modified and the modifier), and in the meantime, the modifier adds more descriptive information to the modified antecedent. Therefore, it was predicted to occur more frequently than the other subgroups. However, the researchers' prediction only holds true for type, not for token. If the strings beginning with 'person' and 'day' are carefully examined, their alteration and coordination subgroups' token frequencies in respective order are higher.

6. NATIVE SPEAKERS' INTUITIVE JUDGMENT

6.1 Procedure

The participants were 30 native Thai language users who voluntarily took part in the survey. They are 18 males and 12 females; six persons aged between 20-25 y/o, fourteen persons 26-30 y/o, eight persons 31-35 y/o, two persons 36-40 y/o. They have at least bachelor's degrees from various fields of study, but none of them did Linguistics Science or Language and Literature programmes as the researchers were afraid that their previous language study might bias their judgments in the experiment.

The survey forms were distributed online. Each survey form contained 28 strings, seven of which are Group A, another seven are randomized from Group B, and the rest are fillers. The fillers, i.e. items that are not related to the research questions, were added to reduce participants' conscious awareness of a certain topic being tested. In this experiment, the fillers feature multi-syllable words, e.g. *naːlíkaː* 'clock', *málákɔ́ː* 'papaya', *ɛ̀akkràjaːn* 'bicycle' and proper names, e.g. *hɔ̂ːkkajdoː* 'Hokkaido', *míthùnaːjon* 'June', *pìkàtsoː* 'Picasso', *roːlèk* 'Rolex'. The order of strings on the survey forms was electronically randomized.

They were told to use their native speaker intuition to rate how autonomous the given word strings are on a 3-point Likert scale. When

any string is rated '*I agree that it is a single-whole*.', '*I don't agree that it is a single-whole*' and '*I am not sure*', it will be given the scores of 1, -1 and 0, respectively. Please note that the researchers avoided using the word 'compound' in the survey forms because the definition and boundary of compounds are still controversial.

6.2 Hypotheses

Hypothesis #3: The seven strings in Group A should receive a score of 1 or very close to 1 because of their strong internal cohesion, while those in Group B should receive a score lower than 0 because of the lack of internal cohesion;

Hypothesis #4: There should be a positive correlation between the percentage of occurrence and the score the participants rated strings of each group;

Hypothesis #5: The participants may give the score of 0 or 1 (rather than -1) to any member of the subgroup whose type and token ratio (henceforth, TTR) is low;

Hypothesis #6: The participants are likely to give lower scores to Group B members, if its Group A counterpart of a similar N exhibits an exceptionally high percentage of occurrence, and vice versa.

6.3 Findings

In this section, the survey results will be presented from two aspects. One (Figure 1) shows the average scores given by the participants. The other (Figure 2) are the average scores given to each of the seven N-V(P) strings.

Figure 1 illustrates the average scores given to Group A and Group B (axis Y) by all 30 participants (axis X). In general, it is obvious that Group A, which strictly complies with the LIH and displays tight internal cohesion, was given higher average scores than Group B. Two thirds of the participants gave the maximum score of 1 to Group A. Only ten participants rated the strings in Group A below 1, but their scores were still considered high (0.71 to 0.85 points). On the contrary, the scores given to Group B were dramatically lower (-0.71 to 0.14 points). There

were four participants giving an average score of 0 or above to Group B. The average scores below -0.5 were from 13 participants, and scores between -0.5 to 0 were from the other 13 participants.

Figure 1. The average scores given by the 30 participants

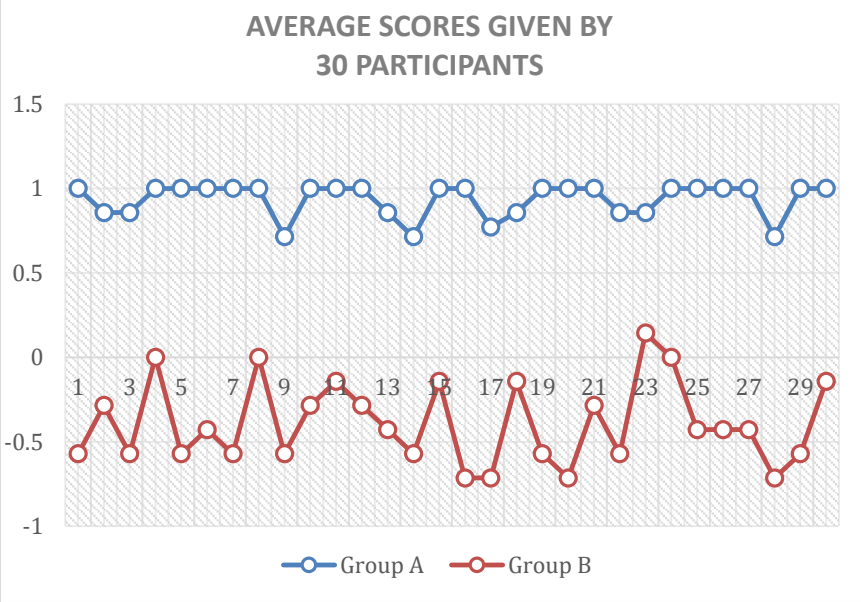
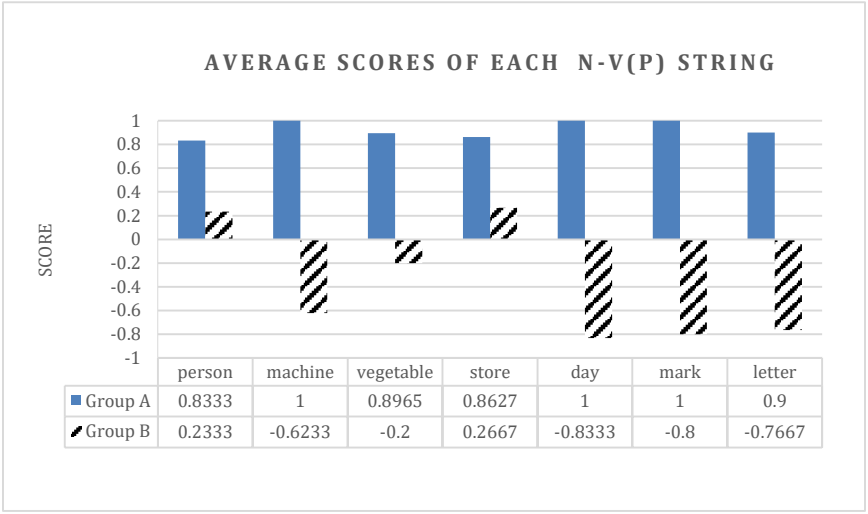


Figure 2 displays the average scores given to each N-V(P) string. If any certain string was rated the maximum score of 1, this means the participants were prone to consider it as a single-whole. In contrast, the lower the score means it is less likely to be understood as a single-whole.

Like the results shown in Figure 1, the strings in Group A received higher average scores than the other group. Three out of seven strings were given full marks. Although the other four in Group A did not receive the score of 1, their scores (0.9, 0.89, 0.86 and 0.83) were close to 1. Overall, it is obvious that native Thai participants agreed that Group A members are single-whole units.

As for Group B, the strings beginning with ‘store’ and ‘person’ are the only two whose scores are above 0 (0.27 and 0.23), while the strings beginning with ‘mark’ and ‘day’ are the bottom lowest and close to -1.

Figure 2. The average scores of each N-V(P) string



	person	machine	vegetable	store	day	mark	letter
Group A	0.8333	1	0.8965	0.8627	1	1	0.9
Group B	0.2333	-0.6233	-0.2	0.2667	-0.8333	-0.8	-0.7667

Hypothesis #3 is interested in native speakers' perception when they were shown the tightly integrated strings in Group A and the loosely integrated ones in Group B. The results in Figure 1 and Figure 2 partially support the researchers' prediction. It is partially accurate that members of Group A achieved the maximum score of 1 or very close to 1, but not all strings in Group B received negative scores. The exception includes the strings beginning with 'person' and 'store' which were given the scores of 0.23 and 0.26.

Hypothesis #4 concerns a positive correlation between the percentage of occurrence and the scores rated by the Thai participants. The results, as presented in Figure 2, strongly support this. Group A members that have significantly high percentages of occurrence were also understood by the participants as single-whole units. Likewise, the positive correlation can be observed in Group B. The strings beginning with 'store', 'person' and 'vegetable' whose percentages of occurrence is higher than those of 'machine', 'day', 'mark' and 'letter' also achieved slightly higher scores from the native speakers' judgment test.

Hypothesis #5 considers the TTR and the score given by the participants. It is clear that each string in Group A has only one type, so its TTR is exceedingly low. All Group A strings were rated 1 or very close to 1, which means the native speakers in this study agreed unanimously that those strings are autonomous enough to be recognized as single-whole units. As for Group B, the strings beginning with 'store' 'person' and 'vegetable' have a relatively lower TTR, and they actually earned satisfying intuition scores (0.26, 0.23 and -0.2) which are about the '*I am not sure*' level. However, the TTR of the string beginning with 'day' seems to contradict the prediction. They exhibit a low TTR, but their native speakers' judgment scores turned out to be the lowest. Perhaps the factor that determines how likely a certain string would be understood as a single-whole unit is not limited to the TTR alone. The researchers wonder whether the percentage of occurrence of its Group A counterpart may play a role too, and this point is elaborated in the next paragraph.

The last hypothesis aims to see if the high percentage of occurrence of a certain Group A string would affect the scores given to its Group B counterparts beginning with a similar N. The strings beginning with 'day', 'machine', 'mark' and 'letter' in Group A exhibit the top four highest percentages of occurrence, as follows: 96.50%, 96.39%, 92.59% and 83.93%. The researchers noticed at least two things in common amon these four. One, their Group B counterparts were rated in the bottom four lowest scores from the participants (-0.62, -0.83, -0.8, and -0.76). Two, the intuition scores given to these strings in Group A were exceptionally high (1, 1, 1 and 0.9). Next are the strings beginning with 'person', 'vegetable' and 'store' in Group A. Their percentages of occurrence (80.23%, 74.50% and 50.07%) are slightly lower than the first four words mentioned earlier. They likewise share two similarities. One, their Group B counterparts ranked as the top three highest in the native speakers' judgment test, as follows: 0.23, -0.2 and 0.26. Two, these strings in Group A received a marginally lower point (0.83, 0.89 and 0.86). The results suggest a predictable reverse correlation. If the percentage of occurrence and the intuitive judgment score in Group A is high, then its Group B counterpart gets the opposite, i.e. lower intuitive judgment scores.

7. CONCLUSION AND DISCUSSION

This section provides a synopsis of the research goal, methods, interpretation of results, limitations and suggestions for future research.

The current study addresses the controversial question about the status of synthetic compounds in Thai and whether it is possible to draw a distinction between them and syntactic phrases. The idea of the Lexical Integrity Hypothesis (LIH) was adopted as a criterion to elicit the data from the Thai National Corpus (TNC). Lexicalists claim that internal cohesion is able to tear apart lexical and syntactic units, so the researchers would like to test if it also works with the Thai language. The data were electronically extracted and separated into two sets, namely Group A and Group B. The former neatly complies with the LIH, while the latter features some sort of linguistic operations acting upon one of their components.

The study was conducted in two phases starting off with the analysis of corpus. Both groups' percentages of occurrence, type frequencies, token frequencies and Type/Token Ration (TTR) were measured and compared. The comparison confirms hypothesis #1 that the degree of internal cohesion is positively associated with frequency. Yet the researchers hesitate in stating that strong internal cohesion might not be the only factor that accounts for higher frequency of occurrence. The reason is the word strings in Group A are less complex and thus shorter than those in Group B, so it can be the case that the word length negatively correlates with frequency (Grzybek 2003 cited in Strauss et al 2007). Aroonmanakun's (2005) recent computational research also supports this point. He mentions the correlation between word length and the likelihood the parser will consider a word sequence in question a compound. If word sequences are longer than four, the parsers tend to presume and analyse them as phrases.

Hypothesis #2 predicted that the kind of linguistic operation which most commonly looks into the components of Group B is modification. The reason the highest token frequency was expected to be in the modification subgroup is because it involves both syntactic and semantic manipulation. In other words, modification deals with combining elements (the modified and the modifier), and in the meantime, the

modifier adds more descriptive information to the modified antecedent. Therefore, it was predicted to occur more frequently than the other subgroups. However, the token frequencies of the strings beginning with 'person' and 'day' in the alteration and coordination subgroups are greater than those in the modification subgroup (see Table 3).

Here are the explanations for the unexpected results. In the alteration subgroup of the strings beginning with 'person', two other predicates *khìː*'ride' and *khàp khìː* 'drive/ride' from a similar semantic domain are found. The predicates *khìː* and *khàp khìː* refer to an act of throwing one's leg over the middle of the vehicles, e.g. bicycle, motorcycle, to mount it. The real world makes these instances possible because these small vehicles are ubiquitous in Thailand, where the language is spoken, so language community members or Thai people are used to them.

As for the strings beginning with 'day', the coordination variants include *wan* 'day' + *duːan* 'month'+ *kə̀ːt* 'be born' (day and month of birth) and *wan* 'day'+ *duːan* 'month' + *piː* 'year' + *kə̀ːt* 'be born' (DMY of birth). These are common and pervasive in many languages though.

In conjunction with the analysis of corpus frequency, the survey of native Thais' judgment is deemed crucial to add as another phase to the study because the researchers noticed something counter-intuitive in Group B, e.g. the strings beginning with 'store' in the modification subgroup cover a large number of members, despite the violation to the LIH. Therefore, four more hypotheses were set up to investigate.

The answer to hypothesis #3 supports the researchers' prediction that Group A is likely to receive scores of 1 or close to 1. Unexpectedly, the strings beginning with 'person' and 'store' in Group B also received scores above 0 too (see Figure 2). The findings surely are surprising, since Group B members violate the LIH and their lessening internal cohesion should negatively affect native speakers' judgment. The researchers found at least three similarities shared between the strings beginning with 'person' and 'store' (see Table 1). One, their Group A counterparts' scores (0.83 and 0.86) rank as the bottom two, when compared with the five other strings in Group A. Two, their modification and alteration subgroups' token frequencies are high (216 and 228 for 'person', 1022 and 90 for 'store'). Three, in both subgroups the TTR is noticeably low as well (18/216, 2/228, 16/1022, and 2/90). Low TTR can be interpreted that a

certain number of words are highly associative with a particular construction. The lower lexical variation allows language users to get a fix on what accounts for possible constituents and thus gradually forming mental entrenchment.

The native speakers' judgment results also reveal the correlation between the higher percentages of occurrence, low TTR, and the degree of lexical autonomy, as shown in hypotheses #4 and #5. Additionally, hypothesis #6 reveals that when certain strings in Group A have considerably high percentages of occurrence, it affects the judgment scores given to their Group B counterparts beginning with similar words. The researchers assumed that language users develop a firm hold of highly co-occurring word sequences. Once they see any unfamiliar variant of such word sequences, they probably do not find it fits well with their language repertoire and tend not to recognize it as a single-whole unit. It can be said that every encounter of any linguistic unit can strengthen the degree of entrenchment (Schmid 2007, 2010). When language users were asked to rate how autonomous the given word strings were, they are prone to state that the frequently occurring ones with slightly fixed word sequences are single-whole units.

What we learned from both phases of the current research is that strong internal cohesion contributes to the quality of being compound, but it is not the only property. In fact, it is the interplay among the frequency of occurrence, length, complexity and language users' familiarity (Caldwell-Harris et al. 2012). Group A denotes the referents which are much more generic. For example, *khon khàp rót* 'driver' does not refer to any specific person who is driving a vehicle, but it refers to an occupation of operating a motor vehicle. If a doctor, a teacher or a singer is seen driving a car, s/he will not be referred to as *khon khàp rót*. The genericity and recursion are properties that sweep over all the strings in Group A.

The significant role of recursion in fortifying mental entrenchment of linguistic behaviours was described under the term "ritualization" analogous to non-human behaviours. Haiman (1994) points out that oftentimes complicated actions contain a series of (non-)compositional elements. When the actions are performed habitually, those elements are stored little by little in an agent's mind as routines. Eventually, agents overlook their individual meaning and execute them as a pre-packaged

unit. In the same vein, recursion drives the word strings in the current study to be understood more or less as single-whole units. This explains why the native participants tend to consider Group A members and some of Group B with high corpus frequency more autonomous than others.

As for the main question, which is also the paper title, *Is there a dichotomy between synthetic compounds and phrases in Thai?*, the results indicate the lexical-syntactic continuum, instead of two discrete realms. In other words, synthetic compounds exhibit a gradient degree between members of lexical category on one end and syntactic category on the other. Therefore, it is yet impossible to draw a sharp distinction.

The researchers hope that the current study contributes to the linguistic field by exploring the status of synthetic compounds from a mixed methods point of view. Yet, the limitation is that the word strings in this study were not presented with contexts. Doing so may or may not yield a different result. In this instance, the researchers also would like to encourage further research on how humans process synthetic compounds. 'Processing' in the sense that covers both storage and retrieval strategies, e.g. Do we store and retrieve a minimal unanalysable unit individually to save cognitive loads, and then combine each unit to form a compound? Or do we store and retrieve them in a big chunk? Moreover, the merit of tackling complicated issues like this is that it potentially brings about collaboration among researchers from diverse disciplines.

REFERENCES

Anderson, Stephen R. 1992. *A-morphous Morphology*. Cambridge: Cambridge University Press.

Aroonmanakun, Wirote. 2005. Collocation Extract: A Tool for Extracting Collocation. *Journal of English Studies* 2:28-39.

Aroonmanakun, Wirote. 2007. Thoughts on word and sentence segmentation in Thai. Paper presented at the 7th International Symposium on Natural Language Processing (SNLP-2017), December 13-15, 2007, Pattaya, Thailand.

Aroonmanakun, Wirote. 2015. The use of context vectors in determining Thai compounds. *Linguistic Research*. 32(1):1-20.

Bauer, Laurie. 1998. When is a sequence of noun + noun a compound in English? *English Language and Linguistics*. 2:65-86.

Bauer, Laurie. 2003. *Introducing Linguistic Morphology*. Washington DC: Georgetown University Press.

Bauer, Laurie. 2009. Typology of compounds. In *The Oxford handbook of compounding*, eds. by Rochelle Lieber and Pavol Štekauer, pp.343-356. Oxford: Oxford University Press.

Bybee, Joan and Sandra Thompson. 1997. Three frequency effects. Paper presented at the 23rd Annual Meeting of the Berkeley Linguistics Society (23 BLS), February 14-27, 1997, Berkeley, the United States.

Caldwell-Harris, Catherine, Jonathan Berant and Shimon Edelman. 2012. Measuring mental entrenchment of phrases with perceptual identification, familiarity ratings, and corpus frequency statistics, In *Frequency Effects in Language Representation (Vol 1): Statistical Effects in Learnability, Processing and Change*, eds. by Dagmar Divjak and Stefan Thomas Gries, pp.165-194. The Hague: De Gruyter Mouton.

Cinque, Guglielmo. 1993. The null theory of phrase and compound stress. *Linguistic Inquiry*. 24(2):239-297.

Di Sciullo, Anna Maria and Edwin Williams, 1987. *On the Definition of Word*. Cambridge: MIT Press.

Diller, Anthony. 1992. Thai. In *International Encyclopaedia of Linguistics (Vol. 4)*, ed. by William Bright, pp.149-156. New York: Oxford University Press.

Dressler, Wolfgang U. 2006. Compound types. In *The Representation and Processing of Compound Words*, eds. by Gary Libben and Gonia Jarema, pp.23-44. Norfolk: Oxford University Press.

Giegerich, Heinz. 2009. Compounding and Lexicalism. In *The Oxford Handbook of Compounding*, eds. by Rochelle Lieber and Pavol Štekauer, pp.178-200. Oxford: Oxford University Press.

Gries, Stefan and Nick C. Ellis. 2015. Statistical measures for usage-based linguistics. *Language Learning*. 65(1):228-255.

Haiman, John. 1994. Ritualization and the development of language. In *Perspectives on Grammaticalization*, ed. by William Pagliuca, pp.3-28. Amsterdam: John Benjamins.

Jackson, Howard and Etienne Ze Amvela. 2000. *Words, Meaning and Vocabulary: An Introduction to Modern English Lexicology*. London: Athenaeum Press.

Kari, Etherbert Emmanuel. 2012. Endocliticization and the lexical integrity hypothesis: insights from Degema. In *Pronouns and Clitics in Early Language*, eds. by María Pilar Larrañaga and Pedro Guijarro-Fuentes, pp.257-282. Göttingen: Mouton de Gruyter.

Langacker, Ronald. 1987. *Foundations of Cognitive Grammar*. Stanford: Stanford University Press.

Lapointe, Steven. 1980. *A Theory of Grammatical Agreement*. University of Massachusetts, Amherst. Doctoral dissertation.

Lieber, Rochelle and Pavol Štekauer. 2009. Introduction: Status and definition of compounding. In *The Oxford Handbook of Compounding*, eds. by Rochelle Lieber and Pavol Štekauer, pp.18-30. Oxford: Oxford University Press.

Lieber, Rochelle and Sergio Scalise. 2006. The Lexical Integrity Hypothesis in a new theoretical universe. *Lingue e linguaggio* 1:7-32.

Müller, Peter O., Ingeborg Ohnheiser, Susan Olsen and Franz Rainer. 2015. *Word-Formation: An International Handbook of the Languages of Europe*. Berlin: de Gruyter Mouton.

Olsen, Susan. 2000. Compounding and stress in English: A closer look at the boundary between morphology and syntax. *Linguistische Berichte*. 181:55-69.

Plag, Ingo. 2003. *Word-Formation in English*. Cambridge: Cambridge University Press.

Prasithrathsint, Amara. 2010. Lexicalization of syntactic constructions in Thai. Paper presented at the 20th Anniversary Meeting of the Southeast Asian Linguistics Society (SEALS 20), June 10-11, 2010, Zurich, Switzerland.

Richards, Johnson. 1987. Type/token ratios: what do they really tell us? *Journal of Child Language*. 14(2):201-209.

Schmid, Hans-Jörg. 2007. Entrenchment, salience, and basic levels. In *The Oxford Handbook of Cognitive Linguistics*, eds. by Dirk Geeraerts and Hubert Cuyckens, pp.117-138. New York: Oxford University Press.

Schmid, Hans-Jörg. 2010. Does frequency in text instantiate entrenchment in the cognitive system? In *Quantitative Methods in Cognitive Semantics: Corpus-driven Approaches*, eds. by Dylan Glynn and Kerstin Fischer, pp.101-133. Berlin: de Gruyter Mouton.

Selkirk, Elisabeth O. 1982. *The Syntax of Words*. Cambridge: MIT Press.

Singnoi, Unchalee. 2000. *Nominal Constructions in Thai*. University of Oregon. Doctoral Dissertation.

Singnoi, Unchalee. 2005. *Compound Nouns: Sciences and Arts in Thai Word Formation*. Bangkok: Chulalongkorn University Press. [in Thai]

Strauss, Udo, Peter Grzybek and Gabriel Altmann. 2007. Word length and word frequency. In *Contributions to the Science of Text and Language*, ed. by Peter Grzybek, pp.277–294. Dordrecht: Springer.

Snyder, William. 2016. Compound word formation. In *The Oxford Handbook of Developmental Linguistics*, eds. by Jeffrey L. Lidz, William Snyder and Joe Pater, pp.89-110. Oxford: Oxford University Press.

[Received 7 March 2017; revised 15 October 2017; accepted 23 November 2017]

Kamolchanok Hongthong
Department of Linguistics
Faculty of Arts
Chulalongkorn University
h.kamolchanok@gmail.com

Kingkarn Thepkanjana, Ph.D
Professor of Linguistics
Department of Linguistics
Faculty of Arts
Chulalongkorn University
thepkanjana@gmail.com, kingkarn.t@chula.ac.th

Wirote Aroonmanakun
Associate Professor of Linguistics
Department of Linguistics
Faculty of Arts
Chulalongkorn University
awirote@chula.ac.th

泰文的合成複合詞與短語中是否具有對分

Kamolchanok Hongthong
Kingkarn Thepkanjana
Wirote Aroonmanakun
朱拉隆功大學

本研究探討泰文中的綜合複合詞與短語之間的結構歧義，由於兩者之間在語言中顯示非常相同。該研究在泰國泰語語料庫中（Thai National Corpus）抽取產出性名詞與動詞（及謂語）為樣本。抽取來的詞語將分為兩組。A 組為具有七層關係且具有最強搭配的名動詞語（及謂語），各展示不同語義關係。而 B 組中的詞語是與 A 組具有相同結構的名動詞語（及謂語），而此組詞語的特點為僅具有五個詞彙廣度的干預、協作、修辭以及變動。為了分類詞彙和句法單位，該研究以詞彙完整性假設進行建立內部結合。此外詞彙引用機率及本地人的感受也將考慮為分析條件之中。研究顯示 A 組詞語出現機率多於 B 組。本地人傾向把 A 組詞語了解為一個單位詞語而 B 組詞語如何將歸納為怎麼樣的單位仍在探討中。總之泰語中的合成複合詞是否將歸類為詞彙或句法仍具有模糊的境界。

關鍵字：合成複合詞、泰文、內部結構、語料庫機率、固守