An Information Structure Annotation of Thai Narrative Fiction

Glenn Slayden
University of Washington
gslayden@u.washington.edu

Abstract

Information structure is concerned with the management and organization of elaboration in discourse. Speakers and writers have at their disposal a variety of techniques for controlling the presuppositions that they wish to maintain and the new relationships that they wish to assert about them. These techniques vary cross-linguistically, encompassing, *inter alia*, prosody, morphology, and syntactic structure. Because prosody is a key marker of information structure in many languages, the information structure of written texts is less-studied than that of spoken discourse. The study of written texts offers some research benefits, however, such as ease of annotation and the disentanglement—from superimposed phonological phenomena—of those markers that do manifest in writing. Additionally, the increased linguistic load placed on syntactic marking in written texts may more readily elicit the full spectrum of syntactic use cases. The present study explores the information structure of written Thai, guided by the annotation of a Thai short story.

1 Introduction

The study of linguistic information structure is subtle, owing to complex interactions between communicative goals and the available marking strategies, many of which are overloaded with competing linguistic functions. Until recently, the field of study has also been at the disadvantage of lacking research consensus and uniform terminology (Paggio 2009: 138). An important catalyst for resolving this impasse was Lambrecht's (1996) exhaustive work, which brought, within the tradition of theoretical linguistics, a welcome degree of rigor. At the broadest level of research consensus are the dual qualia of TOPIC—the mental representation of a discourse referent that a proposition is "about" (*ibid.*: 127), and FOCUS—the unpredictable portion of an assertion that represents new information (*ibid.*: 213).

Despite this progress, relatively little of the information structure literature embraces quantitative methods and modern computational capabilities. Naturally, some theoretical study should precede corpus investigation, lest the latter have a haphazard character, but after several decades of proposals it seems appropriate to put promising ideas to test. Even intuitively compelling proposals can benefit from the exercise of corpus study.

One ambitious and worthwhile quantitative methodology is to adopt a hypothesis wholesale and integrate it into a computational grammar system. In this approach, the hypothesis is falsified when the overall system parses or generates incorrectly. Alternative hypotheses can be evaluated by comparing the detailed performance metrics that such systems produce. Within the Head-driven Phrase Structure Grammar (HPSG) formalism (Pollard and Sag 1994), authors pursuing sketches of computational information structure representations include Engdahl and Vallduví (1996) (Catalan), Wilcock (2005), Jokinen (2005) (Finnish), and Paggio (2009) (Dutch).

There is also need for—and value in—more preliminary studies which simply aim to document, characterize, and establish the cross-linguistic sufficiency of a theoretical model of linguistic information structure for naturally occurring discourse, and this is the objective of this study. By annotating a short written text in the Thai language for information structure, I hope to document the occurrence, utility, and sufficiency of the TOPIC and FOCUS qualia for Thai, and to make some quantitative observations about their surface manifestation (marking) in a simple written narrative.

2 Prior Work

Few English-language results have been published in Thai information structure. A literature survey (Burusphat 2002) confirms only a handful of relevant papers in the field of discourse studies, many of them in Thai.

Meepoe-Baron (Meepoe 1997, Meepoe-Baron 1998) has a number of studies on how Thai interlocutors tie their utterances together by selecting mirroring or complementing lexemes and syntactic structures. While such a

shared common ground—not of discourse content, but of discourse process—may introduce noise into the information structure signal by further competing for its scarce marking resources, I do not control for any such effects in this study. Although Meepoe-Baron's studies involve spontaneous conversation, a similar effect in composed text might originate from poetic or stylistic considerations limiting or dictating the expressive means available for information structure marking.

Meepoe-Baron's research used a discourse corpus collected by Iwasaki, himself the co-author of a major reference grammar of Thai which presents a brief overview of discourse phenomena in Thai, including the *TOPIC*-COMMENT syntactic construction, TOPIC marking, referent tracking, resumptive pronouns (LEFT-DISLOCATION), expletive use, and non-canonical constituent order (Iwasaki and Ingkaphirom 2005: 359-376).

Warotamasikkhadit (1979) is a transformational-grammar account of LEFT- and RIGHT-DISLOCATION and *TOPIC*-COMMENT structures. This pioneering work in the study of Thai information structure strives to disentangle syntactic *TOPIC* from information structure TOPIC.

Some years later, the same author presented examples of fronting and backing constructions in Thai, contrasting them with canonical forms (Warotamasikkhadit 1997). His observations on the role of demonstratives in establishing the rightwards boundary in *TOPIC*-COMMENT structures are elaborated by Singnoi, who provides a good analysis of the subtle relationship between the demonstrative determiners and demonstrative pronouns[1] and their discourse function (Singnoi 2004). In particular, she posits five pragmatico-discourse functions for the demonstratives, which span a range from information status to information structure: drawing attention, switching attention, tracking entities, managing discourse referents, and reintroducing topics (Singnoi 2004: 651).

Both of these authors posit interactions between non-canonical structures and definiteness. The former asserts that fronted topics must be definite, while suggesting that Thai demonstrative determiners do not inherently mark an NP for definiteness (Warotamasikkhadit 1997: 304). Singnoi concurs, citing Christopher Lyons' analysis of *such* in English, as well as separate work by Marianne Mithun and Talmy Givón to note that additional pragmatic context is required for fully establishing the definiteness of NPs in Thai (Singnoi 2004: 650). For the purposes of this study, her most pertinent remarks concern the characterization of rigidly right-bounded noun phrases (Singnoi 2004: 650), work that I summarize in Section 4.5.1.5.

More recently, Singnoi presents results of corpus study of the occurrence of six non-canonical structures (existential, fronting, cleft, passive, and left- and right-dislocation), cross-tabulated against eight discourse functions (marking new or given information, or the discourse topic; substituting agency; marking contrast; shifting topic; reactivating referents; and marking unknown or unimportant agents) (Singnoi 2007). She finds that the existential construction tends to be used to present new information, while FRONTING is used to mark given information or refer to the discourse TOPIC, and LEFT-DISLOCATION is used to reactivate past discourse referents.

3 Source Text

This project will consist of the annotation of the short story *Building Sandpiles* (ก่อกองทราย) (Thanya 1955) written in 1955 by the Thai poet, Thanya Sangkapanthanon (ธัญญา สังขพันธานนท์, 1926-), writing under the penname Paitoon Thanya (ไพฑูรย์ ธัญญา). This text consists of approximately 2,357 words in 197 sentences. A complete copy of the source text is included as Appendices A (Thai) and B (English).

The central character of the fictional narrative is an unnamed small girl who is playing in the sand by herself on a river bank while a group of boisterous boys swim nearby. Playing alone, she forms structures from the wet sand of the beach, but waves repeatedly wash up to destroy them. At first she is frustrated, but after one of the boys cautiously joins her, the two find renewed optimism despite the adversity. The allegorical meditation is well known for its rustic portrayal of uncomplicated innocence, and its southern Thai folkloric sensibility.

[1] The sets of Thai demonstrative determiners and demonstrative pronouns are often confused because their orthographies and pronunciations parallel closely along two dimensions: first, both sets of monosyllables are of the form /n-*v*-n/, where *v* is a high vowel whose backness is proportional to distal significance; and second, the pronouns have falling tone while the determiners have high tone.

4 Annotation Methodology

The annotation work presented here comprises: full glossing and a new English translation; Thai-English sentence alignment and sense selection; identification of dropped subjects, dropped objects, and pronoun occurrence; coarse syntactic constituent analysis; and judged assignment of TOPIC and FOCUS qualia. The guidelines for each of these annotation categories will be discussed in the following sub-sections.

During the annotation procedure, grammatical judgments were elicited from a native speaker as necessary to resolve points of ambiguity. For this project, the native Thai speaker is a 42-year-old, Seattle-resident, non-linguist, bilingual male 'A', who lived ten childhood years in Thailand, and maintains fluency through exclusive Thai speaking with Thai-resident family members and friends, and by frequent travel to Thailand.

4.1 Gloss and Translation

An English translation of the text by was available in bitext format (Glass n.d.) and used for reference, but as with all excellent literary translations, it did not have the specific goal of facilitating linguistic study by, for example, preserving unusual syntactic constructions. With guidance from—and respect to—this source text, it was decided to develop a new full gloss and technical English translation as part of this project. In the new translation, constituent structures of putative information structural interest, such as FRONTING, BACKING, LEFT-DISLOCATION, and RIGHT-DISLOCATION, are paralleled in English, where possible. Where it was not possible to fully capture these structures (as with verb serialization and certain *TOPIC*-COMMENT patterns, for example), the interlinear glossed text (IGT) lines in the detailed annotation provides an indication of the constituent structure of the Thai source.

4.2 Thai Sentence Disambiguation

While sentence breaking is not specifically necessary for information structure analysis, nevertheless an initial goal of this project was to develop consistent guidelines for the treatment of sentences in order to facilitate organized study of the corpus. In Thai orthography, sentence-final punctuation is not used. The space character is a necessary but insufficient marker of a "sentence break," because it is also used prescriptively in certain typographic situations and may be used to separate ideas within such "sentences" (Wathabunditkul 2003). While some of the more specific prescriptive uses can be identified, this overloading renders the space character's marking function fundamentally lossy, and the "sentences" in running Thai text cannot be consistently disambiguated, even by fluent natives (Aroonmanakun 2007).

(1)

ลำน้ำนั้นไหลมาจากทิศตะวันตก					
ลำน้ำ	นั้น	ไหล	มา	จาก	ทิศตะวันตก
lamnám	nán	lăj	ma:	tɕà:k	tʰíttàʔwantòk
stream	[dmnst]	flow	[asp]	from	west
[1.][2] The stream flowed from the west					

(2)

ลำน้ำนั้นไหล มาจากทิศตะวันตก					
ลำน้ำ	นั้น	ไหล	มา	จาก	ทิศตะวันตก
lamnám	nán	lăj	ma:	tɕà:k	tʰíttàʔwantòk
stream	[dmnst]	flow	[asp]	from	west
The stream flowed. (It) came from the west.					

There are subtle feedback relationships between sentence-determination (as informed by space characters), asyndetic coordination, grammaticalized aspect marker disambiguation[3], verb serialization, and subject-dropping. To illustrate these interactions, consider (1), the sum of text which appears before the first space character in the narrative. As attested by informant 'A', the correct interpretation of this fragment is shown, but this text also supports an asyndetic coordination, pro-drop reading where the aspect marker **มา** /ma:/ is taken to be a main verb

[2] Square-bracketed numbers in the text indicate the sequence number of the sentence in the annotated story. When a single sentence generates multiple annotation units, alphabetic suffixes 'a', 'b', etc. are used.

[3] Thai aspect markers are grammaticalized from main verbs from which they are orthographically indistinguishable.

"come," from which it was grammaticalized, giving the acceptable (if pragmatically unusual), "The stream flowed, (and) (it) came from the west." Furthermore, by inserting a space before มา as shown in (2), this is now the preferred semantics.

For this project, the text was divided into sentences according to a subjective balance of the above criteria. To avoid the ambiguity described above, I did not impute a dropped subject for serialized verbs. With similar motivation, I avoid ascribing overt English coordination when there is no overt coordination lexeme in Thai, even if this results in stylistically impoverished English. For example, in (3) the translation, "She was nearly crying, and paralyzed with anger," is not used. Thai orthography containing a space, followed by resumptive serialization (without an overt subject) was preferentially treated as a single sentence, rather than a new sentence with subject drop (3).

(3)

เธอแทบร้องไห้ หันรีหันขวางด้วยความขุ่นเคือง						
เธอ	แทบ	ร้องไห้	หันรีหันขวาง	ด้วย	ความ	ขุ่นเคือง
tʰɤ̂ː	tʰêːp	rɔ́ːŋhâj	hănriːhănkʰwăːŋ	dûaj	kʰwaːm	kʰùnkʰɯːaŋ
she	almost	cry	paralyzed	by	[→state]	angry

[113.] She was nearly crying, paralyzed with anger.

A single continued sentence was less favored when the preceding serial verb had adjuncts or clauses to its right, especially when these intervening constituents contained a (non-subject) noun or noun phrase (4, continuing to 5). As shown, a new sentence is designated with an initial dropped subject marker.

(4)

หลุมพอตันใหญ่งอกงามชิดตลิ่ง					
หลุมพอ	ตัน	ใหญ่	งอกงาม	ชิด	ตลิ่ง
lŭmpʰɔː	tôn	jàj	ŋɔ̂ːkŋaːm	tɕʰít	tàlìŋ
lumpaw	[cls]	big	sprout	near	bank

[4.] A lumpaw tree sprouted from the bank.

(5)

แผ่กิ่งก้านสาขาทอดเงาไปยาวไกล...								
	แผ่	กิ่งก้าน	สาขา	ทอด	เงา	ไป	ยาว	ไกล
	pʰɛ̀ː	kìŋkâːn	săːkʰăː	tʰɔ̂ːt	ŋaw	paj	jaːw	klaj
Øs	spread-out	bough	branch	extend	shadow	[asp]	long	far

[5.] It spread branches, cast a shadow long and far,..

Semantic considerations also influence sentence breaking. The space between (6) and (7) was judged to be sentence-breaking despite strong bias for serialization, because it seems unlikely that the girl's repositioning is an additional consequence, beyond the redesigning, of her displeasure with her efforts.

(6)

ตรงไหนที่เห็นว่ายังไม่ชอบใจเธอก็เข้าไปเสริมแต่ง															
ตรง	ไหน	ที่		เห็น	ว่า		ยัง	ไม่	ชอบ	ใจ	เธอ	ก็	เข้า	ไป	เสริม	แต่ง
troŋ	nǎj	tʰîː		hěn	wâː		jaŋ	mâj	tɕʰɔ̂ːp	tɕaj	tʰɤ̂ː	kɔ̂ː	kʰâw	paj	sɤ̌ːm	tɛ̀ŋ
at	where	that	Øs	see	[→comp]	Øs	still	not	like	heart	she	[link]	[asp]	[asp]	reinforce	decorate

[78.] Wherever she saw that it still wasn't pleasing, she went back and enhanced her design.

(7)

กระถดไปมาอยู่พักหนึ่งจึงหันไปทางกลุ่มเด็กผู้ชายอีกที												
	กระถด	ไป	มา	อยู่	พักหนึ่ง		จึง	หัน	ไป	ทาง	กลุ่ม	เด็กผู้ชาย	อีกที
	kràʔtʰòt	paj	maː	jùː	pʰáknɯ̀ŋ		tɕɯŋ	hăn	paj	tʰaːŋ	klùm	dèkpʰûːtɕʰaːj	ʔìːktʰiː
Øs	reposition	[asp]	[asp]	[asp]	for a while		therefore	turn	go	towards	group	boy	again

[79.] She repositioned herself for a moment; consequently turning towards the group of boys again.

In order to more usefully generalize about structural forms in a small corpus, certain top-level structural forms were decomposed. In this step, each sentence identified according to the criteria described above yielded one or more *annotation units*, each of which had the form of a matrix sentence. The decomposed forms include: independent sentences occurring within quoted dialog (these instances were tagged as originating in quoted speech); overt S-coordination; if-then; S-complement; and the like. Function words related to these forms were ignored, even

if they were embedded in the decomposed parts. For example, in (8), **จึง** /tɕɯŋ/ "therefore" is embedded in the second of two annotation units but is not considered part of the annotation unit. Decomposition was not performed for coordination which is due to monolithic verb serialization, as in (3).

(8)

ตะวันเที่ยงลอยคว้างตรงหัว เงาหลุมพอจึงหดสั้นเข้ามาทุกที

ตะวัน	**เที่ยง**	**ลอย**	**คว้าง**	**ตรง**	**หัว**	**เงา**	**หลุมพอ**	**จึง**	**หด**	**สั้น**	**เข้า**	**มา**	**ทุกที**
tàʔwan	tʰîːaŋ	lɔːj	kʰwáːŋ	troŋ	hŭːa	ŋaw	lŭmpʰɔː	tɕɯŋ	hòt	sân	kʰâw	maː	tʰúktʰiː
sun	noon	float	adrift	straight	head	shadow	lumpaw	therefore	retract	short	[asp]	[asp]	increasingly

[96.] The noontime sun loitered overhead, so the shadow of the lumpaw tree retracted increasingly shorter.

In (9), the three independent annotation units that are categorized for the sentence are shown in square brackets. The function words, outside the bracketed annotation units, are discarded. Because of this decomposition, there are more annotation units than there are sentence designations in this study; the study comprises 296 annotation units derived from 197 sentences, an average of 1.5 units per sentence. This ratio is relatively high because each of the 33 sentences with quoted dialog generated at least two annotation units, one for the quotation frame, and one for the quoted speech. Often a single quotation frame contained multiple sentences of quoted speech.

(9) When [one of them dove down and stuck in the branch, hiding it in the base of the stream, the others would dive hunting for it], and if [anyone was able to find it], then [he had the right to chase and hit his friends].

After sentence breaking, the identified annotation units are independently analyzed for syntactic features (subject- and object-drop, Sections 4.3 and 4.4; syntactic structure, Section 4.5) and information structure (Section 4.6). Note that this methodology presumes that the information structure phenomena of interest are localizable within the matrix sentence, a hypothesis which is adopted without examination.

4.3 Subject- and Object-Drop Recovery

The symbol Øs was inserted to represent the grammatical word position for each dropped subject in the text. A firm annotation guideline used here is that drop symbols may *only* appear in places where an overt Thai expletive or pronoun would be grammatical. Consultant 'A' found elicitation for some of these covert-recovery judgments difficult, so strong is the mandate for the dropping of recoverable arguments. (10) illustrates dropped pronouns which, when made overt, 'A' found cumbersome but acceptable. In some cases, I constructed parallel structures which introduce new referents, and then substituted-in one pronoun at a time until the insertion position of all drop symbols as pronouns could be simultaneously validated in the target sentence.

(10)

มองดูแต่ไกลเหมือนตุ๊กตาตัวใหญ่ไม่สวมเสื้อ

มองดู		**แต่**	**ไกล**		**เหมือน**	**ตุ๊กตา**	**ตัว**	**ใหญ่**	**ไม่**	**สวม**	**เสื้อ**
mɔːŋduː		tɛ̀ː	klaj		mŭːan	túkkàtaː	tuːa	jàj	mâj	sŭːam	sûːa
Øs	examine Ød.o.	from	far	Øs	resemble	doll	[cls]	big	not	wear	shirt

[30.] Seen from afar, she looked like a big doll with no shirt.

(11)

จากนั้นก็ดำลงไปปักซ่อนไว้อีก

จาก	**นั้น**	**ก็**	**ดำ**	**ลง**	**ไป**	**ปัก**	**ซ่อน**	**ไว้**	**อีก**
tɕàːk	nán	kɔ̂ː	dam	loŋ	paj	pàk	sɔ̂ːn	wáj	ʔìːk
from	there Øs	[link]	dive	descend	go	stick-in Ød.o.	conceal	[asp]	again

[58.] After that, he dove down to hide it again.

The object drop symbol Ød.o. was only introduced where the verb strongly or necessarily transitive, or when the referent was deemed significant and identifiable. For example, in (11), the referent of "stick" figures prominently in the discourse at this point, and the verb **ปัก** /pàk/ "to stick in" seems strongly transitive. Similarly, consultant 'A' was not able to form an intransitive use of the verb **ตบแต่ง** /tòptɛ̀ŋ/, "to decorate," so the position of the dropped pronoun in (12) is annotated. Its recoverable referent is the "big pile of sand."

(12)

...และรีบตะล่อมทรายขึ้นเป็นทรงสูงใช้มือตบแต่งอย่างระมัดระวัง

และ	รีบ	ตะล่อม	ทราย	ขึ้น	เป็น	ทรง	สูง	ใช้	มือ	ตบแต่ง		อย่าง	ระมัดระวัง
lɛ́ʔ	riːp	taʔlɔ̂ːm	saːj	kʰûn	pen	soŋ	sŭːŋ	tɕʰáj	mɯː	tòptɛ̀ŋ		jàːŋ	ráʔmátráʔwaŋ
and	Øs hurry	rake	sand	[asp]	[→caus]	style	tall	use	hand	adorn	Ød.o.	[→adv]	careful

[124c.] (...and) she scooped up sand, making a big pile, using her hands to decorate it carefully.

As noted in the discussion of sentence disambiguation, drop-recovery annotation decisions interact with sentence-breaking, and can also interact with the assignment of primary constituent type. The latter can occur, for example, when the putative topic marking particle (Section 4.5.2.6), which permits an optionally dropped pronoun to its left (13), is alternatively treated as an adverbial linker, which does not (14). Both readings, plus the reading where the dropped pronoun is overt, are judged acceptable by consultant 'A'. Because of interactions such as these, the annotation process could not be considered final when each stage was initially completed, as rippling effects had to be considered when completing annotations in other stages, or when reconsidering previously-completed annotations.

(13)

ปากก็ร้องเสียงดังลั่น

ปาก		ก็	ร้อง	เสียง	ดัง	ลั่น
pàːk		kɔ̂ː	rɔ́ːŋ	sǐːaŋ	daŋ	lân
mouth	Øs	then	exclaim	sound	loud	loud

[135.] His mouth, (it) then cried out with a loud sound

(14)

ปากก็ร้องเสียงดังลั่น

ปาก	ก็	ร้อง	เสียง	ดัง	ลั่น
pàːk	kɔ̂ː	rɔ́ːŋ	sǐːaŋ	daŋ	lân
mouth	then	exclaim	sound	loud	loud

[135.] His mouth then cried out with a loud sound.

4.4 Subject and Object Coding

The previous section explained how the recovered *positions* of dropped entities are represented. In addition to inserting null tokens as described, the *facts* of subject- and object-dropping—plus other phenomena related to *entity coding*—were recorded as *ad-hoc feature-based annotations* for the annotation unit as a whole. Overt, non-pronomial subjects were given no ad-hoc features. Beyond this, features were expressed with forms such as SUBJECT=DROP, which designates that a dropped subject token was imputed. Overt pronomial subjects were designated SUBJECT=PRONOUN. The existential expletive was given the annotation SUBJECT=DUMMY and second-person impersonal subjects were denoted with SUBJECT=IMPERSONAL; these last two designations could additionally be marked SUBJECT=DROP when implied but uninstantiated.

4.5 Syntactic Structure

In order to test the hypothesis that information structure is marked by sentence-scale syntactic structure in Thai, each annotation unit was categorized according to the general ordering of its grammatical constituents. With such a task, there is a danger of presupposing, *a priori*, syntactic correlations for the information structure phenomena we are trying to study. To minimize this, I avoid using terms such as "topicalization" when describing constituent structure, instead using terms that attempt to characterize the actual direction of reordering from canonical Thai constituent order. Specifically, I differentiate LEFT- and RIGHT-DISLOCATION, movements which introduce an overt pronoun, from FRONTING and BACKING, which do not. Unfortunately, in *TOPIC*-COMMENT, the field has a well- and long-established name for an important syntactic phenomenon which conflicts with the information-structural use of the term, TOPIC. On this point, it would be tedious for my readers to accommodate an alternative presentation. To avoid confusion, I will italicize *TOPIC* when it refers to the syntactic constituent.

As noted, The annotation system was ad-hoc and feature-based, allowing an arbitrary number of feature strings to be assigned to each annotation unit. I developed a closed set of *primary constituent structure types*, exactly

one of which designated for each annotation unit. These basic types are detailed in Section 4.5.2. Optional features, described in the previous section and in Section 4.5.3, were applied for cross-cutting secondary phenomena.

During the course of this work, the Thai particle, ก็ /kɔ̂ː/ was gradually identified as a lexeme of great interest since it may have information-structural function as a *fronting marker*. An intriguing hypothesis is that, when ก็ occurs in a position to the right of a subject, it causes that subject to be summarily cleft from its predicate, either with or without a resumptive pronoun. Its absence renders the construction indistinguishable from CANONICAL sequence. This role is discussed throughout my remarks in this section, and specifically in Section 4.5.2.6.

4.5.1 **Subject- and Object Movement**

I begin by discussing key constituent movements: RIGHT-DISLOCATION, BACKING, LEFT-DISLOCATION, and object FRONTING. When an annotation unit manifests one of the described structures, the phenomenon is notated as an ad-hoc feature. Because of the way the feature set evolved, these are considered modifications of the 'CANONICAL' designation. Subject FRONTING is treated as an instance of the *TOPIC*-COMMENT syntactic pattern, see Section 4.5.1.4. Subject- and object-FRONTING will receive extended discussion, and I also present a brief overview of an interesting structural property of noun phrases that renders them more amenable to FRONTING in Section 4.5.1.5.

4.5.1.1 Subject Right-Dislocation

Matrix subjects can be dislocated to the right. When this leaves an overt pronoun in place, the phenomenon is called RIGHT-DISLOCATION (Singnoi 2007: 8). The canonical form of (15) is shown in (16). Other names for RIGHT-DISLOCATION include *backing topicalization* (Warotamasikkhadit 1997: 304) and *anti-topic* (Lambrecht 1996: 203 after Chafe 1976). The feature SUBJECT=RIGHT is coded for this annotation unit.

(15)

พวกเขามากันห้าคน					
พวกเขา	มา	กัน	ห้า	คน	
pʰûːakkʰăw	maː	kan	hâː	kʰon	
they	come	together	five	person	
[13.] They came together, five of them.					

(16)

ห้าคนมากัน			
ห้า	คน	มา	กัน
hâː	kʰon	maː	kan
five	person	come	together
Five people came together.			

4.5.1.2 Backing

In SUBJECT BACKING, a rightwards-moved subject leaves no pronoun trace (Warotamasikkhadit 1997: 304). The annotation text contains one case of subject BACKING. In (17), the subject is dropped from CANONICAL position and positioned to the right. Of syntactic interest, perhaps, is that the polite particle is not outflanked by the BACKING movement.

Warotamasikkhadit also categorizes a certain benefactive alternation as an instance of BACKING, but the handful of benefactive constructions in our text (ให้ /hâj/ "give" in 18, e.g.) seem to exhibit the canonical order of complements for Thai ditransitives, that is, the direct object followed by the indirect object. Therefore, in the study text, only (17) receives SUBJECT=BACK.

(17)

เป็นไงแกน่ะ				
เป็น	ไง	แก	น่ะ	
pen	ŋaj	kɛː	nâʔ	
Øs is	how	[2s]	[polite]	
[24a.] What about you?				

(18)

ริ้วคลื่นที่โหมกระหน่ำมาไม่ยั้งหยุดทำให้เธอไม่อาจก่อขึ้นได้ง่าย ๆ																
ริ้ว	คลื่น	ที่	โหมกระหน่ำ	มา	ไม่	ยั้ง	หยุด	ทำ	ให้	เธอ	ไม่	อาจ	ก่อ	ขึ้น	ได้	ง่าย ๆ
ríw	kʰlûːn	tʰîː	hǒːmkràʔnàm	maː	mâj	jáŋ	jùt	tʰam	hâj	tʰɤ̌ː	mâj	ʔàːt	kɔ̀ː	kʰûn	dâj	ŋâːjŋâːj
furrow	wave	which	buffet	[asp]	not	halt	halt	make	[benef.]	she	not	[pot]	construct	[asp]	[pot]	easy

[147.] The lines of waves which buffeted did not stop and made it not easy for her to be able to successfully build.

4.5.1.3 Subject Left-Dislocation

In LEFT-DISLOCATION, a leftwards-moved constituent is followed by a resumptive pronoun. Example (19) is the lone instance of LEFT-DISLOCATION in the narrative. Without the deictic demonstrative pronoun, the sentence would still be grammatical, as the overt pronoun มัน /man/ "it" could refer to any felicitous antecedent. Although this pattern could be considered a fundamental instance of the *TOPIC*-COMMENT syntactic pattern, I reserve that designation for cases with dual subjects or an enhanced semantic relation between the subject and the nucleus, and prefer the syntactic analysis for cases of simple subject reduplication such as (19). Accordingly, that example is denoted CANONICAL, SUBJECT=LEFT.

(19)

นั่นมันทรายต่างหาก				
นั่น	มัน	ทราย	ต่าง	หาก
nân	man	saːj	tàːŋ	hàːk
[dmnst-prn]	it	sand	each	separate

[102c.] That, it's just some sand.

4.5.1.4 Subject Fronting

In FRONTING, a leftwards-moved constituent leaves no overt pronoun trace (Singnoi 2007: 8). I hasten to note, however, that absent a syntactic marker, FRONTING of a subject—even when it is rightwards-bounded by a determiner phrase complement (see Section 4.5.1.5)—is syntactically indistinguishable from a subject in canonical constituent order. Therefore, in Thai, for syntactic FRONTING to exist, it must be overtly marked. Later in this paper, I advance the hypothesis that the particle ก็ /kɔ̂ː/ can perform this function. When this word, which has many functions in Thai, acts to demarcate a FRONTED constituent, I will call it the *fronting marker*. However, for reasons discussed in Section 4.5.2.6, I elect to annotate all instances of FRONTING that are overtly marked with ก็ as instances of the *TOPIC*-COMMENT pattern.

To summarize, since *unmarked* SUBJECT FRONTING is degenerate (except perhaps when assisted by semantic infelicity, see Section 4.5.1.6 and example 33), and *marked* subject FRONTING is considered to be a subtype of the *TOPIC*-COMMENT syntactic pattern, the annotation SUBJECT=FRONT is not used in this study. OBJECT FRONTING, on the other hand, can be syntactically self-marking and widely attested in Thai. Before continuing to that discussion, however, it will be helpful to discuss noun phrases which are rigidly rightwards-bounded. This is the topic of the following section.

4.5.1.5 Rigidly-Rightwards-Bounded Noun Phrases

Here, I briefly address the claim of one author that an extra-syntactic dependency may apply to leftwards-moved noun phrases. The discussion will lead to an examination of what may be an important structurally distinguishing feature of NPs. In any case, the claim requires examination if a purely syntactic treatment of NP movement is to be licensed.

I begin by examining the case of LEFT-DISLOCATION. Although the resumptive pronoun might be thought sufficient demarcation for LEFT-DISLOCATING a subject, we must account for the observation that bare nominals are not felicitous here. That is, while CANONICAL (20), and LEFT-DISLOCATED (21) are acceptable, *

(22), is judged marginal by informant 'A', and is generally rejected in reference grammars (Iwasaki and Ingkaphirom 2005: 361). To account for this, Warotamasikkhadit suggests that leftwards-moved arguments "must be made definite" by the presence of one of the definite determiners, or the resumptive pronoun "which is always definite" (Warotamasikkhadit 1997: 304). This language seems to suggest that he believes this constraint to be pragmat-

ically motivated, but I favor a simpler syntactic motivation which is easily tested here. My view may be shared by Singnoi, who notes that one "quite rigid" syntactic function of Thai demonstratives is to establish the rightwards boundary of an NP (Singnoi 2004: 647-648), and by Iwasaki and Ingkaphirom, who go so far as to distinguish the demonstrative lexemes as "topic markers" when they function in this role (Iwasaki and Ingkaphirom 2005: 361). To test the view that it is syntactic rigidness that is required, and not pragmatic definiteness, I formed a non-corpus sentence (23) with a LEFT-DISLOCATED subject and the non-universal quantifier "some." Informant 'A' judges this sentence acceptable, so assuming that a resumptive pronoun can be no more definite than its antecedent, I ascribe the well-formedness requirements for LEFT-DISLOCATED constituents to syntactic—and not pragmatic—considerations.

(20)

ลำน้ำนั้นไหลมาจากทิศตะวันตก					
ลำน้ำ	นั้น	ไหล	มา	จาก	ทิศตะวันตก
lamnám	nán	lăj	maː	tɕàːk	tʰíttàʔwantòk
stream	[dmnst]	flow	[asp]	from	west
[1.] The stream flowed from the west.					

(21)

ลำน้ำนั้นมันไหลมาจากทิศตะวันตก						
ลำน้ำ	นั้น	มัน	ไหล	มา	จาก	ทิศตะวันตก
lamnám	nán	man	lăj	maː	tɕàːk	tʰíttàʔwantòk
stream	[dmnst]	it	flow	[asp]	from	west
The stream, it flowed from the west.						

*
(22)

ลำน้ำมันไหลมาจากทิศตะวันตก					
ลำน้ำ	มัน	ไหล	มา	จาก	ทิศตะวันตก
lamnám	man	lăj	maː	tɕàːk	tʰíttàʔwantòk
stream	it	flow	[asp]	from	west
The stream, it flowed from the west.					

(23)

คนบางคนเขาทำให้ผมปวดหัว						
คน	บาง	คน	เขา	ทำให้	ผม	ปวดหัว
kʰon	baːŋ	kʰon	kʰăw	tʰamhâj	pʰŏm	pùːathŭːa
person	some	[cls]	they	cause	I	headache
Some people, they give me a headache.						

As for unmarked SUBJECT *FRONTING*, even rigid rightwards NP edges—that is, determiner phrases that end with a demonstrative—are insufficient for establishing contrast with CANONICAL subject position (recall that the fronting marker is typically introduced to achieve this purpose), so Warotamasikkhadit's claim is not so easily settled in the *absence* of a resumptive pronoun, and this is left for future work. For purposes of such investigations, I occasionally make note of the "rigid" or "non-rigid" characteristic of fronted NPs in my discussions below.

4.5.1.6 Object Fronting

OBJECT FRONTING, authors describe widely (Singnoi 2007: 5, Smyth 2002: 116, Warotamasikkhadit 1997: 303), and I now return to a discussion of this phenomenon. In Section 4.5.1.4 I noted that OBJECT FRONTING can be self-marking, and it is now possible to note that this is due to the rigid rightwards boundary requirement which I elaborated in the previous section. With a FRONTED object and overt subject, two NPs will be adjacent, and the rigid rightwards boundary on the former helps establish the FRONTING pattern.

It would be interesting to exhibit some examples from the study corpus, but the only use of object FRONTING in the study corpus—including instances within quoted text—is for the dialog frame itself, where the quoted speech as a whole, being the object of a speech-act verb, is FRONTED. In fact, this pattern is used every time speech is attributed to a speaker in the text. Quoted dialog appears to manifest a different form of NP rightwards rigidity, as—even without the use of quotation mark punctuation[4]—quoted utterances can usually be disambiguated from the

[4] Although Western orthographic punctuation is sometimes used in Thai, it is not mandatory and may be considered prescriptively or stylistically poor practice.

adjacent dialog-frame subject, probably via biased consideration of context, orthographic space, or usage conventions of common speech act verbs.

4.5.2 Primary Constituent Structure Type

As discussed in 4.5, each annotation unit is assigned exactly one *primary constituent structure type*. The complete set used in this annotation project is: CANONICAL, BENEFACTIVE, IMPERATIVE, INTERJECTION, PASSIVE, QUESTION, *TOPIC*-COMMENT, COPULA, and S-COMPLEMENT. Some of these support or require further specification, as follows: the CANONICAL designation can be extended by noting the chaining of multiple, semantically related clauses (but note that serialization over the same subject is considered an ordinary CANONICAL pattern); the COPULA designation can be extended to denote that the copular word is uninstantiated with the extended notation COPULA=DROP; the S-COMPLEMENT designation *requires* specification of one of the other primary constituent structure types, using the notation S-COMP=*type*; and the *TOPIC*-COMMENT designation *requires* that the *TOPIC*-to-COMMENT relation type be specified as either *possessor*, *secondary-cause*, or *subject*. I now review these types in detail.

4.5.2.1 Canonical

The canonical constituent order in Thai is SUBJECT – VERB – OBJECT, where additional predicates may be serialized over the same subject:

[S] V [O] [[V [O]] [...]]

Verb serialization constructions in Thai include directional, temporal, conjunctive, and resultative varieties, which I will not classify here; refer to Muansuwan (2001, 2002) and Thepkanjana (1986) for treatments. Grammatical aspect markers assume positions both before and after main verbs and interact with verb serialization; I limit the scope of the current work to coding their presence via sense-selection; the analyses of Koenig and Muansuwan (2000, 2005) guided this aspect of the task. Monolithic serialization, where action proceeds uniformly through a string of verbs and aspect markers, is annotated as a CANONICAL structure. In (24), a prepositional phrase adjunct is seen in CANONICAL position (to the right of the verb); as with syntactic subjects and objects (Section 4.5.1), ad-hoc features can be added to the "CANONICAL" designation should a verbal adjunct be positioned left-wise. This is discussed in 4.5.3.1.

(24)

เด็กหญิงทรุดนั่งบนผืนทราย					
เด็กหญิง	ทรุด	นั่ง	บน	ผืน	ทราย
dèkjǐŋ	sút	nâŋ	bon	pʰɯ̌ːn	saːj
girl	kneel	sit	on	surface	sand

[115.] The girl sank back on the sand.

In addition, the story exhibits cases of *clause chaining*, where a semantic meta-relation is expressed through the juxtaposition of two (or more) semantically complete events expressed as clauses. When these chains of semantically related clauses did not meet the requirements for partition into distinct annotation units, they were given the designation CANONICAL=CHAIN. With this annotation, ad-hoc features can be independently attached to the subject of the chained clause by using notation of the form, (e.g.) SUBJECT-2=DROP.

4.5.2.2 Benefactive

The syntax of the BENEFACTIVE construction in Thai is a complex subject, and documenting it meaningfully is beyond the scope of this study. In order to permit the four sentences found in the text to be handled accordingly during analysis, they were identified with the primary constituent structure type BENEFACTIVE. Without further comment, I show a simple instance in (25).

(25)

เธอกอบทรายให้ดู					
เธอ	กอบ	ทราย	ให้		ดู
tʰɤː	kɔ̀ːp	saːj	hâj		duː
she	scoop	sand	[benef.]	Ød.o.	look

[180.] She scooped up some sand for him to see.

4.5.2.3 Imperative and Interjection

In this story, imperative declarations and interjections occur only within quoted dialog. Corresponding instances are marked with the primary type IMPERATIVE or INTERJECTION. In the case of the latter, annotation features related to dropping or movement are not coded. This allows for consistent treatment of utterances such as (26).

(26)

ใช่
ใช่
tɕʰâj
yes

[23a.] Yeah.

4.5.2.4 Passive

Because the phenomenon manipulates subjects and objects in distinctive ways, I chose to annotate occurrences of the overt "adversative" passive marker **ถูก** /tʰùːk/ (Singnoi 2002: 76) as a primary constituent structure type. I do not attempt to insert a dropped subject marker for the demoted subject of a PASSIVE. For example, in (27), the boys who jumped in the pool are not syntactically positioned. Likewise in (45), the animate actor ("the girl") is not made explicit. As for Singnoi's "non-adversative" type, the overt[5] preposition **โดย** /doːj/ "by" does not appear in the study text, so I conclude that there are no instances of this form of passive construction.

(27)

ผืนน้ำราบเรียบถูกแรงกระแทกแตกกระจายเป็นฟองพราย...										
ผืน	น้ำ	ราบเรียบ	ถูก	แรง	กระแทก	แตก	กระจาย	เป็น	ฟอง	พราย
pʰɯ̌ːn	náːm	râːprîːap	tʰùːk	rɛːŋ	kràʔtʰɛ̂ːk	tɛ̀ːk	kràʔtɕaːj	pen	fɔːŋ	pʰraːj
surface	water	flat	[pass]	power	crash	break	disperse	[→caus]	froth	brilliant

[142.] The smooth surface of the water was smashed into a sparkling froth...

4.5.2.5 Question

Beyond designating QUESTION as the primary constituent structure type, and annotating the co-occurrence of certain structural movements (the one example of subject BACKING in the text appears in the context of an interrogative, see Section 4.5.1.2), the syntax of interrogative forms is not investigated in this study.

4.5.2.6 *Topic*-Comment

The so-called *TOPIC*-COMMENT syntactic pattern[6] is attested in Thai (Smyth 2002:117). Hewing to convention, I exclude clausal *TOPICS* from this discussion, but will close this section with some remarks on this point. In the present treatment, *TOPIC*-COMMENT is distinguished from simple FRONTING by either: an enhanced semantic relationship of the left-wise constituent to the COMMENT—*TOPIC*-COMMENT generally permits greater latitude here; or the presence of an overt fronting marker. [7] In other words, although the most basic *TOPIC*-COMMENT pattern is indistinguishable from a canonical SUBJECT-VERB sentence (or FRONTED subject), the former is more flexible in allowing the referent of the syntactic *TOPIC* to enter enhanced semantic relationships with the COMMENT. Being purely syntactic, the LEFT-DISLOCATION and FRONTING analyses are simpler than the more semantically-adept *TOPIC*-COMMENT. Therefore, when unmarked and when extended semantics is not signaled, I select the syntactic phenomenon, rather than *TOPIC*-COMMENT, as the constituent structure for the annotation unit. I further define the COMMENT as requiring a verb (NP-NP structures are discussed in 4.5.2.7).

---

[5] Singnoi considers **โดย** /doːj/ a preposition and not a function word, and so she calls this construction "unmarked." (Singnoi 2002: 86)

[6] The use of the term "topic" in describing the *TOPIC*-COMMENT syntactic pattern should not be construed as implying that the pattern necessarily marks for information structure topic. To avoid confusion, we italicize topic when it refers to the left-positioned syntactic constituent in the *TOPIC*-COMMENT pattern.

[7] Here, I admit the possibility of additional syntactic clefting mechanisms beyond the one that I propose at the end of this section, such as those proposed by Iwasaki and Ingkaphirom (2005: 361).

I propose a topic-demarcation role for the common Thai particle ก็ /kɔ̂ː/, while noting that the particle has a wide range of uses—including that of discourse marker—and it may at times simultaneously satisfy multiple roles from this set. For example when the particle appears to separate clauses, it must be relevant to clause-chaining, a role which I won't investigate here. While collapsing multiple distinct roles of the particle ก็ may seem like an appealing generalization, it is clear from detailed studies on Thai verb serialization (Thepkanjana 1986, Muansuwan 2001, 2002), and the vast literature on the (nominal-)*TOPIC*-COMMENT pattern that—beyond the structural the structural similarity discussed at the end of this section—multiple phenomena are involved.

แล้วเด็กชายทั้งหมดก็รีบถอดเสื้อผ้าผลัดกองไว้บนเนินทรายเหลือตัวล่อนจ้อน

แล้ว	เด็กชาย	ทั้ง	หมด		ก็	รีบ	ถอด	เสื้อผ้า	ผลัด	กอง	ไว้	บน	เนิน	ทราย	เหลือ	ตัว	ล่อนจ้อน
(28) lɛ́ːw	dèktɕʰaːj	tʰáŋ	mòt		kɔ̂ː	rîːp	tʰɔ̀ːt	sûːapʰâː	pʰlàt	kɔːŋ	wáj	bon	nɤ̌ːn	saːj	lŭːa	tuːa	lɔ̂ːntɕɔ̂ːn
then	boy	all	total	Øs	[link]	hurry	strip	clothes	switch	pile	keep	on	mound	sand	remain	self	naked

[36.] Then, all of the boys, (they) took off their clothes and threw them on a hump of sand, leaving themselves naked.

แล้วเด็กชายทั้งหมดรีบถอดเสื้อผ้าผลัดกองไว้บนเนินทรายเหลือตัวล่อนจ้อน

แล้ว	เด็กชาย	ทั้ง	หมด	รีบ	ถอด	เสื้อผ้า	ผลัด	กอง	ไว้	บน	เนิน	ทราย	เหลือ	ตัว	ล่อนจ้อน
(29) lɛ́ːw	dèktɕʰaːj	tʰáŋ	mòt	rîːp	tʰɔ̀ːt	sûːapʰâː	pʰlàt	kɔːŋ	wáj	bon	nɤ̌ːn	saːj	lŭːa	tuːa	lɔ̂ːntɕɔ̂ːn
then	boy	all	total	hurry	strip	clothes	switch	pile	keep	on	mound	sand	remain	self	naked

Then, all of the boys took off their clothes and threw them on a hump of sand, leaving themselves naked.

In (28), the particle is shown cleaving-off the (non-rigid) NP subject of a CANONICAL sentence. An unusual aspect of this phenomenon is that the lexeme does not precisely partition constituents to its left and right but rather its presence to the *right* of the optional resumptive pronoun—a position that might be considered *within* the syntactic COMMENT—designates the syntactic *TOPIC* as FRONTED or LEFT-DISLOCATED. The canonical version of the sentence is shown in (29). I do not consider the presence of the overt fronting marker to be the type of trace that qualifies (28) as an instance of LEFT-DISLOCATION, because an overt pronoun is still felicitous *before* the linking particle, as indicated by Øs. Unfortunately, in this study corpus, there are no such instances—ก็ co-occurring with both an overt pronoun and a LEFT-DISLOCATED subject—but consultant 'A' judges the case grammatical. His comment is that the reduplication feels to him like colloquial speech, the particle in this case perhaps serving as a discourse marker. Example (30) shows another example of FRONTING marked by ก็, this time where a subject pronoun is dropped, leaving only its quantifying specifier in place. A final example (31) shows that the marking pattern is compatible with a rigid subject NP. Here the demonstrative นั้น /nán/ marks the rightwards boundary of an NP which has a classifier-demonstrative complex as its specifier.

แล้วทั้งสองก็หัวเราะขึ้นพร้อมกันดัง ๆ

แล้ว		ทั้ง	สอง		ก็	หัวเราะ	ขึ้น	พร้อมกัน	ดัง ๆ
(30) lɛ́ːw	Øs	tʰáŋ	sɔ̌ːŋ	Øs	kɔ̂ː	hŭːarɔ́ʔ	kʰûn	pʰrɔ́ːmkan	daŋdaŋ
then	Øs	all	two	Øs	[link]	laugh	[asp]	concurrently	loudly

[194.] Then, the two of them, (they) laughed loudly, together.

เพียงอึดใจลำคลองสายนั้นก็สำสลอนไปด้วยเด็กชายตัวเปล่าเปลือย

เพียง	อึดใจ	ลำคลอง	สาย	นั้น		ก็	สำ	สลอน	ไป	ด้วย	เด็กชาย	ตัว	เปล่า	เปลือย
(31) pʰiːaŋ	ʔɯ̀ttɕaj	lamkʰlɔːŋ	sǎːj	nán		kɔ̂ː	sǎm	sàlɔ̌ːn	paj	dûaj	dèktɕʰaːj	tuːa	plàːw	plɯːaj
only	moment	canal	[clas]	[dmnst]	Øs	[link]	mixed	profuse	[asp]	with	boy	[cls]	empty	bare

[64.] In just a moment, that canal, (it) became profuse with naked boys.

One way to categorize subtypes of *TOPIC*-COMMENT is according to the semantic relation between the *TOPIC* NP and the COMMENT clause. In fact, in combination with a rigid NP (Section 4.5.1.5), a certain degree of syntactic infelicity between the *TOPIC*-COMMENT pair appears to be sufficient for signaling the *unmarked TOPIC*-COMMENT pattern (33) that we were unable to obtain with FRONTING. I speculate that such semantically-motivated patterns are most felicitous when the pair suggests a single semantic relationship that is oblique, but minimally so. For this study, subtypes of *TOPIC*-COMMENT were identified based on: rigidness of an NP's rightwards boundary, occurrence

of the syntactic fronting marker ก็ /kɔ̂ː/, and judgment of the semantic relationship between *TOPIC* and COMMENT. This collection of subtypes is by no means intended to be exhaustive, but rather just the set that appeared in the study text.

Subject

As discussed in the preceding section, the fronting marker ก็ /kɔ̂ː/ can be used to demarcate what is otherwise a traditional subject constituent and subject role in a canonical sentence (28, 30, and 31 above). When the fronting marker is not present and there is no semantic obliqueness between the subject and the putative COMMENT, this subtype of *TOPIC*-COMMENT will not be designated, since a CANONICAL or LEFT-DISLOCATION analysis will be sufficiently and preferentially explanatory.

Possessor

A leftwards NP can be in a possession relationship with the main sentence. This subtype can be marked or unmarked by the fronting marker. In an unmarked case (32), the topic "each person" denotes the (several) possessors of asyndetically conjoined "sickly hands and feet" and "wai-ling-green mouths" (a drop marker records the position of the pronoun required by Thai NP classifier-demonstrative complex). Note that the use of space in the original Thai orthography is not consistent with the partitioning of *TOPIC* and COMMENT, an issue I note without further comment.

(32)

แต่ละคนมือเท้าซีดเซียว ปากเขียวเหมือนยอดหวายลิง											
	แต่ละ	**คน**	**มือ**	**เท้า**	**ซีด**	**เซียว**	**ปาก**	**เขียว**	**เหมือน**	**ยอด**	**หวายลิง**
	tɛ̀ːlá?	kʰon	mɯː	tʰáːw	sîːt	siːaw	pàːk	kʰǐːaw	mɯ̌ːan	jɔ̂ːt	wăːjliŋ
Øs	each	[cls]	hand	foot	pale	sickly	mouth	green	similar	top	wai-ling

[98.] Each person, (their) hands and feet were sickly pale and mouth as green as the top of a wai-ling tree.

Secondary Cause

The syntactic *TOPIC* in a *TOPIC*-COMMENT structure can have a role of secondary causation with respect to the COMMENT (33). As with other subtypes, the fronting marker ก็ may follow the COMMENT'S (overt or covert) subject.

(33)

เกมนี้ พวกเขาสนุกสุดเหวี่ยง				
เกม	**นี้**	**พวกเขา**	**สนุก**	**สุดเหวี่ยง**
keːm	níː	pʰûːakkʰăw	sànùk	sùtwìːaŋ
game	[dmnst-det]	they	fun	extremely

[167.] Thanks to this game, they really had fun.

Putative "clausal-*TOPICS*"

In the introduction to this section, I noted that clausal topics are generally not admitted. I will close this discussion of the *TOPIC*-COMMENT syntactic pattern with a brief look at how leftwards clausal structures marked by the putative fronting marker ก็ /kɔ̂ː/ mirror the more traditional nominal *TOPIC*-COMMENT structure. The most compelling motivation for this treatment is that having to discard ก็ as euphonic or vacuous weakens the alternative resultative serialization analysis. Comparing putative *TOPIC*-COMMENT structure (34) with a conventional NP *TOPIC*-COMMENT structure (30, 31), and with a serialization reading (35), which draws upon the widely cited sense of ก็ as meaning "then," one notes that the positions of ก็ and imputed subjects correlate closely between the former pair.

(34)

พูดจบก็ผลุบหายลงไปพร้อมกิ่งไม้ในมือ													
	พูด	**จบ**		**ก็**	**ผลุบ**	**หาย**	**ลง**	**ไป**	**พร้อม**	**กิ่ง**	**ไม้**	**ใน**	**มือ**
	pʰûːt	tɕòp		kɔ̂ː	pʰlùp	hăːj	loŋ	paj	pʰrɔ́ːm	kìŋ	máːj	naj	mɯː
Øs	speak	end	Øs	[link]	dive	lose	descend	go	with	branch	wood	in	hand

[48.] When (he) had finished speaking, (he) dove under the water...

Another treatment of ก็ suggests that it is an overt conjunction here *

(36), but this analysis is rejected on the basis that the explicit conjunction **และ** /lɛ́ʔ/ "and" can still be added—arguably changing the semantics—immediately before the (overt or covert) subject of the main (rightwards) clause. Note that *

(36)—including the overt conjunction—is acceptable under the homonymous sense of **ก็**, "then:" "He finished speaking, and he <u>then</u> dove..."

(35)

พูดจบก็ผลุบหายลงไปพร้อมกิ่งไม้ในมือ												
พูด	**จบ**	**ก็**	**ผลุบ**	**หาย**	**ลง**	**ไป**	**พร้อม**	**กิ่ง**	**ไม้**	**ใน**	**มือ**
pʰûːt	tɕòp	kɔ̂ː	pʰlùp	hăːj	loŋ	paj	pʰrɔ́ːm	kìŋ	máːj	naj	mɯː
Øs speak	end	[then]	dive	lose	descend	go	with	branch	wood	in	hand
[48.'] He finished speaking, then (he) dove under the water...												

* (36)

พูดจบและก็ผลุบหายลงไปพร้อมกิ่งไม้ในมือ													
พูด	**จบ**	**และ**		**ก็**	**ผลุบ**	**หาย**	**ลง**	**ไป**	**พร้อม**	**กิ่ง**	**ไม้**	**ใน**	**มือ**
pʰûːt	tɕòp	lɛ́ʔ		kɔ̂ː	pʰlùp	hăːj	loŋ	paj	pʰrɔ́ːm	kìŋ	máːj	naj	mɯː
Øs speak	end	and	Øs	[and]	dive	lose	descend	go	with	branch	wood	in	hand
[48.''] (He) finished speaking, and (he) and? dove under the water...													

Semantic relation types between *TOPIC* and COMMENT in (34) could trivially be characterized along the lines of the accepted serialization types in Thai: directional, temporal, conjunctive, resultative, etc. In short, the proposal has mild allure, but for the purposes of annotation, here I adopt only the conventional *TOPIC*-COMMENT pattern, which does not admit clausal *TOPICS*. This concludes the discussion of *TOPIC*-COMMENT and I now continue with the review of the primary syntactic constituent structure types that are ascribed to each annotation unit in this project.

4.5.2.7 Copula

As noted above, NP-NP structures are not considered to be in the spirit of the *TOPIC*-COMMENT pattern; such an analysis seems undue, as it would require imputing both a dropped dummy subject and a dropped existential verb to the putative COMMENT. Instead, supported by the additional evidence of an immediately preceding parallel structure (37)—in which the copula is overt—I suggest a COPULA-DROP analysis for the equated NPs in (38). This analysis is consistent with the six instances of NP-NP in the text. In fact, the study text is consistent with the additional constraint that the first NP be deictic.

(37)

"ตรงนี้เป็นห้องครัว			
ตรง	**นี้**	**เป็น**	**ห้องครัว**
troŋ	níː	pen	hɔ̂ːŋkʰruːa
at	[dmnst-det]	[copula]	kitchen
[72a.] "This is the kitchen.			

(38)

นั่นห้องพ่อกับแม่"				
นั่น	**ห้อง**	**พ่อ**	**กับ**	**แม่**
nân	hɔ̂ːŋ	pʰɔ̂ː	kàp	mêː
[dmnst-prn]	room	father	and	mother
[72b.] That's mommy and daddy's room.				

4.5.2.8 S-Complement

Structures that enable whole-sentence complements are given the feature S-COMP=*type*, where *type* corresponds to the primary constituent structure type of the complement sentence. In (39), the complement sentence is shown in red; it entails the overall designation S-COMP=CANONICAL.

(39)

ตรงไหนที่เห็นว่ายังไม่ชอบใจเธอก็เข้าไปเสริมแต่ง														
ตรง	**ไหน**	**ที่**	**เห็น**	**ว่า**	**ยัง**	**ไม่**	**ชอบ**	**ใจ**	**เธอ**	**ก็**	**เข้า**	**ไป**	**เสริม**	**แต่ง**
troŋ	năj	tʰîː	hěn	wâː	jaŋ	mâj	tɕʰɔ̂ːp	tɕaj	tʰɤː	kɔ̂ː	kʰâw	paj	sɤ̌ːm	tɛ̀ŋ

at	where?	which	Øs	see	[→comp]	Øs	still	not	like	heart	she	[link]	[asp]	[asp]	reinforce	decorate

[78.] Wherever she saw that it still wasn't pleasing, she went back and enhanced her design.

4.5.3 **Additional Features**

After assigning a primary syntactic type to each annotation unit, an arbitrary number of additional syntactic features could be assigned, based on the presence of orthogonal syntactic phenomena. These features are described in the following sections.

4.5.3.1 Verbal Adjunct Fronting

The text contains numerous instances of fronted verbal adjuncts. Most are temporal modifiers for the verbs of the main clause (40), although prepositional phrases are also found (24). Annotations units with fronted verbal adjuncts are given the feature VERBAL-ADJUNCT=FRONT, and this feature appears in the story for the CANONICAL, PASSIVE, S-COMPLEMENT, and TOPIC-COMMENT constituent structure types. When dropped pronouns are recovered, semantically-rich verbal adjuncts can begin to resemble chained clauses. In (41), if we are willing to elide the covert impersonal subject in the introductory clause, "(If) you saw (her) from afar..." the introductory clause could be considered a verbal modifier for the main verb "resembled," giving an approximate de-gapped reading akin to (allowing for severe abuse of English) "She resembled, *afar-seen-ly*, a doll." As interesting as this is, the sentence is nevertheless given the CANONICAL=CHAIN—rather than the CANONICAL, VERBAL-ADJUNCT=FRONT—annotation.

(40)

สักครู่จึงโผล่พรวดขึ้นมา							
สัก	**ครู่**	**จึง**	**โผล่**	**พรวด**	**ขึ้น**	**มา**	
sàk	kʰrûː	tɕɯŋ	pʰlòː	pʰrûːat	kʰûn	maː	
just	moment	therefore	Øs	visible	suddenly	[asp]	[asp]

[49.] In a moment, he came up again.

(41)

มองดูแต่ไกลเหมือนตุ๊กตาตัวใหญ่ไม่สวมเสื้อ										
มองดู	**แต่**	**ไกล**	**เหมือน**	**ตุ๊กตา**	**ตัว**	**ใหญ่**	**ไม่**	**สวม**	**เสื้อ**	
mɔːŋduː	tɛ̀ː	klaj	mɯ̌ːan	túkkàtaː	tuːa	jàj	mâj	sǔːam	sûːa	
Øs	examine	Ød.o.	from	far	Øs	resemble	doll	[cls]	big	not	wear	shirt

[30.] Seen from afar, she looked like a big doll with no shirt.

4.5.3.2 Serial Predicate Fronting

As with verbal adjunct fronting, fronting of a portion of a predicate that serializes over the single subject of a nuclear sentence is considered independent of that sentence's primary syntactic constituent type. In most cases, SERIAL-PREDICATE=FRONT is found in combination with existential constructions. In (42), I note that the LEFT-DISLOCATED constituent **เดินนำหน้า** "walk lead" is the matrix predicate that might more canonically occur after the subject, as shown in (43).

(42)

เดินนำหน้าเป็นชายตัวโตกว่าเพื่อนตามด้วยเพื่อนชายอีกสามคน														
เดิน	**นำหน้า**	**เป็น**	**ชาย**	**ตัว**	**โต**	**กว่า**	**เพื่อน**	**ตาม**	**ด้วย**	**เพื่อน**	**ชาย**	**อีก**	**สาม**	**คน**
dɤːn	namnâː	pen	tɕʰaːj	tuːa	toː	kwàː	pʰûːan	taːm	dûaj	pʰûːan	tɕʰaːj	ʔìːk	săːm	kʰon
walk	lead	Øs	is	male	self	large	[compr]	friend	follow	by	friend	male	another	three	person

[15.] Walking in front, it was a boy larger than his friends, followed by three more friends, all boys.

(43)

ชายตัวโตกว่าเดินนำหน้าเพื่อนตามด้วยเพื่อนชายอีกสามคน													
ชาย	**ตัว**	**โต**	**กว่า**	**เดิน**	**นำหน้า**	**เพื่อน**	**ตาม**	**ด้วย**	**เพื่อน**	**ชาย**	**อีก**	**สาม**	**คน**
tɕʰaːj	tuːa	toː	kwàː	dɤːn	namnâː	pʰûːan	taːm	dûaj	pʰûːan	tɕʰaːj	ʔìːk	săːm	kʰon
male	self	large	[compr]	walk	lead	friend	follow	by	friend	male	another	three	person

[15.'] A boy larger than his friends walked in front, followed by three more friends, all boys.

In order to informally test whether the existential verb must be overt in this construction, I presented informant 'A' with a copy of (42) with the existential verb deleted. He judged it grammatical only after having to backtrack, a fact

that he volunteered without cue. This is consistent with the "garden-path," pronoun-drop reading, "Øs walked, leading a boy larger than his friends...," where the nuclear subject has turned into a direct object. An overt existential verb blocks this parse with the result that the subject of the sentence becomes a (dropped) dummy expletive, as positioned in (42).

Unusual sentence (44) is the only instance of a fronted serial predicate which does not co-occur with an existential construction. It may in fact be a typographical error in the source.

(44)

เปลือยกิ่งก้านอาบแดดอยู่เคร่งขรึม							
เปลือย	กิ่ง	ก้าน	อาบ	แดด	อยู่	เคร่งขรึม	
plɯːaj	kìŋ	kâːn	ʔàːp	dɛ̀ːt	jùː	kʰrêŋkʰrɯ̌m	
strip	branch	stem	bathe	sun	[asp]	solemn	

[8.] Bared, the branches bathed in solemn sunlight.

4.5.3.3 Existential

We saw that some instances of existential constructions are associated with a particular type of predicate fronting. Others occur in simple CANONICAL sentences, with either an overt existential verb (45) or a predicative adjective applied to directly to the dummy expletive (46). Both receive the designation SUBJECT=DUMMY. In the study corpus, the dummy pronoun is always dropped in existential constructions, so the annotation is always accompanied by SUBJECT=DROP.

(45)

มีเสียงคึก ๆ ดังมาจากท่าลงน้ำ									
มี	เสียง	คึก ๆ		ดัง	มา	จาก	ท่า	ลง	น้ำ
miː	sǐːaŋ	kʰɯ́kkʰɯ́k		daŋ	maː	tɕàːk	tʰâː	loŋ	náːm
Øs	[exist]	sound	clamor	loud	come	from	harbor	descend	water

[60.] There was a loud clamor from the river landing.

(46)

เงียบเหงาลมสงัด				
เงียบ	เหงา	ลม	สงัด	
ŋîːap	ŋǎw	lom	sàŋàt	
Øs	quiet	lonely	wind	tranquil

[9.] It was quiet and the wind was still.

This concludes the review of the syntax annotation methodology. I now turn a discussion of the information structure annotation methodology.

4.6 Information Structure

Because a goal of the project was to be able to characterize the structural manifestation of TOPIC and FOCUS, it was desirable to adopt guidelines for these features that did not make reference to surface form. This was in order to avoid the illogic of characterizing TOPIC and FOCUS marking based on an annotation of TOPIC and FOCUS marking. Accordingly, it was determined that the criteria for the information structure component of the project would be *subjective discrimination*—by this author—of TOPIC and FOCUS *qualia*: the ineffable mental representations of discourse referents and propositions. There appears to be no alternative toehold for initiating a program of quantitative research in information structure. Since qualia are necessarily represented indirectly, the structural elements which seem to most closely correspond to the TOPIC and FOCUS are what are, in fact, annotated, but it is important to bear in mind that neither these surface forms nor their denotata are intended to be necessarily functionally salient in formulating this component of the annotation. Concerns about including a subjective component in the methodology are somewhat allayed by noting that subjective observations are well-contained, and thus easy for other researchers to exclude.

The approach to information structure was strongly influenced by the FOCUS-centric ideas of Lambrecht (1996). To Lambrecht, TOPIC and FOCUS are obliquely-related pragmatic relations where three well-motivated subtypes of FOCUS have primacy. Deviations from this approach arose as a result of practicalities of application, as discussed in the following sections.

I accede at times in this paper to the convention of annotating surface constituents with square brackets and subscripted TOPIC and FOCUS indications. The system has patent inadequacies; for example, "a referent which is topical in a discourse is not necessarily coded as a TOPIC expression in a given sentence or clause" (Lambrecht 1996: 130); and there is no reason to assume that the surface representations must be contiguous.

4.6.1 **Focus**

Lambrecht's three-way categorization of FOCUS (Lambrecht 1996: 223) is used in this project. This approach being FOCUS-centric, it was straightforward to assign one of his three FOCUS types to each annotation unit. The annotation procedure involved multiple sequential reading passes through the story. Unresolved annotation units from earlier passes were identified as either: elaborating what an identifiable TOPIC referent *did* (PREDICATE FOCUS, *ibid.*: 226); resolving a free variable in the TOPIC RELATION, which is necessarily an open proposition (ARGUMENT FOCUS, *ibid.*: 228); or asserting something entirely non-presupposed (SENTENCE FOCUS, *ibid.*: 233). In Lambrecht's view, all pragmatically felicitous sentences necessarily belong to one of these FOCUS types. In the following subsections, each FOCUS type is discussed in turn. Examples in these sections include TOPIC RELATION annotations, which assume a form that will be explained in Section 4.6.3.

4.6.1.1 Sentence Focus

The first FOCUS category in Lambrecht's scheme concerns *event-reporting* or *presentational* sentence types. When a proposition does not seem to make reference to lexicogrammatical presuppositions (Lambrecht 1996: 233), it is called SENTENCE FOCUS, "wide focus," (Casielles-Suarez 2004: 142) or "all-focus" (Paggio 2009: 145) structure. The first sentence of a text (47) is typically an instance of SENTENCE FOCUS (*ibid.*: 145).

(47)

ลำน้ำนั้นไหลมาจากทิศตะวันตก ทอดตัวเลื้อยเลี้ยวเข้าเขตหมู่บ้าน...

ลำน้ำ	นั้น	ไหล	มา	จาก	ทิศตะวันตก	ทอด	ตัว	เลื้อย	เลี้ยว	เข้า	เขต	หมู่บ้าน
lamnám	nán	lăj	ma:	tɕà:k	tʰíttà?wantòk	tʰɔ̂:t	tu:a	lɯ́:aj	lí:aw	kʰâw	kʰè:t	mù:bâ:n
stream	[dmnst]	flow	[asp]	from	west	drop	self	slither	turn	enter	limits	village

[1.] The stream flowed from the west, dropped down, slithered, turned, and entered the village limits, ...
FOCUS-TYPE: sentence
TOPIC RELATION: x such that x happened
FOCUS: all

The TOPIC RELATION of SENTENCE FOCUS sentences is designated { *x* such that *x* happened }, as many authors have made the observation that SENTENCE FOCUS is a special case of ARGUMENT FOCUS which invokes the "stage topic" (Erteschik-Shir 2007: 16) or "here and now" (Casielles-Suarez 2004: 160).

4.6.1.2 Predicate Focus

Lambrecht calls syntactically CANONICAL sentences, where the proposition can be construed as a comment about the TOPIC referent "topic-comment" sentences (Lambrecht 1996: 226). Because I use this same terminology to refer to the specific syntactic pattern described in 4.5.2.6, I will not adopt this usage, instead referring to sentences which manifest PREDICATE FOCUS as having PREDICATE FOCUS STRUCTURE. Asserting that this type of FOCUS structure is "the most useful pragmatic articulation," Lambrecht concludes that PREDICATE FOCUS is the "most natural... pragmatic construal" for isolated sentences (Lambrecht 1996: 132). Sentence (48) is judged to be in predicate FOCUS.

(48)

หลุมพอ	ต้น	ใหญ่	งอกงาม	ชิด	ตลิ่ง
lŭmpʰɔ:	tôn	jàj	ŋɔ̂:kŋa:m	tɕʰít	tàlìŋ
lumpaw	[cls]	big	sprout	near	bank

[4.] A lumpaw tree sprouted from the bank.
FOCUS TYPE: predicate
TOPIC RELATION: x such that tree did x

4.6.1.3 Argument Focus

In Lambrecht's work, sentences which identify an argument in an open proposition are identificational sentences with ARGUMENT FOCUS. For each instance of ARGUMENT FOCUS, I annotate an open proposition of the type, { *x* such that *x* ... } in the annotation unit's TOPIC field. This proposition informally codes a presupposition that may involve a "topical" discourse entity, but Lambrecht does not admit the use of the term TOPIC here, instead suggesting that a "pragmatic subject"—additionally distinct from the "semantic subject"—is manifested (Lambrecht 1996: 229-230). In his scheme, the *pragmatic subject* and *pragmatic predicate* of (49) are, THE X THAT WAS KEEPING UP THE REAR / WAS A SMALL GIRL, which contrast with the *semantic subject* and *semantic predicate* A SMALL GIRL / KEPT UP THE REAR. The mapping to syntactic constituents is illustrative; the syntactic predicate "keeping up the rear" corresponds to the pragmatic predicate and the semantic subject, while the syntactic subject "a small girl" codes the pragmatic predicate and the semantic subject (*ibid.*: 231). So, in ARGUMENT FOCUS, it is the FOCUS referent which resolves *x* in the open proposition. The information structure annotation for the example is shown below the English translation.

(49)

เหลือ	รั้ง	ท้าย	เป็น	เด็กหญิง	ตัว	เล็ก	ไม่	สวม	เสื้อ
lŭ:a	ráŋ	tʰá:j	pen	dèkjĭŋ	tu:a	lék	mâj	sŭ:am	sû:a
remain	occupy	back	Øs	is	girl	[cls]	small	not	wear	shirt

[15.] Keeping up the rear there was [a small GIRL, wearing no shirt.]$_{FOC}$
FOCUS TYPE: argument
TOPIC RELATION: *x* such that *x* was walking in the rear
FOCUS: girl

For the purposes of this project, TOPIC RELATION is considered "contemporaneous" with its resolution (see Section 4.6.3). This means that the open proposition which is coded as the TOPIC RELATION in an ARGUMENT FOCUS sentence need not be recoverable exclusively from prior discourse context; it has access to information in its in-sentence resolution, which is the FOCUS. That is, in (49), the TOPIC RELATION seems to gratuitously presuppose that someone was "walking in the rear." Such presuppositions may require accommodation of entities or placeholder predicates that may not be entirely established in the preceding discourse. I construe the intentional content of the TOPIC proposition to be instant and holistic with its FOCUS resolution, this pair being interrelated through both their unified syntactic realization and their accessibility in the mind of the speaker, prior to utterance. I now turn to a description of the subtypes of ARGUMENT FOCUS that were encountered in the narrative.

Dialog

The special ability of quoted dialog to partition inner- from outer-structure seems to suggest special information structure treatment for the outer frame itself. Accordingly, I arbitrarily designate quoting frames as an instance of ARGUMENT FOCUS where the TOPIC is { *x* such that *X* said *x* }, and the dialog itself is the FOCUS which resolves this open proposition. Consider sentence [21], which consists of annotation units [21a] and [21b]. Annotation of the outer frame is shown in (50), and in this case, the inner dialog comprises just one annotation unit, which is independently annotated (51). A dialog frame may contain several annotation units.

(50)

"น้ำใสจัง" เด็กชายร่างเล็กพูดขึ้น

น้ำ	ใส	จัง	เด็กชาย	ร่างเล็ก	พูด	ขึ้น
náːm	sǎj	tɕaŋ	dèktɕʰaːj	râːŋlék	pʰûːt	kʰûn
water	clear	extremely	boy	small	speak	[asp]

[21b.] [“The water's really clear,”]$_{FOC}$ a smaller boy said.
FOCUS TYPE: argument
TOPIC RELATION: x such that a smaller boy said *x*
FOCUS: dialog

(51)

"น้ำใสจัง"

น้ำ	ใส	จัง
náːm	sǎj	tɕaŋ
water	clear	extremely

[21a.] “[The water]$_{TOP}$ [is really clear]$_{FOC}$”
FOCUS TYPE: predicate
TOPIC RELATION: x such that the water had the attribute x
FOCUS: clear

Copular resolution

Pragmatic equation can manifest in ARGUMENT FOCUS in a few different ways. The simplest form is the copular presentation of the resolving FOCUS entity for the free variable in the TOPIC proposition (52). I call this the *resolution* subtype of ARGUMENT FOCUS. This subtype also admits examples which use deixis to resolve the TOPIC proposition (53). The copular function, proper, should only enter the FOCUS when it is establishing a relation that is not presupposed (this is discussed in the next section). In (53) both copular arguments are instantiated entities that must sensically be located somewhere (e.g., a stick and “here”), and in (52) one argument is, by construction, a part of the other (e.g., a stream necessarily has a deepest part). Therefore, because these two locative sentences relate two entities that are presupposed to have fixed locations, they are seen as instances of ARGUMENT FOCUS.

(52)

...แต่ตรงที่ลึกที่สุดของลำน้ำช่วงนี้คือวังวนเหนือหาดทรายใต้ต้นหลุมพอนั่นเอง

แต่	ตรงที่	ลึกที่สุด	ของ	ลำน้ำ	ช่วง	นี้	คือ	วังวน	เหนือ	หาดทราย	ใต้	ต้น	หลุมพอ	นั่นเอง
tèː	troŋtʰîː	lúktʰîːsùt	kʰɔ̌ːŋ	lamnám	tɕʰûːaŋ	níː	kʰɯː	waŋwáná	nɯ̌ːa	hàːtsaːj	tâj	tôn	lŭmpʰɔː	nânʔeːŋ
but	place	deepest	[poss]	stream	portion	[dmnst-det]	[copula]	pool	above	beach	under	plant	lumpaw	surely

[41b.] (...but) surely the deepest part of the stream was a [POOL by the beach, under the lumpaw tree.]$_{FOC}$
FOCUS TYPE: argument (resolution)
TOPIC RELATION: *x* such that *x* is the deepest part of the stream
FOCUS: the pool by the beach

(53)

ไม้อยู่นี่"

ไม้	อยู่	นี่
máːj	jùː	nîː
wood	[loc-cop]	[dmnst-prn]

[80b.] The stick is [here!]$_{FOC}$
FOCUS TYPE: argument (resolution)
TOPIC RELATION: *x* such that the stick is discovered to be located at *x*
FOCUS: deictic “here”

Performative

Alternatively, it can be the equation of two entities *itself* that is in FOCUS. In this eventuality, I ascribe ARGUMENT FOCUS with a TOPIC RELATION of the form, e.g., { *x* such that the relation between *y* and *z* is *x* }, and the FOCUS resolves the free variable by designating the *type* of this relation. In this way, the speech act is performative, effecting a new relationship by its utterance. One variation of this pattern that appears in the text is *locative designation* (54), which reflects the *aboutness* notion of “what shall be (performatively) designated where.” Extending the analysis from the previous section, we would expect that this type of designation require that one of the copular arguments be

unbound, that is, that it not carry a presupposition of fixed location, and this is what we see in (54), where one of the copular arguments, "kitchen" is an uninstantiated generic. We could further speculate that *exactly* one of the copular arguments must be unbound, since it seems infelicitous to pin one unbound entity to another, but this takes us a bit afield; in any case, there are no such instances in our small corpus, so we leave the issue for further investigation.

(54)

ตรงนี้เป็นห้องครัว

ตรง	**นี้**	**เป็น**	**ห้องครัว**
troŋ	níː	pen	hɔ̂ːŋkʰruːa
at	[dmnst-det]	[copula]	kitchen

[72a.] THIS is the KITCHEN. [?]$_{FOC}$
FOCUS TYPE: argument (performative designation)
TOPIC RELATION: *x* such that the spatial relationship between "this" location and a kitchen shall be *x*
FOCUS: "one and the same"

In this example, there are no surface elements which readily and obviously summarize the TOPIC or FOCUS. This points out the insufficient rigor of annotating bracketed grammatical constituents as information structure TOPIC or FOCUS. Another issue with bracketing is that, in Thai, a predicative adjective lexeme may need to split its two pragmatic roles between TOPIC and FOCUS. This is discussed at the end of Section 4.6.5.

Corrective

In (55), it is neither a copular argument nor the equating property of a copula sentence that is in FOCUS, but rather the *type* of one of the copular arguments. This subtype is related to Gussenhoven's "corrective focus" (Gussenhoven 2007: 11), although that work does not discriminate the correcting of a *referent entity* from the correcting of the *type of a referent entity*, which is what we see here. The location of the entity as established by the implicit copula is not contested, so the FOCUS is on re-categorizing the type of this entity. This implies a subtle distinction whereby the FOCUS is neither the referent entity, nor instance of a generic type itself, but rather the abstract *type* of the referent entity. In other words, it is not the house referent itself that is new, it is the assertion that the designated entity is, in fact, a *thing* of type "house."

(55)

นี่บ้านฉัน

นี่	**บ้าน**	**ฉัน**
nîː	bâːn	tɕʰǎn
[dmnst-prn]	home	[1s]

[101b.] This is my [HOUSE!]$_{FOC}$
FOCUS TYPE: argument
TOPIC RELATION: *x* such that *x* is the type of my thing that this is
FOCUS: house-thing

Here, although the context of the story informs the corrective reading, two additional readings could be obtained by alternatively FOCUSing the possessive pronoun "my" or the demonstrative pronoun "this." A fourth, performative reading can also be obtained in accordance with the discussion in the previous section. But this example presents an opportunity to discuss an interesting fifth reading, a reading which combines corrective FOCUS with the relationship-focusing aspect of the performative reading. The TOPIC of a performative reading is a type of relationship, which the very utterance of a performative ARGUMENT FOCUS assertion enacts. However, it is also possible to *correct*, rather than *enact*, a relationship expressed in a TOPIC proposition. With *performative* ARGUMENT FOCUS, a speaker changes world-truth so that it matches her discourse model; conversely, with *corrective* ARGUMENT FOCUS, she changes her hearer's discourse model so that it matches her own. In (56), I give two illustrations of this by placing FOCUS on the *correction* of the presupposition that a certain presupposed place in the world does not have the "one and the same" relationship with the presupposed house of the speaker. The TOPIC here, informally-vetted by the *aboutness* test, is found to be: { *x* such that the relationship between the indicated place and the location of your house is *x* } which correctly identifies the FOCUS—the resolution of the free variable in the TOPIC—as being the "one and the same" (or "*is-ness*") relationship.

"(This [deictic entity] isn't your house.) (Actually,) This $[\text{IS}]_{\text{FOC}}$ my house!"

(56) "(Why don't we go [from where we are now] to your house now?) (Actually,) This $[\text{IS}]_{\text{FOC}}$ my house."

This concludes the discussion of FOCUS types. In Section 4.6.4 I will examine the possibility of unifying all three FOCUS types into a monolithic representation which captures all aspects of a proposition's information structure (with the exception of the FOCUS entity itself). Such a generalized form holds great promise for more naturally revealing—rather than imposing—information structural types that may appear in a corpus. First, I will turn to a discussion of the approach to TOPIC adopted for this study.

4.6.2 **Topic Referent**

For TOPIC, the objective of this study was to capture *aboutness* impressions from the *experience* of reading each sentence, and the primary subjective diagnostic for TOPIC is the *aboutness test*, which considers whether a sentence seems to primarily increase the hearer's knowledge "about" the specified referent (Lambrecht 1996: 131).

Many authors assert that a TOPIC REFERENT must be an entity (Gundel 1988: 210). Lambrecht qualifies his account, noting that "TOPIC referents are for the most part not propositions but entities" (Lambrecht 1996: 151). One difficulty is that different authors use the term to variously refer to both inter- and intra-sentential referents. Büring (1999: 145) is careful to avoid this confusion, designating the former D-TOPIC ("discourse") and the latter S-TOPIC ("sentence"). Carefully observing this dichotomy clarifies much apparent contradiction in the literature. For example, Büring suggests that S-TOPICS can contrast ("contrastive"), narrow down ("narrowing"), or be in part-whole relationship ("partial/residual") with a D-TOPIC (*ibid.*). Identifying the TOPIC subtypes that Büring proposes thus appears to require maintaining a running tabulation of possibly overlapping discourse- and sentence-level TOPICS. While this approach may indeed capture interesting insights, it is left for future elaboration; I instead elected to hew to Lambrecht's observations which are described next.

Lambrecht is less concerned with DISCOURSE TOPIC and attempts to characterize S-TOPIC in isolation, while noting that the concept is fundamentally intertwined with discourse context:

> Strawson's remark that 'there is great variety of possible types of answer to the question what the topic of a statement is' stresses the inherently vague character of the notions of aboutness and relevance. This inherent vagueness has consequences for the grammatical coding of topics in sentences... It is often necessary to take into account the discourse context..., whether the [referent is] 'a matter of standing current interest or concern' (Strawson)..., the communicative intentions the speaker had..., and the state of mind of the addressee with respect to the referent. (Lambrecht 1996: 119-120)

He continues by noting that (S-)TOPIC and FOCUS, as succinctly summarized by another author, "do not form a dichotomy" and are not complementary (Casielles-Suarez 2004: 158). In particular, he explains that his three FOCUS types—PREDICATE FOCUS, ARGUMENT FOCUS, and SENTENCE FOCUS—imply certain "information structure categories," namely: unmarked "topic comment" sentences, identificational sentences, and event-reporting sentences, respectively (Lambrecht 1996: 121-127). Another type, the "background establishing" sentence, may contain a

> non-topical or 'semi-topical' expression which appears within a sentential scene-setting topic expression, which itself [may be] embedded within a matrix clause whose subject is the primary topic of the sentence. (*ibid.*: 125)

Lambrecht's FOCUS types not only imply certain TOPIC "information structure categories," but prototypical TOPIC RELATIONS as well For example, SENTENCE FOCUS, it is claimed, has no readily identifiable TOPIC (*ibid.*: 124). I continue in the next section with a more detailed examination of TOPIC RELATION, which introduces the form for the annotation of TOPIC that was adopted for this project.

4.6.3 **Topic Relation**

Lambrecht introduces the concept of the TOPIC RELATION: the *aboutness* relation that holds between TOPIC REFERENTS and propositions (Lambrecht 1996: 130). He is careful to distinguish TOPIC RELATION from TOPIC REFERENT

(and TOPIC EXPRESSION), and notes that, "while a TOPIC expression always... designates a TOPIC referent, a referent which is TOPICAL in a discourse is not necessarily coded as a TOPIC expression" (Lambrecht 1996: 130). Despite his care, in places, confusion can arise from ambiguous uses of the word, "topic." For example, even in his ultimate definition, he seems to first characterize "topic" as a referent, only to immediately call it a relation:

> *A referent is interpreted as the topic of a proposition* if IN A GIVEN DISCOURSE the proposition is construed as being ABOUT this referent, i.e. as expressing information which is RELEVANT to and which increases the addressee's KNOWLEDGE of this referent. Following Reinhart (1982), we may say that the relation "topic-of" expresses the pragmatic relation of aboutness which holds between a referent and a proposition with respect to a particular discourse. The term "pragmatic relation" should be understood as meaning "relation construed within particular discourse contexts." *Topic is a pragmatically construed sentence relation*. (Lambrecht 1996: 127, emphasis added)

Since the primary criterion for TOPIC annotation was *aboutness*, and since this citation links *aboutness* with TOPIC RELATION, an annotation of *aboutness* primarily entails an annotation of TOPIC RELATION, and not TOPIC REFERENT. Furthermore, since TOPICS cannot function in a proposition without entering a TOPIC RELATION (*ibid.*: 151, 130), there is merit in adopting TOPIC RELATION as the object of TOPIC annotation. Lambrecht seems to agree; regarding his example, "(Who went to school?) The CHILDREN went to school," he notes that

> the best candidate [for a topic is] the presupposed open proposition "[*x* such that] *x* went to school," concerning which the asserted proposition can be said to add a relevant new piece of information (Lambrecht 1996: 122).

Table 1 illustrates some examples of the application of this view of TOPIC RELATION, along with the corresponding FOCUS, and suggests that the phenomena are not entirely independent. This is the idea that will be developed further in the remainder of this section.

Assertion	**TOPIC RELATION**	**FOCUS**
Prawit ran.	*x* : Prawit did *x* *x* : *x* ran	running-activity Prawit (referent)
Prawit RAN.	*x* : Prawit did *x*	running-activity
STEVE ran.	*x* : *x* ran	Prawit (referent)
Prawit's SKINNY.	*x* : Prawit has the attribute *x*	skinny-attribute
Prawit's DEAD.	*x* : Prawit has aliveness-status *x*	dead-aliveness-status
Prawit ran to the PARK.	*x* : Prawit ran to *x*	park (referent)
They HAD great fun.	*x* : the *corrected* event-status of their great-fun-activity is actually *x*	completed-event-status
They had GREAT fun.	*x* : their fun had attribute *x*	great-attribute
This is my HOUSE!	*x* : the type of thing that "this" entity of mine is, is *x*	house-type
This is MY house!	*x* : the relationship between myself and this house is *x*	ownership-relationship
THIS is my house.	*x* : the one house, selected from a set, that belongs to me is *x*	[deictic] (referent)
THIS is my HOUSE.	*x* : the *performative* relationship between the deictic location and my house is *x*	one-and-the-same-relationship
This IS my house.	*x* : the *corrected* relationship between the deictic location and my house is *x*	one-and-the-same-relationship

Table 1. TOPIC RELATION examples

The approach to TOPIC that was adopted was that of annotating a single open proposition which captures its presuppositional structure and its TOPIC referents; TOPIC annotations for this project were constructed according to this form. In the adopted approach, the TOPIC quale of an asserted proposition is stated in terms of a proposition with a free variable which is taken to be contemporaneous with the assertion. TOPIC RELATIONS expressed in this way express subtypes of aboutness. The reference definition used for this project is:

TOPIC RELATION is an open proposition that expresses its canonical, contemporaneous presupposition, by using FOCUS-exclusive predications to express relationships between one[8] free variable and one or more FOCUS-exclusive bound variables.

The intuitive view of TOPIC RELATION is that it is the traditional TOPICAL REFERENT of a sentence, *plus* a relation or set of relations that describe the role that the TOPIC REFERENT is "presupposed to play... in the given proposition" (Lambrecht 1996: 151). In an ARGUMENT FOCUS proposition, the TOPIC relation may contain a non-FOCUSED predicate from the proposition itself. In this view PREDICATE FOCUS and SENTENCE FOCUS propositions are expessed in the same way, and one of several generic placeholder predicates is always imputed. This will be examined later on. TOPIC RELATIONS expressed in this annotation project have the form shown in (57).

(57) "(What did Supaphan ride?) Supaphan rode the $[\text{BUS}]_{\text{FOC}}$." { *x* such that Supaphan rode *x* }

I would be remiss in proceeding without mentioning that, while Lambrecht's idea of stating TOPIC in terms of a presupposed open proposition was exciting, *alas*, he presented it in the context of a straw-man argument. Since his idea was adopted for this project I must briefly attempt to rebut Lambrecht's arguments for dismissing it.

His first argument against the use of propositions with a free variable—that is, propositions that I denote TOPIC RELATIONS—to represent TOPICS is that such propositions are semantically incomplete and so "cannot be said to have [referents, and] therefore [that] the asserted proposition cannot be construed as being *about* its referent" (Lambrecht 1996: 122). Firstly, it's not clear that the free variable in "*x* such that *x* went to school" is *not* a referent, any less so than "Supaphan," just because it's semantically unresolved. Considering that some types of referents are semantically compatible with the resolution of *x* and other types are not, it is not the case the free variable is completely unconstrained. Furthermore, defining the TOPIC RELATION as existing contemporaneously with its assertion—which is justified on the basis of the fact that, to the speaker of an assertion, there is no dichotomy between "new" and "old" information (more specifically, that a speaker begins an utterance with the presumption that he will have timely access to everything he will need to complete it)—enables full access to the resolution arguments of *all* free variables.

It is also possible to rebut Lambrecht's conclusion that these considerations disqualify "the asserted proposition [from being] construed as being about its referent" by noting that, in the absence of the very sort of open proposition he disqualifies, an asserted proposition cannot be construed as being about *a referent* either. Lambrecht admits as much when he notes that TOPIC RELATION "*aboutness*" must hold between a TOPIC REFERENT and its proposition (*ibid.*: 127). In short, without the assistance of being incorporated into a TOPIC RELATION, where it necessarily becomes associated with some form of predicate (see Section 4.6.5), it seems impossible for a TOPIC REFERENT, in isolation, to function as a "topic" for *any* proposition.

Finally, Lambrecht writes that "since the presupposition cannot be identified with a syntactic constituent..., there is no structural element which can be identified as a TOPIC EXPRESSION" (*ibid.*: 122). To this point, I only note that it is not clear to me why the requirement should be necessary.

4.6.4 **Unification of Focus Types**

In 4.6.1.1, we saw that SENTENCE FOCUS is a special case of ARGUMENT FOCUS. But it is also the case that, when much of a predicate can be contextually presupposed, PREDICATE FOCUS can be stated in terms of ARGUMENT FOCUS. In (58), it is already well-established in the story that the boys are playing, which entails that the activity was necessarily being perceived as any of: fun, tedious, monotonous, unpleasant, etc. Even though the boys are coded with a pronoun, the sentence is not really *about* them (more on this follows below). Neither is the sentence *about* what the boys were doing, but it is rather *about* a characterization of an activity that we already know they are engaged in. The new information that is conveyed by this sentence is the "narrowing" of the numerous possibilities (Büring 1999: 145) by selection of the first as the resolution of the ARGUMENT FOCUS TOPIC RELATION { *x* such that the (presupposed) playing had the attribute *x* }. If we contrast this with the prototypical PREDICATE FOCUS TOPIC RELATION {

[8] Multiple-FOCUS assertions are not investigated.

x such that *x* did *y* } we see that ARGUMENT FOCUS is just more specifically isolating an attribute than PREDICATE FOCUS, which tends to conflate action with its attributes.

(58)

พวกเขากำลังสนุกเต็มที่			
พวกเขา	**กำลัง**	**สนุก**	**เต็มที่**
pʰûːakkʰăw	kamlaŋ	sànùk	temtʰîː
they	[asp]	fun	fully

[56.] They were having a [really-fun]$_{FOC}$ time.
FOCUS TYPE: argument
TOPIC RELATION: *x* such that their playing had attribute *x*
FOCUS: fun

This discussion suggests that Lambrecht's range of FOCUS types describes a cline according to the universality of the required presupposition. For SENTENCE FOCUS, where we willingly accommodate, at any time or place, the TOPIC RELATION { *x* such that *x* happened }, we are actually eliding the TOPIC because it is so familiar and common; it is not necessary to say, "[As for what happened,]$_{TOP}$ [Prawit ran after the bus,]$_{FOC}$" because our experience suggests that, in utterances, things always happen—otherwise there is nothing to utter. Although such a TOPIC may seem gratuitous since it incorporates nothing motivated by the surface form, it certainly has pragmatic motivation and does embody a characteristic which may facilitate generalization in the discussion below.

Comparing PREDICATE FOCUS to SENTENCE FOCUS, I suggest that it is only slightly less universal to infer { *x* such that Prawit did *x* }, than it is to infer { *x* such that *x* happened }, because our cognitive sensibility seems to support a fundamental distinction between entities and events, such that the former chronically participate in the latter. To say that referent "Prawit" is the TOPIC in "Prawit [RAN after the bus]$_{FOC}$," is inconsistent with our predilection for accommodating the presupposition that subjects typically *do* things, because it excludes from the TOPIC this inherent characteristic of subjects. In this view, the TOPIC in the PREDICATE FOCUS example is more properly stated, { *x* such that Prawit did *x* } which also intuitively agrees better with our primary annotation heuristic for TOPIC RELATION, namely the *aboutness* test. This sentence is clearly *about* "what Prawit did;" it's not just *about* Prawit. That he took *some* action appears to be necessarily presupposed, and TOPIC *RELATION* more accurately captures *aboutness*.

Lastly on this cline, ARGUMENT FOCUS is the least-universal type because it incorporates the greatest amount of situationally-specific material into its TOPIC RELATION. In "Prawit ran after [the BUS]$_{FOC}$," I agree with Lambrecht's characterization of the proposition as { *x* such that Prawit ran after *x* }.

To summarize these observations, I note the number of situationally-specific elements that are bound into each TOPIC RELATION: the SENTENCE FOCUS example makes reference to none; for PREDICATE FOCUS there is one ("Prawit"), and with ARGUMENT FOCUS there are two ("Prawit" and "run"). If a generalizable relationship does hold between FOCUS (as captured by Lambrecht's three categories) and TOPIC RELATION (when the latter is uniformly expressed as a proposition with free and bound variables), then Lambrecht's approach of de-emphasizing TOPIC in favor of FOCUS-centricity has appeal. It appears that, when universally applicable presuppositions are made explicit, a single approach to formulating TOPICS in terms of free and bound variables may sufficiently characterize information structure.

4.6.5 **Obligatory Predicative Cues in Topic Relation**

The harmonization of Lambrecht's well-motivated FOCUS types discussed in the previous section leads to the conclusion that felicitous proposition must always entail, at a minimum, the presupposition that *something happened*. It may not be possible for a sentence to be "about" a referent without additionally presupposing, at a minimum, a degenerate form of implicit predication, a result that would explain the infelicity of bare NPs in many languages. This argument was introduced in the individual sections on FOCUS subtypes (4.6.1.1 - 4.6.1.3) and unification of FOCUS subtypes (Section 4.6.4); in this section, I detail it further.

Examining the range of open proposition forms in the FOCUS type cline that was developed in 4.6.4, we saw that the number of bound variables in the increases from zero, with SENTENCE FOCUS; to one, with PREDICATE FOCUS; and finally to two, with ARGUMENT FOCUS. This corresponds to SENTENCE FOCUS requiring the least presupposition and ARGUMENT FOCUS the most. Note, however, that these figures represented the number of bound variables

that explicitly appeared in the surface forms. For example, the SENTENCE FOCUS TOPIC RELATION { x such that x happened } still has exactly one predicate, "happened," and I have attempted to show throughout the analysis of information structure, that all TOPIC RELATIONS necessarily have a predicate.

A different approach towards establishing this same conclusion begins by noting that, in more elaborate forms of presupposition, entities have properties, states, or attributes or do things, but they never do *nothing*. When things do nothing, it is not worthy of comment and further, it appears that it is not possible to comment, without framing the occurrence of nothing as "something happening." If all utterances necessarily have a FOCUS, as Lambrecht believes, then we cannot construct an utterance without a (recoverable) predicate. This confirms the observation that bare NPs in isolation are not felicitous in English and Thai (at least). In other words, for an assertion to have a simple REFERENT as a TOPIC necessarily presupposes that the TOPIC will either be in a state, have or acquire properties or attributes, or do something.

Lambrecht's diagnostic for PREDICATE FOCUS hints at this; I have added emphasis to his diagnostic to illustrate my point: "(What did the children *do* next?) The children went to SCHOOL" (Lambrecht 1996: 121). That all TOPICS entail some sort of predicate is also suggested by Lambrecht's basic premise that all felicitous utterances must present new information. According to this, if nothing happened, then there's nothing to say, so the very existence of an utterance implies that something happened. Even with an alternative diagnostic such as, "What about the children?" one cannot hide the presupposition that the children either did something, were in some state, or had some property[9], because this very diagnostic is a concise *aboutness* test for the assertions "They went [to school]$_{FOC}$," "They were [content]$_{FOC}$," or "They were [overweight]$_{FOC}$." These sentences are not *about* the children but rather are about what the children *did* or *how they were*.

This discussion is easily extended to states signaled by a stative; the diagnostic "What about Prawit?" is felicitous for the assertion "He's dead," which here confirms the TOPIC RELATION { *x* such that Prawit is in the state of aliveness *x* }. Clearly the assertion is *about* Prawit's state, and not Prawit himself. In other words, it is so overwhelmingly likely that subjects do things and have states and attributes, that any attempt to conceive of a PREDICATE FOCUS proposition which has its syntactic subject as its pragmatic TOPIC morphs into an ARGUMENT FOCUS assertion of the form { *x* such that *X* is in state *x* }, { *x* such that *X* has the property *x* } or { *x* such that *X* did *x* }, and this latter form is—somewhat arbitrarily—dubbed PREDICATE FOCUS.

Therefore, for this project, if a predicate is not available in the "old" information, i.e., in the case of PREDICATE FOCUS where the assertion's predicate is reserved for the resolving the free variable as the FOCUS of the assertion, then a well known placeholder is imputed. Contemporaneousness allows a TOPIC RELATION to be formulated while taking into account predicative cues may be inaccessible to the hearer prior to the time of utterance, in effect allowing it to "peek" at the FOCUS predicate, to ensure that the imputed placeholder is semantically compatible but more general than the free variable (FOCUS) predicate.

Recall the proposed the diagnostic context "(What about the children?)...," which reveals no information about the intended form of the proposition. The speaker clearly knows what she intends to say prior to saying it, and it is this fact that permits the principle to assert that the TOPIC RELATION is singular and well-formed. The result is that the TOPIC RELATION—by construction—contains a predicate that is compatible with the free variable predicate. I also showed in that paragraph how "do" is more general than "go," as shown by the diagnostic pair: "(What did the children do?) They [went to the zoo]$_{FOC}$."

It is not an arbitrary requirement that TOPIC RELATIONS have a predicate, for it seems impossible to craft an utterance that does not—at least collaterally—invoke an *unfocused* predicate, that is, that does not place into the presupposition either: a predication that is already in the discourse common ground and which it is not the goal of the utterance to elaborate (i.e. *old*, as in "It was [MIEW]$_{FOC}$ who ran," where the TOPIC RELATION { *x* such that *x* ran } makes reference the predicate 'ran'), or a placeholder predication that performs the crucial role of situating the TOPIC referents with respect to each other and the free variable. The actual FOCUS of the proposition substitutes for the placeholder by answering the felicitous diagnostic question: "Who was it who ran?"

---

[9] It's possible that, in adopting *aboutness* as our test for topichood, we have unintentionally incorporated a pragmatic quirk of the word "about" into our approach, since it's not obvious why the same word should pragmatically permit all of—states, attributes, and actions—to be associated with entities.

Well known placeholder predications take the form: { *x* such that *X* did *x* }, { x such that *X* has attribute x }, { x such that *X* is in state x }. We should not let the SENTENCE FOCUS diagnostic "What happened?" mislead us here, as it conflates these placeholders, giving the impression that TOPIC RELATION'S predicative cue is perhaps unavailable or non-existent, a problem whose solution is again provided by the contemporaneousness principle: the fact that a unique placeholder can always be recovered from the utterance after the fact seems to suggest that the speaker must have intended to disambiguate "What happened?" along one of these lines. In any case, I was not able to conceive an assertion that felicitously answers "What happened?" without selecting (or implying) exactly one placeholder predication.

This analysis may be especially important in Thai, where predicative adjectives allow the syntactic conflation of actions with attributes. We saw this in (58), a predicative adjective sentence, where the single word สนุก /sànùk/ "fun" encodes two pragmatic functions: first, the experience by some subject of an activity, and second, the attribute that this activity is enjoyable. In ARGUMENT FOCUS, only the latter function should be associated with the FOCUS, while the predicative pragmatic function should be in the TOPIC RELATION. Since the word is a single lexeme, such a designation cannot be approximated by bracketing surface constituents in the Thai orthography. And one last unrelated point on the pitfalls of bracketing: in (55), we see that the designation of the lexeme "house" as "FOCUS" encourages the misperception that the FOCUS quale corresponds to the referent, a particular sand house (or pile of sand) in the world, when it in fact invokes neither, but rather only the set of qualities and properties of that we associate with "house-*ness*."

5 Summary and Results

This concludes the methodology review of this project, the fully supervised (manual) annotation of a small Thai-English bitext for large-scale syntactic structure and information structure. The corpus that was developed for this project is available at the website http://www.thai-language.com/id/590220/information-structure. The materials include the full English and Thai texts of the story, and the breakdown of each annotation unit with its accompanying annotations. This small gold standard corpus may have wide application as a training or evaluation resource in a broad range of Thai language natural language processing (NLP) tasks. Work is already underway to use the corpus as a reference in the development of a new approach to automatic sentence alignment, a critical task in statistical machine translation.

		focus type		
primary syntactic constituent type	**N**	**argument**	**predicate**	**sentence**
*topic-comment (all)	7	4.4824	0.7619	1.0388
*s-comp (all)	4	1.1448	0.3333	0.1502
*copula (all)	11	19.5834	3.6667	4.1852
*canonical (all)	237	0.0101	0.0506	0.0152
*benefactive	4	1.1448	0.3333	0.1502
*imperative	8	2.2896	2.0417	0.0006
*interjection	9	2.5758	1.3333	6.1145
*passive	4	0.6389	1.3333	0.1502
*question	13	3.7205	1.6410	0.2245
		85	99	113

Table 2. Focus type manifestation of primary syntactic constituent types

Having independently annotated the text for syntactic features and information structure, correlative studies were possible. The dataset enables many more quantitative studies than were possible within the scope of this work. Here, I investigate only the broadest characteristics in the data, and it is hoped that the dataset will be used by other researchers in Thai syntax, information structure, and statistical applications to test their hypotheses and further develop their respective fields.

The primary investigative methodology used for this analysis is a chi-squared correlation of syntactic features with information structure phenomena. In the first chart, the primary syntactic constituent structure types are displayed with chi-squared values against each of the three Lambrecht FOCUS types, as judged. Each annotation unit is represented exactly once in this chart. We note strong correlation between ARGUMENT FOCUS and the COPULA pattern. We also see high correlation between the TOPIC-COMMENT pattern and ARGUMENT FOCUS.

Table 3 shows *ad-hoc* subject features versus FOCUS type. Dummy subject appears to imply ARGUMENT FOCUS. A surprising result is that of the impersonal pronoun, but we note that N=3. Surprisingly, overt pronouns don't seem to be predictive in this corpus. Pronoun drop is slightly biased for ARGUMENT FOCUS.

		focus type		
subject feature	**N**	**argument**	**predicate**	**sentence**
subject=back	1	0.3184	0.4469	2.4965
subject=drop	86	3.2166	1.1211	0.3944
subject=dummy	8	7.7821	3.5754	0.4098
subject=gerund	1	0.3184	0.6844	0.2346
subject=impersonal	3	0.9553	1.3408	7.4896
subject=left	1	1.4588	0.4469	0.2346
subject=pronoun	59	1.4461	0.0052	1.6947
subject=right	1	0.3184	0.4469	2.4965
subject-2=drop	13	0.0047	0.0062	0.0008
subject-2=dummy	6	0.6211	0.1732	0.1181
		57	80	42

Table 3. Focus type manifestation of ad-hoc subject features.

Table 4 isolates the movement features for specific study. The feature verbal-adjunct=front strongly biases for predicate focus. Table 5 shows that, within the TOPIC-COMMENT syntactic pattern, subject FRONTING, as marked by the fronting particle, favors ARGUMENT FOCUS. Finally, Table 6 summarizes the information structure manifestations of the remaining grammatical and specialty features that were annotated in the study corpus.

		focus type		
movement features	**N**	**argument**	**predicate**	**sentence**
*topic-comment (all)	7	0.0570	0.0136	0.1059
object=front	33	6.5890	5.2963	6.5185
prep-phr=front	4	0.0727	0.6420	0.0557
serial-predicate=front	5	0.0137	0.8025	1.0377
subject=back	1	0.6420	0.1605	3.2600
verbal-adjunct=front	29	7.2492	11.5933	3.1853
subject=left	1	0.1997	0.1605	0.1975
subject=right	1	0.6420	0.1605	3.2600
		52	13	16

Table 4. Focus type manifestation of TOPIC-COMMENT subtypes.

		focus type		
topic-comment subtypes	**N**	**argument**	**predicate**	**sentence**
*topic-comment=possessor	1	0.1143	0.1429	0.1429
*topic-comment=secondary-cause	1	0.1143	0.1429	0.1429
*topic-comment=subject	5	0.0914	0.1143	0.1143
		5	1	1

Table 5. Focus type manifestation of (syntactic) TOPIC-COMMENT subtypes

		focus type		
grammatical, specialty	**N**	**argument**	**predicate**	**sentence**
*s-comp=*canonical	3	0.7759	1.0529	0.0469
*s-comp=*copula	1	0.2586	0.3276	0.8305
*passive	4	0.9011	1.3103	0.0718
*benefactive	4	1.0345	0.3630	0.0718
*copula	5	10.6264	1.6379	2.0690
*copula=drop	6	12.7517	1.9655	2.4828
*imperative	8	2.0690	2.1602	0.0291
*interjection	9	2.3276	1.2875	4.9093
*question	13	3.3621	1.7647	0.0716
superlative	4	0.9011	0.0735	0.2593
subject=gerund	1	0.2586	1.3802	0.4138
		15	19	24

Table 6. Focus type manifestation of grammatical features and breakdowns of S-COMPLEMENT and COPULA subtypes.

6 Conclusion and Future Work

This work suggested many interesting sub-projects. A deeper and more rigorous methodology for Thai syntactic annotation, facilitated by a precision grammar of the language is a priority for future work. An oversight was the omission of an *ad-hoc* feature to record NEGATION. This would have been valuable because it appears that NEGATION obtains special ARGUMENT FOCUS treatment. A more rigorous corpus-based characterization of the Thai FRONTING marker would be helpful. Further work is also needed to develop a consistent approach to Thai predicative adjectives with regard to ARGUMENT FOCUS *versus* PREDICATE FOCUS treatment.

In this study, correlations between syntactic and information-structural phenomena were confirmed in a small text-only Thai-language corpus. It seems reasonable to speculate that authors who produce texts—in any language—that are not intended to be read aloud necessarily draw more heavily on syntactic information structure effects, amplifying their occurrence in a way that ultimately serves to motivate the evolution of a distinct register that we recognize as unnatural for spontaneous spoken communication. For this reason, it would be interesting to compare the results of this study with comparable syntactic phenomenon figures for a corpus of *spoken* Thai.

As for cross-linguistic information structure, it would be interesting to continue study the sufficiency and utility of the unified predicate-bearing TOPIC representations. Particular attention should be paid to integrating SENTENCE FOCUS with the other focus subtypes. One conclusion of this work is that a comprehensive unified approach to information structure annotation which proceeds along these lines may be within reach.

Related work would examine more carefully the idea that it is impossible for an assertion to be "about" a non-predicated "naked" TOPIC REFERENT, on the basis that such an entity has no TOPIC RELATION to mediate its appearance in the assertion. This summarizes the view presented here that only a TOPIC RELATION (open proposition) may bear *aboutness*. I also attempted to connect this idea to an interdependent argument, namely, that any utterance implies that *something happened*, and conversely, if nothing has happened, there is nothing that can be said "about" that, without invoking an artificial frame around it. While fanciful, such work might yet capture a powerful generalization.

References

W. Aroonmanakun. 2007. Thoughts on Word and Sentence Segmentation in Thai. *Proceedings of the Seventh International Symposium on Natural Language Processing*,Pattaya, Thailand, 85-90.

Daniel Büring. 1999. Topic. In P. Bosch & R. van der Sandt (eds.), *Focus: Linguistic, cognitive, and computational perspectives*, 142–165. Cambridge: Cambridge University Press.

Somsonge Burusphat. 2002. Discourse studies in Thailand. *Discourse Studies* 4.4: 501-510.

Eugenia Casielles-Suarez. 2004. *The Syntax-Information Structure Interface*. New York & London: Routledge.

Wallace Chafe. 1976. Givenness, Contrastiveness, Definiteness, Subjects, Topics and Point of view. In C. N. Li, (ed.), *Subject and topic,* 25-55. New York: Academic Press.

Elisabet Engdahl, & Enric Vallduví. 1996. Information Packaging in HPSG. In Claire Grover & Enric Vallduví, (eds.), *Studies in HPSG (Edinburgh Working Papers in Cognitive Science 12)*, 1–32. Edinburgh: Centre for Cognitive Science, University of Edinburgh.

Nomi Erteschik-Shir. 2007. *Information structure: The syntax-discourse interface*. Oxford: Oxford University Press.

Tom Glass, (tr.). n.d. "*Building Sandpiles" by Paitoon Thanya*. *http://crcl.th.net/bitext/sand-bi.htm* Bangkok: Center for Research in Computational Linguistics.

Jeanette K. Gundel. 1988. Universals of Topic-Comment Structure. in Michael Hammond, Edith Moravcsik, and Jessica Wirth, (eds.), *Studies in Syntactic Typology*, 209–239. Philadelphia: John Benjamins.

Carlos Gussenhoven. 2007. Types of Focus in English. In C. Lee, M. Gordon, & D. Büring (eds.), *Topic and focus: Cross-linguistic perspectives on meaning and intonation*, 83-100. Dordrecht: Kluwer Academic Publishers.

Shoichi Iwasaki & Preeya Ingkaphirom. 2005. *A reference grammar of Thai*. New York: Cambridge University Press.

Kristiina Jokinen. 2005. A Grammar for Finnish Discouse Patterns. In Antti Arppe, et al., (eds.), *Inquiries into Words, Constraints and Contexts: Festschrift for Kimmo Koskenniemi on his 60th Birthday*, 227-240. Stanford: CSLI Publications.

Jean-Pierre Koenig & Nuttanart Muansuwan. 2000. How to End Without Ever Finishing: Thai Semi-Perfectivity. *Journal of Semantics*. 17, 147-184.

Jean-Pierre Koenig & Nuttanart Muansuwan. 2005. The Syntax of Aspect in Thai. *Natural Language & Linguistic Theory*. 23 (2), 335-380.

Knud Lambrecht. 1996. *Information Structure and Sentence Form* (Cambridge Studies in Linguistics). Cambridge: Cambridge University Press.

Amy Meepoe. 1997. Establishing and Maintaining Temporal Frames in Thai Conversational Discourse. *Proceedings of the Seventh Meeting of the Southeast Asian Linguistics Society (SEALS VII).*

Amy Meepoe-Baron. 1998. Tying techniques in Thai: The practices of personal reference. *Crossroads of Language, Interaction, and Culture*, 1:59-69. Los Angeles: Regents of the University of California.

Nuttanart Muansuwan. 2002. Verb complexes in Thai. Thesis (Ph. D.)—State University of New York at Buffalo, 2002.

Nuttanart Muansuwan. 2001. Directional serial verb constructions in Thai. *Proceedings of the 7th International HPSG Conference, UC Berkeley (22-23 July, 2000)*. Dan Flickinger and Andreas Kathol, (eds). Stanford: CSLI Publications.

Patrizia Paggio. 2004. The information structure of Danish grammar constructions. *Nordic Journal of Linguistics* 32.1, 137-164.

Carl Pollard & Ivan A. Sag. 1994. *Head-driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Unchalee Singnoi. 2002. An analysis of passive constructions in Thai. In Dawn Nordquist and Catie Berkenfield, (eds.), *Proceedings of the Second Annual High Desert Linguistics Society Conference, 1999*, 69-86. Albuquerque, New Mexico: High Desert Linguistic Society.

Unchalee Singnoi. 2004. Discourse Functions of Thai Demonstratives: a Case On Pragmatically Controlled Irregular functions. In Somsonge Burusphat, (ed.), *Papers from the Eleventh Annual Meeting of the Southeast Asian Linguistics Society (SEALS XI)*, 645-657. Tempe, Arizona: Arizona State University, Program for Southeast Asian Studies.

Unchalee Singnoi. 2007. On the Pragmatic Functions of 'Marked Clausal Constructions' in Thai. In *Papers from the Seventeenth Annual Conference of the Southeast Asian Linguistic Society (SEALS-XVII), College Park, MD*. (in press).

David Smyth. 2002. *Thai: An essential grammar*. London: Routledge.

Paitoon Thanya [ไพฑูรย์ ธัญญา]. 1955. *Building Sandpiles (abridged)* [ก่อกองทราย (ย่อ)]. *http://203.172.241.178/chaiyaporn/thai/literature3.doc* (accessed on February 26, 2010).

Kingkarn Thepkanjana. 1986. Serial verb constructions in Thai. Thesis (Ph. D.)—University of Michigan, 1986.

Graham Wilcock. 2005. Information structure and minimal recursion semantics. In Antti Arppe, et al., (eds.), *Inquiries into Words, Constraints and Contexts: Festschrift for Kimmo Koskenniemi on his 60th Birthday*, 268–277. Stanford: CSLI Publications.

Udom Warotamasikkhadit. 1979. Thai Sentence Focus. In Theraphan L. Thongkum et al., (eds.), *Studies in Tai and Mon-Khmer Phonetics and Phonology In Honour of Eugénie J.A. Henderson*, 313-324. Chulalongkorn University Press.

Udom Warotamasikkhadit. 1997. Fronting and backing topicalization in Thai. *Mon-Khmer Studies* 27: 303-6.

Suphawut Wathabunditkul. 2003. Spacing in the Thai Language. *http://www.thailanguage.com/ref/spacing* (accessed on March 4, 2010).