

# DELPH-IN Grammar of Thai



Glenn Slayden  
University of Washington  
10<sup>th</sup> DELPH-IN Summit, Tomar,  
Portugal  
July 15 2015

## Technical facts

- Not using “LKB-style morphology”
- Token mapping enabled
  - 1 placeholder rule for now
  - POS tagger: trigram HMM trained on “Orchid” corpus
- Syntax rules: 27
- Native LEs: 223
- Generic LEs: 3 (placeholders)
- ICONS (but no phenomena yet)

## Refactoring a matrix grammar

- 2014: take the plunge and “fork” from matrix
  - /tdl-lex
    - /adjective.tdl, /noun-common.tdl, verb-aux.tdl,...
  - /tdl-main (“entries”)
    - /labels.tdl, /roots.tdl, /rules.tdl, /trigger.tdl,...
  - /tdl-types (“types”)
    - /fundamentals.tdl, /syntax.tdl, lex-rule.tdl,
    - /lex-type.tdl, /tmt.tdl, /icons.tdl ...

## Thai tokenization

- Thai does not use space between words
- Word-breaking (tokenization) can be ambiguous:

...ตากลม...

...ตาก ลม... [tâ:k lom] (exposed to the air)

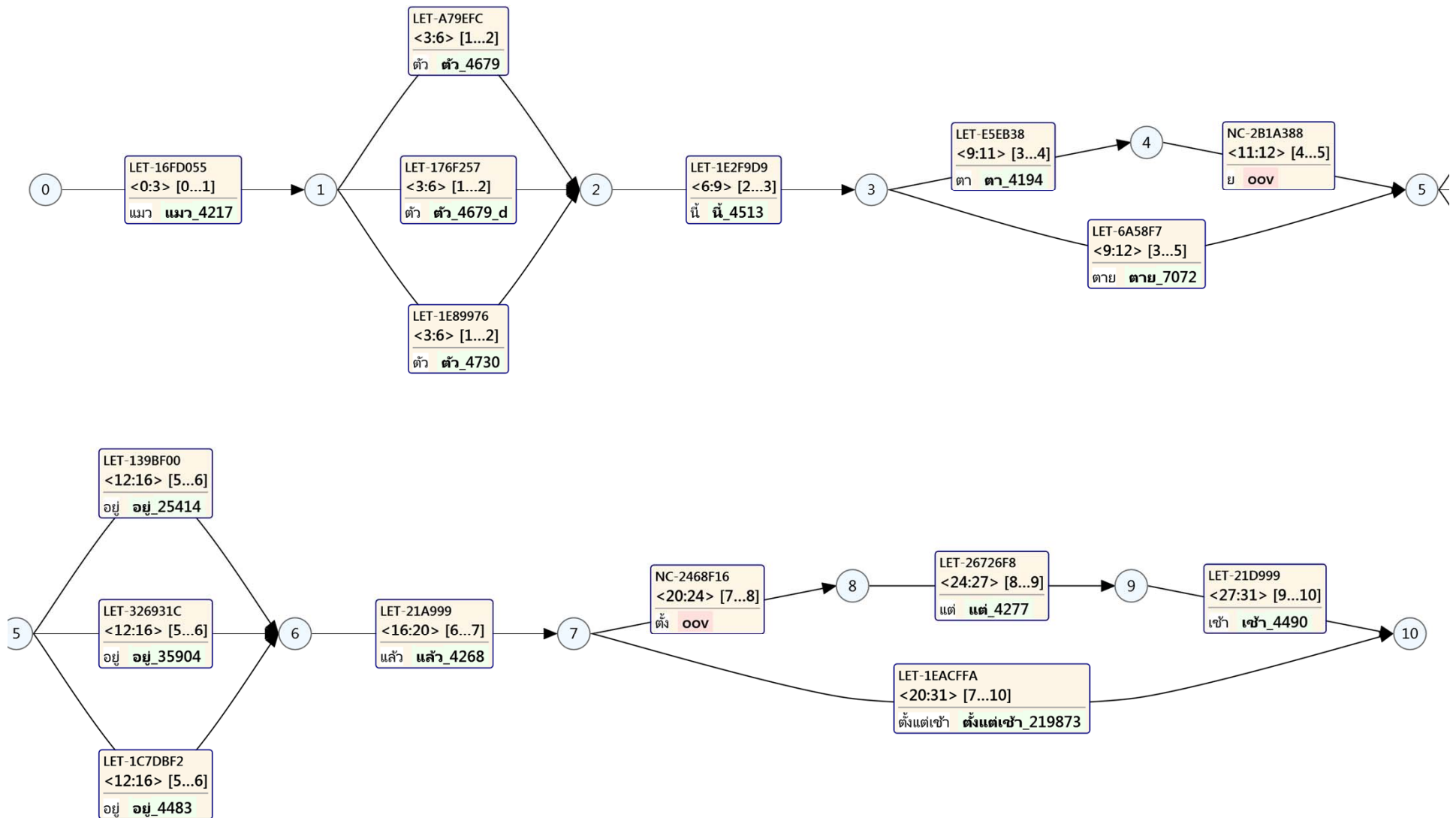
...ตา กลม... [ta: klom] (circular eye)

## Tokenization (continued)

- After several false starts, such as:
  - “conventional” (i.e. statistical) tokenizers
  - “assume cooperative/defer problem” (manual)
  - “whitespace” expedient

## Exhaustive tokenization

- Current: exhaustive based on lexicon
  - when loading grammar, initialize a trie with all lexical items
  - Iterate sentence characters, deploying all possible (overlapping) lexical items into token lattice
  - Skip lookup for certain impossible-initial patterns
  - Postulate generics for uncovered gaps
  - Pass entire lattice as *input* to token mapper
- “ ” (space) added to lattice as an explicit token
  - currently not used



## Lexicon elaboration

- Access to lexicon of *thai-language.com*:
  - 59,218 fully supervised sense entries/categories
  - 15,025 Thai lemmas
  - 22,453 simple English senses (1.49 per lemma)
  - 36,102 compound items with English sense (incl. phrase, sentence)
  - 9,668 complete sentences
  - 587 grouping categories (any items, including categories)



## Lexicon integration: false starts

- Dump entire lexicon into TDL ✘
  - DELPH-IN grammars require precise supervision of part-of-speech (lexical supertype)
- Store TDL text in tldb database ✘
  - Easier to manage TDL all in one place (text file)
  - But we really don't want to drift out of sync

## Lexicon integration: solution

1. Store supertypes only in tldb
2. Website (admin setting) can show TDL for copy/paste
3. Standalone TDL reader maintains sync to/from tldb

4. ลูกแอปเปิ้ล <small>lú:k ʔè:p pʰɯm [noun, loanword, English]</small>	
definition	an apple
item data	D 20688 ลูกแอปเปิ้ล_20688 := common-noun-lex & [ ORTH < "ลูกแอปเปิ้ล" >, SYNSEM.LKEYS.KEYREL.PRED "_apple_n_3_rel" ].
image	
categories	ผลไม้ → Fruits

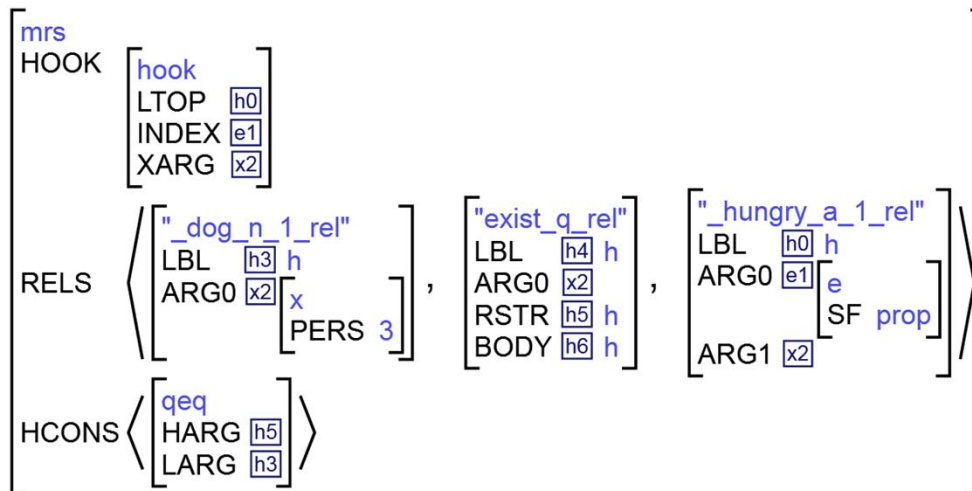
# Thai Grammar: Syntactic Phenomena

- Predicative adjectives

หมาหิว

dog hungry

“The dog’s hungry.”



- Greeting; optional polite particle

สวัสดีครับ

greeting [polite]

“Hello!”

สวัสดีครับ

Hello!

CP cd-comp-head (root\_strict)  
cd-comp-head-phrase

ADJ สวัสดี\_196817  
attr-adjective-lex

C ครับ\_4370  
formality-part-lex

สวัสดี

ครับ  
[polite]

root\_strict

```

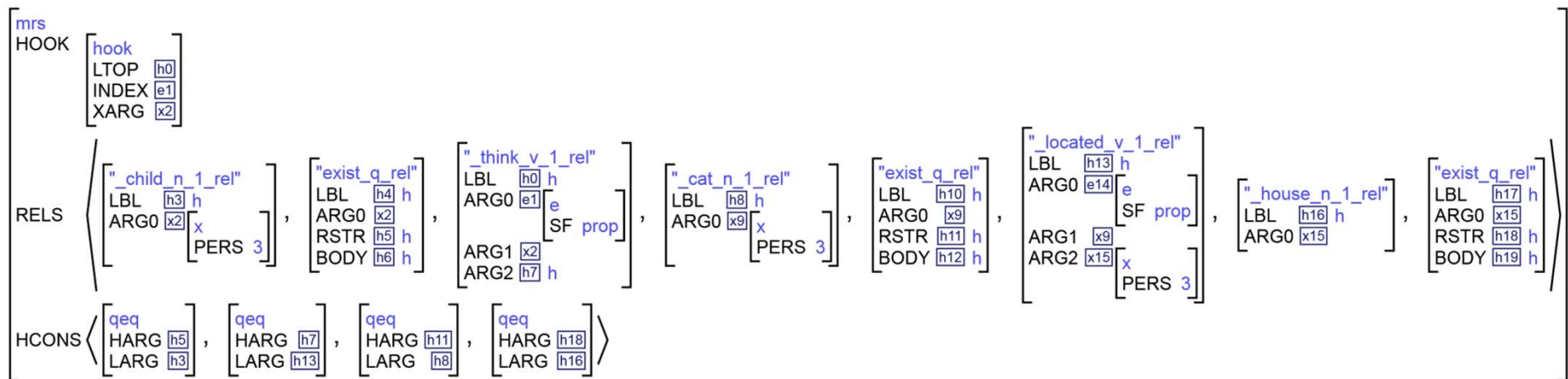
[h0] [e1]
{ [h0] : "_greeting_a_1_rel"([e1], [i2]) }
{ }
    
```

```

[ mrs
  HOOK [ hook
        [ LTOP [h0]
          INDEX [e1]
          XARG [i2] ] ] ]
  RELS [ "_greeting_a_1_rel"
        [ LBL [h0] h
          ARG0 [e1] e
          ARG1 [i2] i ] ] ]
  HCONS bag
    ]
    
```

- Clause embedding

เด็กคิดว่าแมวอยู่บ้าน  
child think [compl] cat [loc] house  
“The child thinks that the cat is in the house.”



- Embedded question

เด็กถามว่าถึงบ้านหรือยัง  
child ask [compl] arrive house or not  
“The kid is asking ‘Are we there yet?’”

```
[h0] [e1] { SF prop }  
{ [h3] : "_child_n_1_rel"([x2] { PERS 3 })  
  [h4] : "exist_q_rel"([x2], [h5], [h6])  
  [h0] : "_ask_v_1_rel"([e1], [x2], [h7])  
  [h8] : "_arrive_v_1_rel"([e9] { SF ques }, [x10] { COG-ST in-foc }, [x11] { PERS 3 })  
  [h12] : "_house_n_1_rel"([x11])  
  [h13] : "exist_q_rel"([x11], [h14], [h15]) }  
{ [h5] qeq [h3], [h7] qeq [h8], [h14] qeq [h12] }
```

## Embedded question

ฉันถามว่าคุณทราบว่ามีรถไหม

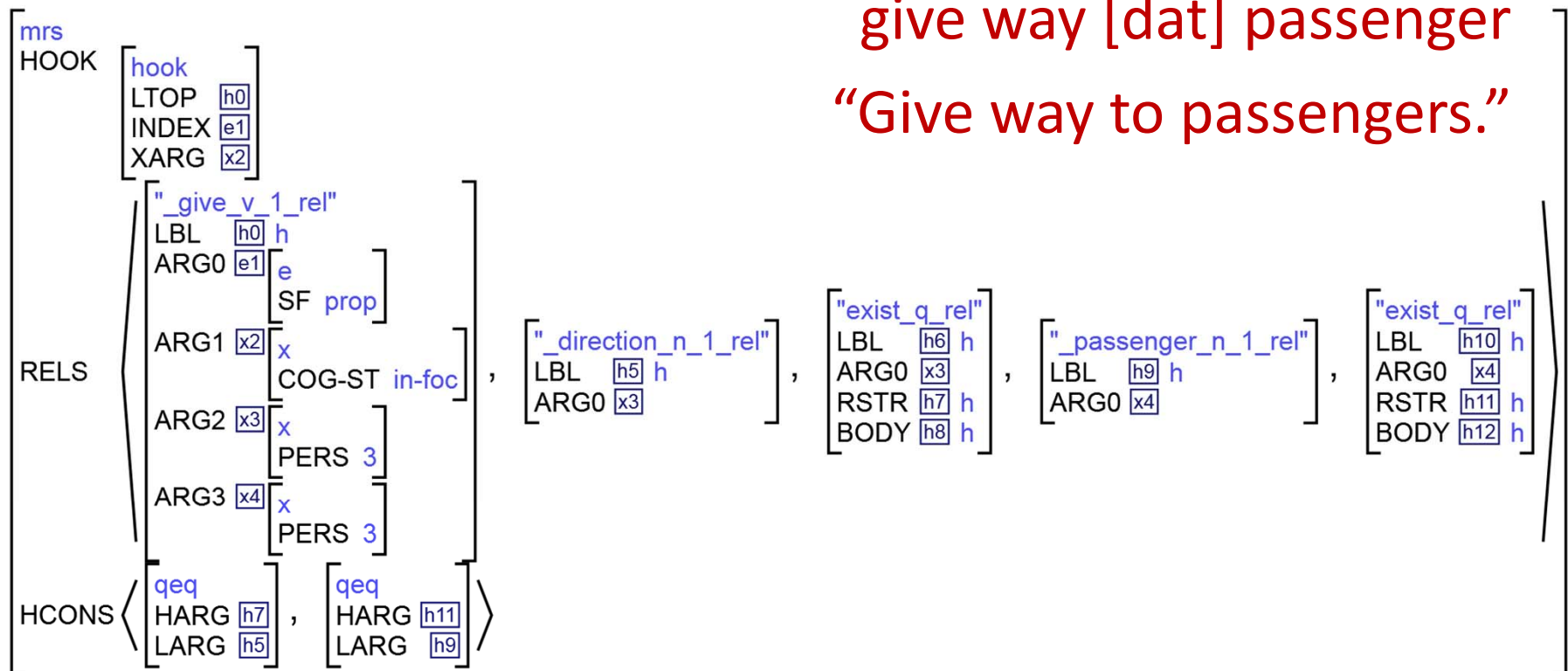
[1sg] ask [compl] [2] know [compl] dog chase car [ques]

“I’m asking if you know whether dogs chase cars.”

```
[h0] [e1] { SF prop }  
{ [h3] : "pron_rel"([x2] { PERS 1, NUM sg, GEND f, SPECI + })  
  [h4] : "exist_q_rel"([x2], [h5], [h6])  
  [h0] : "_ask_v_1_rel"([e1], [x2], [h7])  
  [h8] : "pron_rel"([x9] { PERS 2, NUM sg, SPECI + })  
  [h10] : "exist_q_rel"([x9], [h11], [h12])  
  [h13] : "_know_v_1_rel"([e14] { SF ques }, [x9], [h15])  
  [h16] : "_dog_n_1_rel"([x17] { PERS 3 })  
  [h18] : "exist_q_rel"([x17], [h19], [h20])  
  [h21] : "_chase_v_1_rel"([e22] { SF prop }, [x17], [x23] { PERS 3 })  
  [h24] : "_car_n_1_rel"([x23])  
  [h25] : "exist_q_rel"([x23], [h26], [h27]) }  
{ [h5] qeq [h3], [h7] qeq [h13], [h11] qeq [h8], [h15] qeq [h21], [h19] qeq [h16], [h26] qeq [h24] }
```

- Case marker

ให้ทางแก่ผู้โดยสาร  
give way [dat] passenger  
“Give way to passengers.”



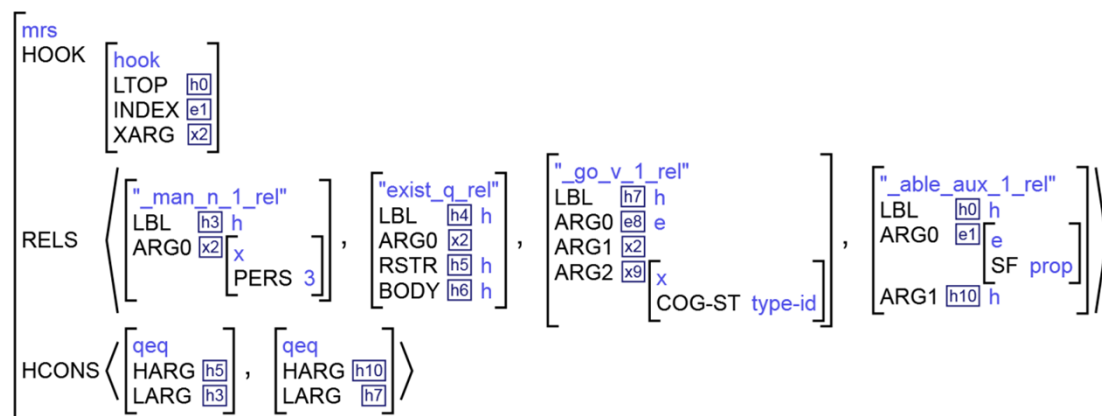


- Modal auxiliaries

ไม่ยอมกินอะไร

[neg] allow eat [anything]

“I didn’t allow (myself) to eat anything.”



ผู้ชายไปไหน

man go [pot]

“He’s able to go.”

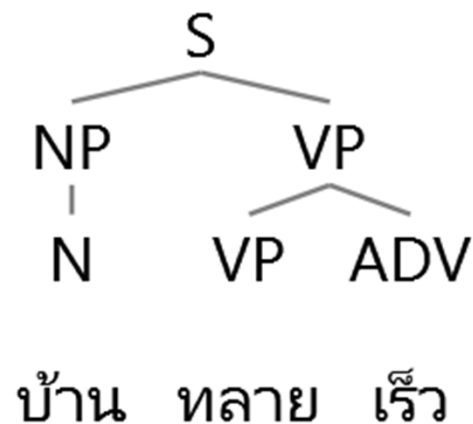
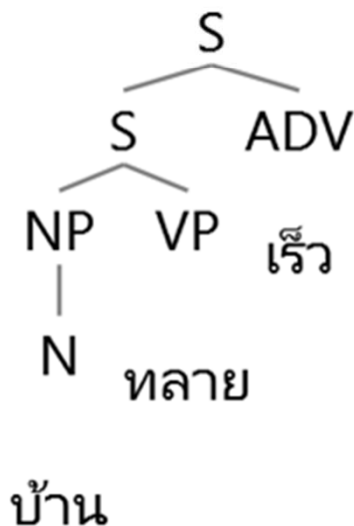
- Adverbs

```
[h0] [e1] { SF prop }  
{ [h3] : "_house_n_1_rel"([x2] { PERS 3 })  
  [h4] : "_exist_q_rel"([x2], [h5], [h6])  
  [h0] : "_collapse_v_1_rel"([e1], [x2])  
  [h0] : "_quick_r_1_rel"([e7], [e1]) }  
{ [h5] qeq [h3] }
```

บ้านทลายเร็ว

house collapse quick

“The house collapsed quickly.”

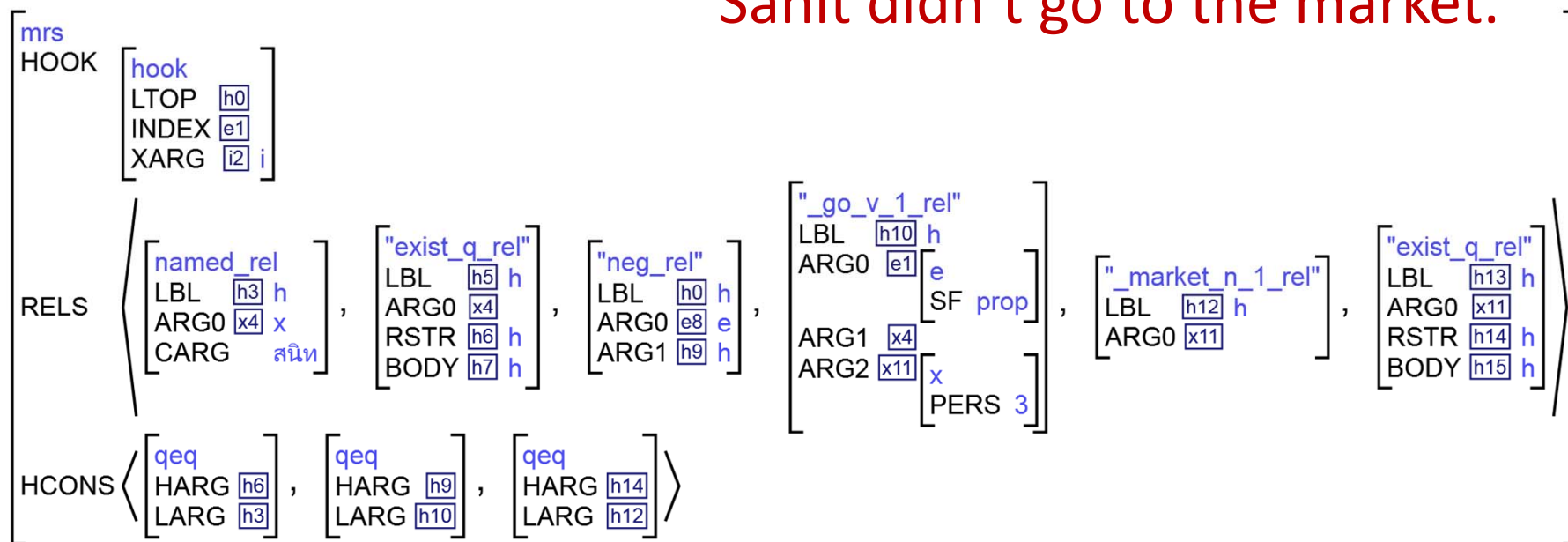


- Negation

สนิทไม่ไปตลาด

[name] [neg] go market

“Sanit didn’t go to the market.”

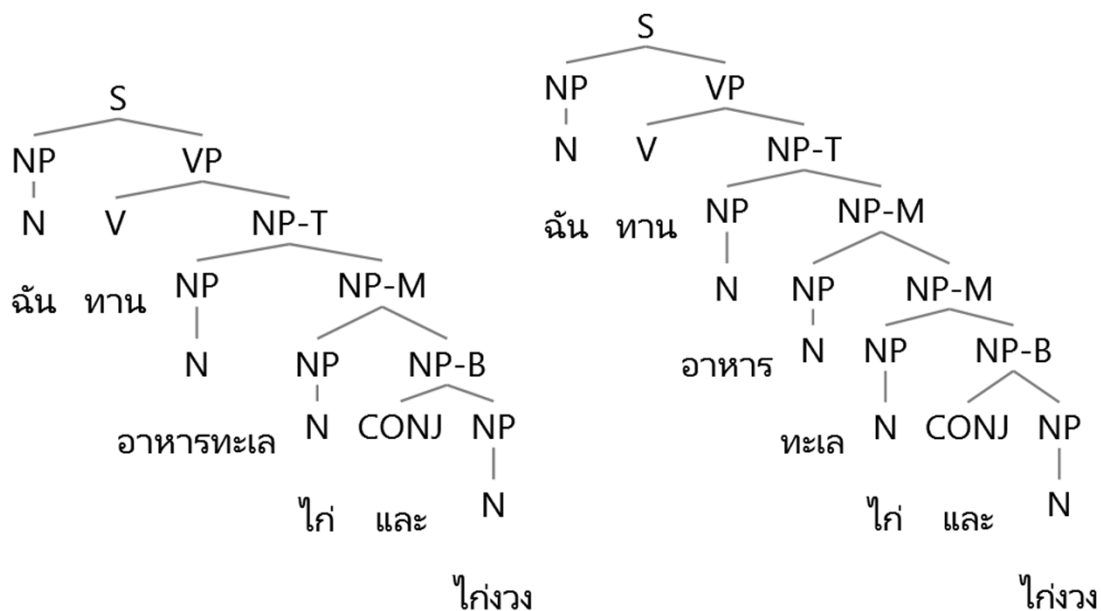


- Nominal coordination

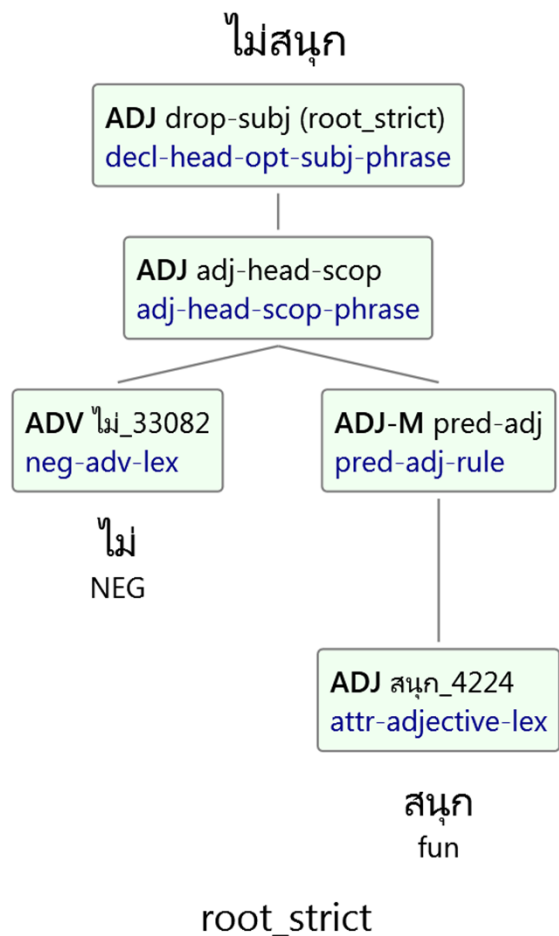
ฉันทานอาหารทะเล ไก่ และ ไก่วง  
[1sg] eat seafood chicken and turkey  
“I eat seafood, chicken and turkey.”

```

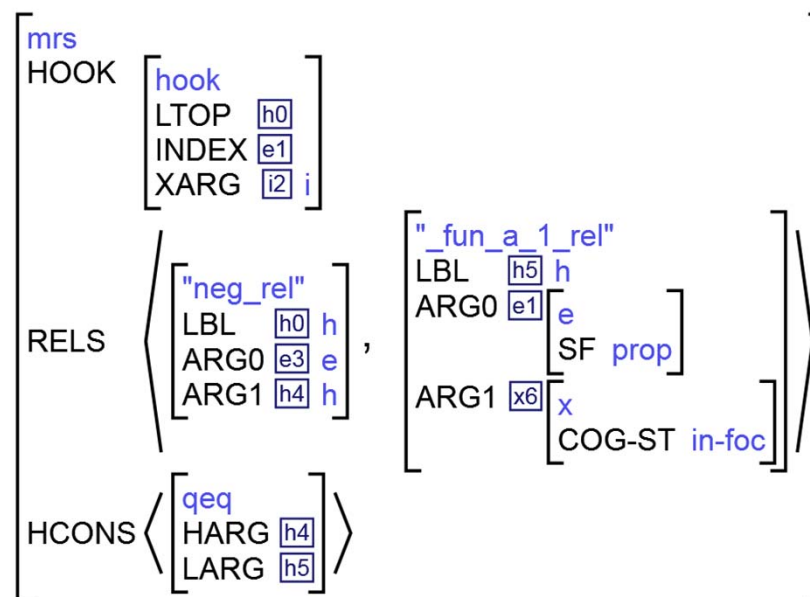
h0 e1 { SF prop }
{ h3 : "pron_rel"(x2 { PERS 1, NUM sg, GEND f, SPECI + })
  h4 : "exist_q_rel"(x2, h5, h6)
  h0 : "_eat_v_3_rel"(e1, x2, x7)
  h8 : "_seafood_n_1_rel"(x9 { PERS 3 })
  h10 : "exist_q_rel"(x9, h11, h12)
  h13 : "_chicken_n_1_rel"(x14 { PERS 3 })
  h15 : "exist_q_rel"(x14, h16, h17)
  h18 : "_and_c_rel"(h19, h20, x21, x14, x22 { PERS 3 })
  h23 : "_turkey_n_1_rel"(x22)
  h24 : "exist_q_rel"(x22, h25, h26)
  h27 : predsor(x21, h18, h28)
  h29 : "implicit_coord_rel"(h30, h31, x7, x9, x21)
  h32 : "exist_q_rel"(x7, h29, h33) }
{ h5 qeq h3, h11 qeq h8, h16 qeq h13, h25 qeq h23 }
    
```



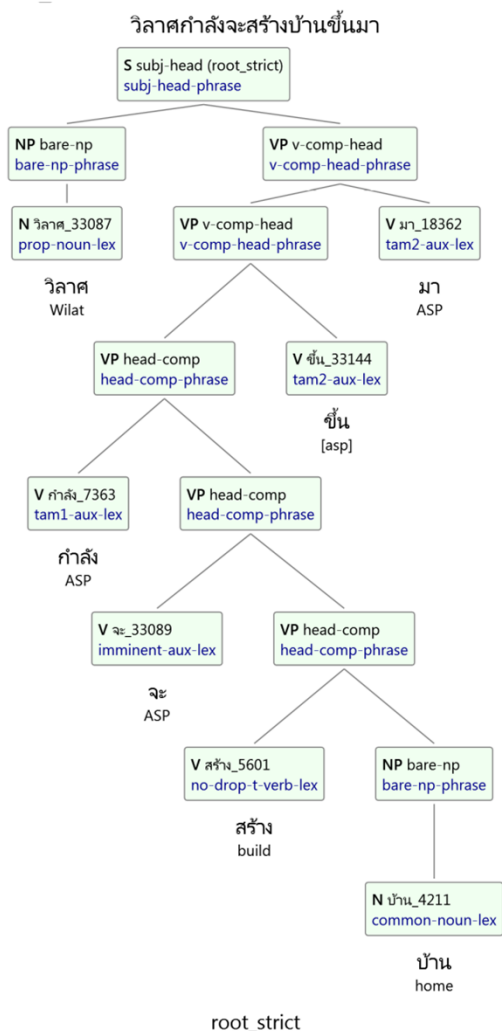
- Subject-drop and object drop



ไม่สนุก  
[neg] fun  
“It’s not fun.”

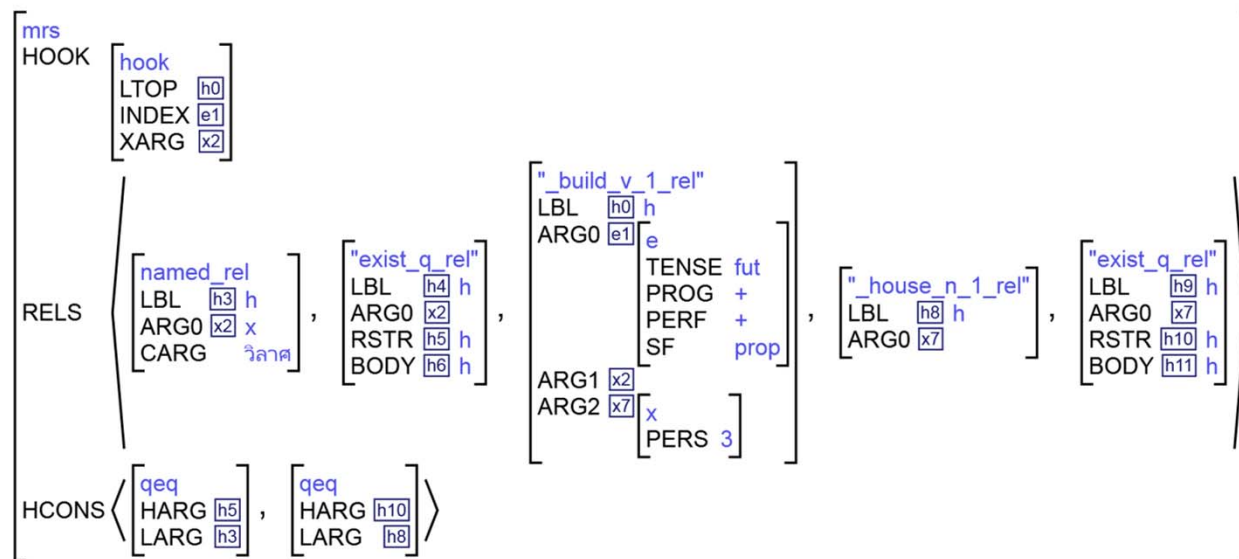


• Positional constraints on aspect markers



วิลาศกำลังจะสร้างบ้านขึ้นมา

[name] [asp] [asp] build house [asp] [asp]  
“Wilat is about to start building a house.”



- Question particles

หรือเปล่า, หรือยัง, ไหม

- Negative copula

ดินสอไม่ใช่ปากกา

pencil [neg] pen

“A pencil is not a pen.”

- Copula, possessive

เขาคือครูของฉัน

[3] [cop] teacher [poss] [1sg]

“He’s my teacher.”

- Verbal coordination (explicit)

เขาไปซื้อดอกไม้ที่ตลาดและไปเยี่ยมเพื่อน

[3] go buy flower at market and go visit friend

“He went to buy flowers and visit a friend.”

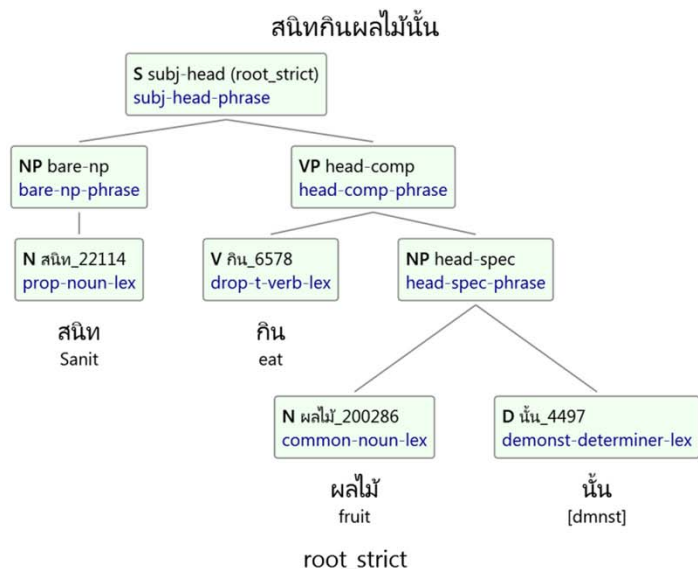
- Verb serialization (implicit coordination)

ต้องรีบกลับไปเรียกให้ มาบอก

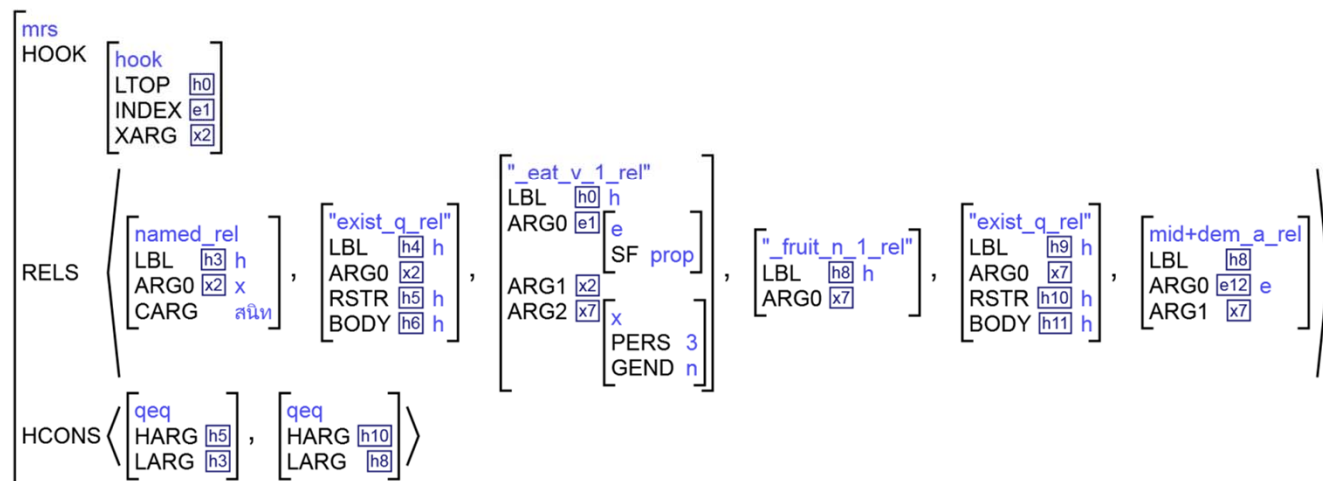
must hurry return go call give come say

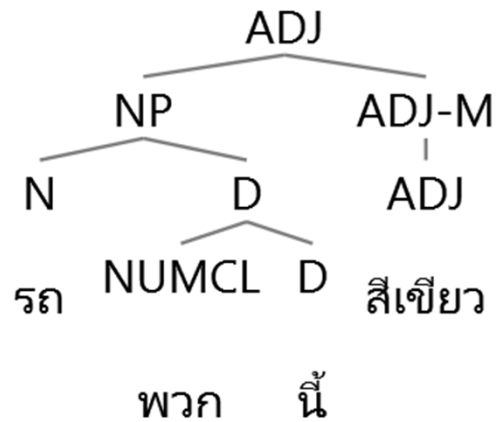
“You must hurry back and call him to come tell me.”



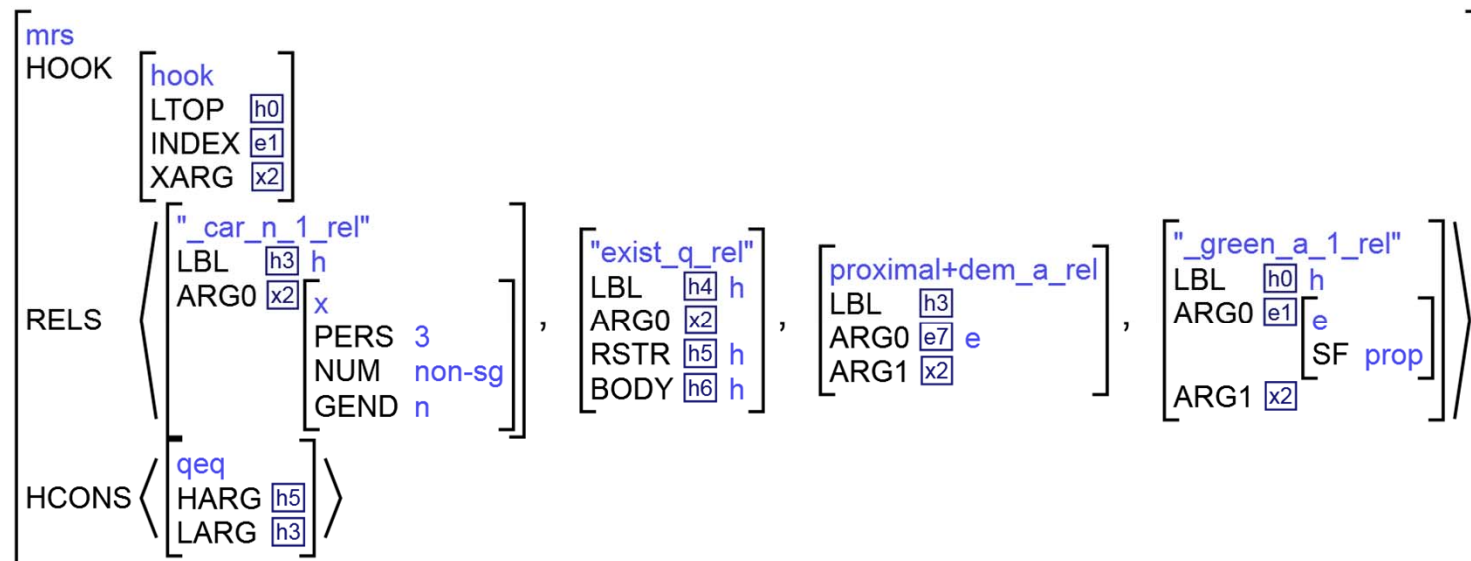


Demonstratives  
สนิทกินผลไม้ นั้น  
[name] eat fruit [dmnst]  
“Sanit ate that fruit.”





รถพวกนี้สีเขียว  
car [numcl] [dmnst] green  
“These cars are green.”

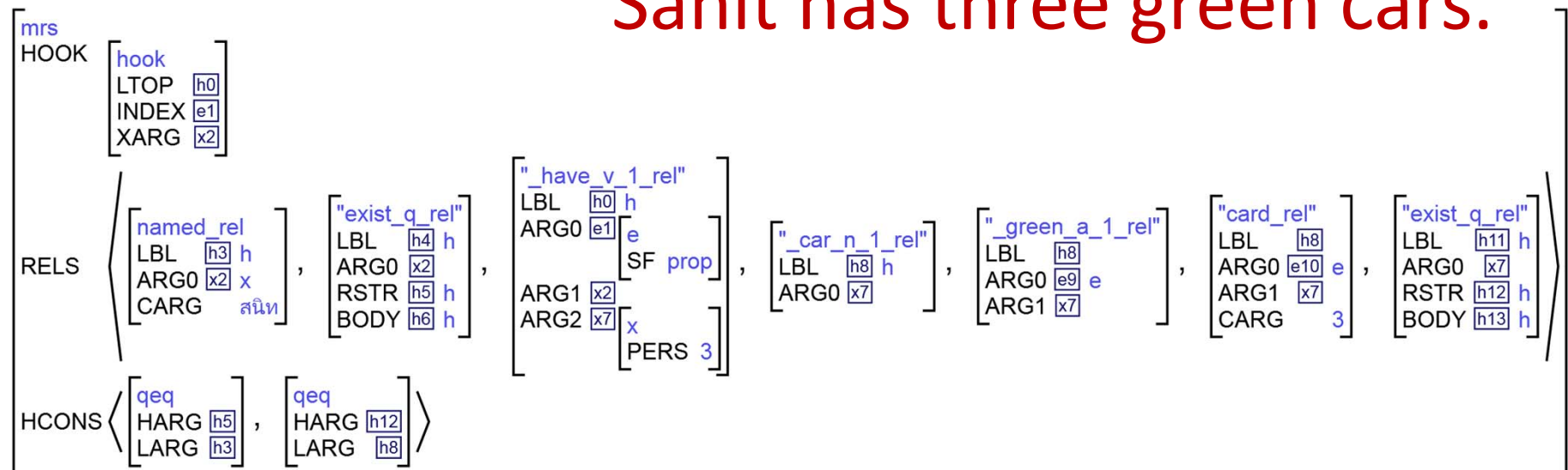


## Numeric classifiers

สนิทมีรถสีเขียวสามคัน

[name] have car green three [clas]

“Sanit has three green cars.”



## Numeric classifier/demonstrative interaction

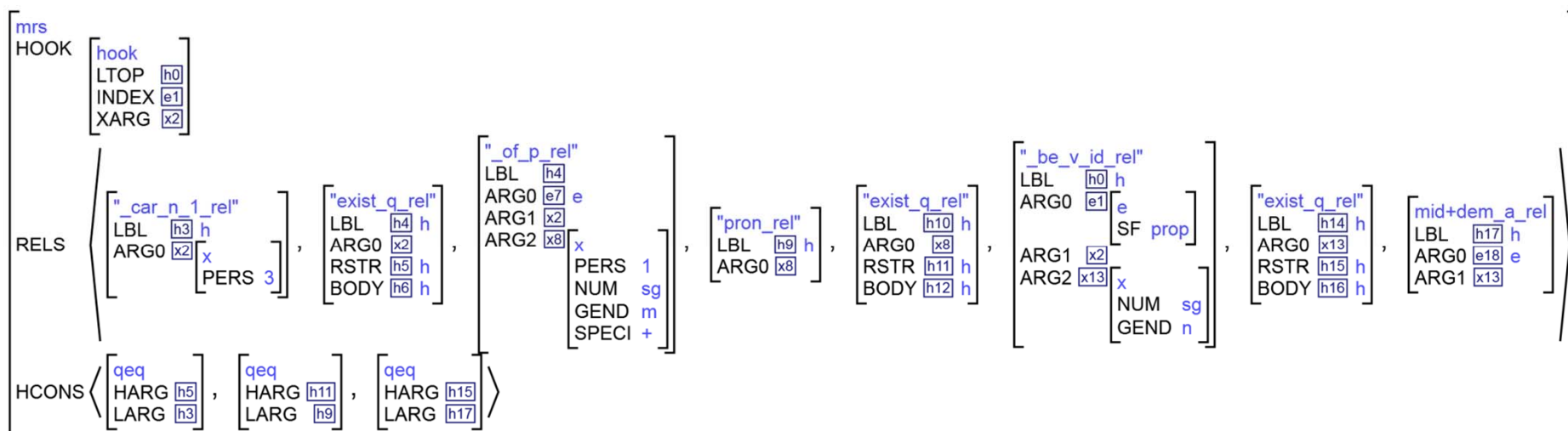
```
h0 e1 { SF prop }  
{ h3 : named_rel(x2, สนิท)  
  h4 : "exist_q_rel"(x2, h5, h6)  
  h0 : "_buy_v_1_rel"(e1, x2, x7 { PERS 3, NUM sg, GEND n })  
  h8 : "_car_n_1_rel"(x7)  
  h8 : "_green_a_1_rel"(e9, x7)  
  h8 : "card_rel"(e10, x7, "3")  
  h11 : "exist_q_rel"(x7, h12, h13)  
  h8 : proximal+dem_a_rel(e14, x7) }  
{ h5 qeq h3, h12 qeq h8 }
```

สนิทซื้อรถสีเขียวสามคันนี้  
[name] buy car green three [clas] [dmnst]  
“Sanit bought these three green cars.”

```
h0 e1 { SF prop }  
{ h3 : named_rel(x2, สนิท)  
  h4 : "exist_q_rel"(x2, h5, h6)  
  h0 : "_buy_v_1_rel"(e1, x2, x7 { PERS 3, NUM sg, GEND n })  
  h8 : "_car_n_1_rel"(x7)  
  h8 : "_green_a_1_rel"(e9, x7)  
  h8 : "card_rel"(e10, x7, "1")  
  h11 : "exist_q_rel"(x7, h12, h13)  
  h8 : proximal+dem_a_rel(e14, x7) }  
{ h5 qeq h3, h12 qeq h8 }
```

สนิทซื้อรถสีเขียวคันหนึ่งนี้  
[name] buy car green [clas] one [dmnst]  
“Sanit bought this one green car.”

รถของผมคือคันนั้น  
car [poss] [1sg] [cop] [clas] [dmnst]  
“My car is that one.”



## Presentational

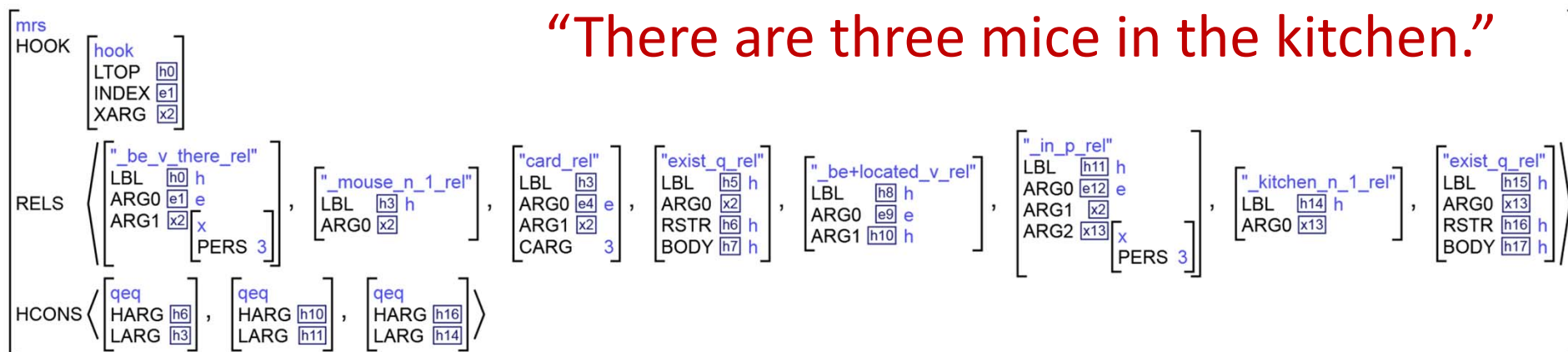
```

h0 e1
{ h0 : "_be_v_there_rel"(e1, x2 { PERS 3 })
  h3 : "_mouse_n_1_rel"(x2)
  h3 : "card_rel"(e4, x2, "3")
  h5 : "exist_q_rel"(x2, h6, h7)
  h8 : "_be+located_v_rel"(e9, h10)
  h11 : "_in_p_rel"(e12, x2, x13 { PERS 3 })
  h14 : "_kitchen_n_1_rel"(x13)
  h15 : "exist_q_rel"(x13, h16, h17) }
{ h6 qeq h3, h10 qeq h11, h16 qeq h14 }
    
```

มีหนูสามตัวอยู่ในครัว

have mouse three [clas] [loc] in kitchen

“There are three mice in the kitchen.”



## Future work

- Unmarked possessive  
บ้านฉันทลาย  
house 1sg collapse
- “About cats”
  - <http://www.thai-language.com/id/590394/>
  - [http://www.thai-language.com/id/590417](http://www.thai-language.com/id/590417/)
- Relative clauses
- ...et cetera ad infinitum