

August 25, 2006

# Texas Linguistics Society 10

CENTER FOR THE STUDY  
OF LANGUAGE  
AND INFORMATION

August 25, 2006

# Two Approaches to Mayan Grammar Development in CCG

## 1.1 Introduction

Computational grammar development often tends towards the development of lexemic grammars, i.e. grammars which posit little or no word-internal grammatical structure, or assume such morphological structure is handled by a system other than syntax proper. However, certain morphological processes do have significant syntactic consequences. Verb-internal incorporated pronouns in Mayan languages and others constitutes one such phenomenon: as incorporated pronouns saturate one grammatical argument and semantic role, they are basically morphological affixes which alter a verb's sub-categorization. This is a particularly acute issue for Categorical Grammar, as syntactic categories—or algebraic characterizations of words' and phrases' sub-categorization—occupy a central role in the theory. As such, incorporated pronouns constitute an appropriate case study in thinking about lexemic vs. morphemic styles of computational grammar development.

This paper addresses this issue by presenting contrasting lexemic and morphemic analyses of incorporated pronouns and ergativity markers in the Mayan language Popti'. Popti' (Craig, 1977) is a configurational VSO language with an ergative/absolutive nominal system articulated by agreement markers and incorporated pronouns on the verb forms. Relative clause formation and focus constructions are both right-branching phenomena, and there are constraints on what constituents may be raised out of either construction. That these constraints are closely related to the modalities presented in §1.1.1 is the subject of

2 /

the next section.

### 1.1.1 CCG

(Multi-modal) Combinatory Categorical Grammar (CCG) (Baldrige, 2002, Baldrige and Kruijff, 2003) is a mathematically constrained radically lexicalist grammatical formalism. In CCG, lexical items are assigned one or more categorial types which are formed from basic categories  $s, np, \dots$  closed under the the directional slash operators  $\{/, \backslash, \dots\}$ . Slashes, in turn, are decorated by modalities  $\{*, \diamond, \times, \cdot\}$  which dictate the applicability of the rules in the system to the category formed by that slash.

The rule schemata over categories are given in Figure 1. Syntactic categories encode sub-categorization and word-order in a concise algebraic manner, and words and phrases are given a unified analysis, as each is simply lent a categorial type, or one derived from the item's constituents.

With respect to applicability to the rules, the modalities are related to each other via the hierarchy in Figure 1. Slash-categories decorated with  $\times$  or  $\diamond$  can enter into the rules for  $*$ , namely  $<, >$ , and slash-categories decorated with  $\cdot$  can in fact enter into any rule above. Moreover, a slash decorated with  $\cdot$  can unify with a slash decorated with any of the other modalities.

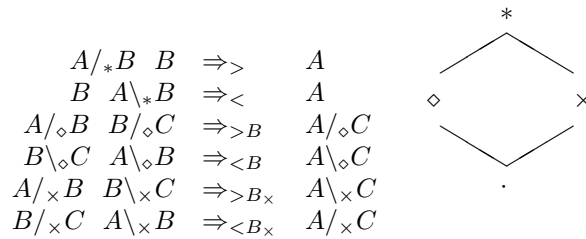


FIGURE 1 CCG rule schemata and modality hierarchy

As a methodological priority, the fewest number of categories possible are assigned to a given lexical item.

## 1.2 Lexemic Analysis

### 1.2.1 Basic Word Order

Popti' features a rich set of determiner-like NP classifiers. Classifiers serve both as pronouns (of type  $np$ ) and as determiners. A possible analysis of the determiner role could be given by a category like  $np/n$ ,

## TWO APPROACHES TO MAYAN GRAMMAR DEVELOPMENT IN CCG / 3

however as Popti' nouns are predicative (1.1), we opt instead for  $s/np$  for nouns and  $np/(s/np)$  for classifiers. (Here and below, when slashes are not marked with a modality, it is to suggest that the particular modality has not been established at this point in the discussion, and the  $\diamond$  is assumed. Likewise, if no rule is given in a CCG step, it should generally be  $<$  or  $>$ .)

$$(1.1) \begin{array}{l} \textit{winaj} \quad \textit{hach} \\ \text{man} \quad \text{you} \\ \text{You are a man} \end{array} \qquad \frac{\frac{\textit{winaj}}{s/np} \quad \frac{\textit{hach}}{np}}{s}$$

$$(1.2) \begin{array}{l} \textit{x'apni} \quad \textit{naj} \quad \textit{winaj} \\ \text{arrived} \quad \text{cl} \quad \text{man} \\ \\ \text{The man arrived} \end{array} \qquad \frac{\frac{\textit{x'apni}}{s/np} \quad \frac{\frac{\textit{naj}}{np/(s/np)} \quad \frac{\textit{winaj}}{s/np}}{np}}{s}$$

Further evidence for this analysis is proved by relative clauses, in §1.2.2.

**1.2.2 Relative clauses**

Relative clauses are formed by finite clauses with the absolutive argument extracted to the left, as in

$$(1.3) \begin{array}{l} \textit{naj} \quad \textit{winaj} \quad \textit{x'apni} \quad \textit{ewi} \\ \text{cl} \quad \text{man} \quad \text{arrived} \quad \text{yesterday} \\ \text{the man who arrived yesterday} \end{array}$$

That is to say, an  $s/np$  verb phrase forms a relative clause. By one analysis, there is a relative clause formation type-changing rule that transforms predicates into modifiers, qua  $s/np \rightsquigarrow np \backslash np$ . However, closer inspection reveals that such a rule is not necessary for this construction, if we assume the  $np/(s/np)$  analysis for classifiers above:

$$(1.4) \begin{array}{l} \textit{ka'} \quad \textit{c'ulch'en} \quad \textit{ch'en} \quad \textit{xaloko} \\ \text{more} \quad \text{pretty} \quad \text{cl/them} \quad \text{you-bought} \\ \text{The ones you bought are more pretty} \end{array}$$

$$\frac{\frac{\frac{\textit{ka'}}{s/s} \quad \frac{\textit{c'ulch'en}}{s/np}}{s/np} \quad >B \quad \frac{\frac{\textit{ch'en}}{np/(s/np)} \quad \frac{\textit{xaloko}}{\ddot{s}/np}}{np}}{s}$$

Coming back to (1.3), we need a story for how *ewi*, “yesterday” combines with the relative clause construction. Consider:

4 /

- (1.5) *x'apni naj winaj ewi*  
 arrived cl/the man yesterday  
 The man arrived yesterday.

Based on (1.2.2) assume *ewi* has categorial type akin to  $s \backslash s$ , then to get the relative clause formation in (1.3) to work, we ascribe the  $\times$  modality to the relevant slashes verbal and adverbial categories:

$$\frac{\frac{naj}{np/(s/np)} \quad \frac{\frac{winaj}{(s/np)/(s/\times np)} \quad \frac{\frac{x'apni}{s/\times np} \quad \frac{ewi}{s \backslash \times s}}{s/\times np}}{s/np}}{np} < B_{\times}$$

This analysis predicts that the adverb *ewi* can shift to the left of the subject in finite clauses, as in

- (1.6) *x'apni ewi naj winaj*  
 arrived yesterday cl/the man  
 The man arrived yesterday

$$\frac{\frac{\frac{x'apni}{s/\times np} \quad \frac{ewi}{s \backslash \times s}}{s/\times np} < B_{\times} \quad \frac{naj \ winaj}{np}}{s}$$

Indeed, this prediction seems to be correct. (Nora England, p.c.).

### 1.2.3 The focus operator *ha'*

The focus operator *ha'* extracts either the object of transitive verbs or the subject of intransitive verbs (i.e. the absolutive constituents) from the VSO verbal nucleus. The categorial type for *ha'* is given by  $(s/(s/\times np))/np$ , as in

- (1.7) *ha' naj smak ix*  
 focus him hit she  
 It's him who she hit

$$\frac{\frac{\frac{ha'}{(s/(s/\times np_1))/np_1}}{s/(s/\times np_1)} \quad \frac{naj}{np_1} \quad \frac{\frac{smak}{(s/\times np_1)/np_2} \quad \frac{ix}{np_2}}{s/\times np_1}}{s}$$

It follows immediately from the categorial analysis of fronting, here, that only the subject of intransitives or the object of transitives may be focused, as indicated in the derivation for (1.7), where  $np_1$  must be the object and  $np_2$  the subject of the clause. Also, the particular type provided for *ha'* prevents multiple instances of focus-extraction.

## TWO APPROACHES TO MAYAN GRAMMAR DEVELOPMENT IN CCG / 5

Popti' has a means to focus the subject of transitives, however. There is a morphologically realized lexical rule that transforms *smak* to *xmakni* and that, in this analysis, basically switches the arguments of the verb. So:

- (1.8) *ha' naj xmakni ix*  
 focus cl/he hit cl/her  
 It's he who hit her

$$\frac{\frac{\frac{ha'}{(s/(s/\times np_1))}/np_1} \quad \frac{naj}{np_1} \quad \frac{xmakni}{(s/\times np_1)/np_2} \quad \frac{ix}{np_2}}{s/(s/\times np_1)} \quad \frac{s/\times np_1}{s}$$

**1.2.4 Complements and quotations**

Verbs of reporting, such as *xal* ("said") take a complement clause object, as in

- (1.9) *xal naj jet-an tato x'apni ya' cumi*  
 said he to-us that arrived cl/the lady  
 He said to us that the lady arrived

This is handled straightforwardly with a new atomic category *cp*:

$$\frac{\frac{\frac{xal}{(s/\times cp)/\diamond np} \quad \frac{naj}{np}}{s/\times cp} > \quad \frac{jet-an}{s \setminus \times s} < B_{\times} \quad \frac{tato}{cp/\diamond s} \quad \frac{x'apni \quad ya' \quad cumi}{s}}{s/\times cp} < B_{\times} \quad \frac{cp}{s} >$$

Under certain conditions, reporting verbs such as *xal* undergo a morphologically realized transformation into a quoting term, *yalni*, that accompanies quotative inversion:

- (1.10) *x'apni ya' cumi yalni naj jet-an*  
 arrived cl/the lady said he to-us  
 He said the lady arrived

$$\frac{\frac{x'apni \quad ya' \quad cumi}{s} \quad \frac{\frac{yalni}{(s \setminus \times s)/\diamond np} \quad \frac{naj}{np}}{s \setminus \times s} > \quad \frac{jet-an}{s \setminus \times s} <}{s} <$$

**1.2.5 mac and long-distance extraction**

The Wh pronoun *mac* ("who") occurs to the left of the verb and induces a question, roughly like English:

6 /

- (1.11) *mac xul ewi*  
 who arrived yesterday  
 Who arrived yesterday?

$$\frac{\frac{mac}{s/\diamond(s/.np)} \quad \frac{xul}{s/\times np}}{s} > \frac{ewi}{s\backslash\times s}$$

*mac* can enter into (somewhat) long-distance dependencies as in:

- (1.12) *mac xawa' ha melyu tet*  
 who you-gave your money to  
 Who did you give your money to

$$\frac{\frac{mac}{s/\diamond(s/.np)} \quad \frac{\frac{xawa'}{(s/\diamond pp)/\times np} \quad \frac{\frac{ha}{np/\diamond(s/.np)} \quad \frac{melyu}{s/\diamond np}}{np}}{s/\diamond pp}}{s} > \frac{tet}{pp/\diamond np} > B$$

However, combining *mac* with the reporting verb *xal* from §1.2.4 does not go through straightforwardly. Crucially, harmonic composition is blocked by the  $\times$  modality on the verbal absolutive nucleus  $s/\times np$ . Consider this attempt for “Who did Peter say hit Mary?”:

- (1.13) \**mac xal naj pel chubil xmakni ix malin*  
 who said cl Peter that hit cl Mary  
 (attempted:) Who did Pater say hit Mary?

$$\frac{\frac{chubil}{cp/\diamond s} \quad \frac{\frac{xmakni}{(s/\times np)/\diamond np} \quad \frac{ix malin}{np}}{s/\times np}}{s} !!$$

In Craig (1977)’s treatment, (1.13) is marked with a (?). This may be because, since *xmakni* is a subject-inverted pseudo-passive, some speakers may treat it without the  $\times$  modal verbal nucleus. Nevertheless, using the quotative construction produces a clean reading:

- (1.14) *mac xmakni ix malin yalni naj pel*  
 who hit cl Mary say cl Peter  
 Who did Peter say hit Mary?

## TWO APPROACHES TO MAYAN GRAMMAR DEVELOPMENT IN CCG / 7

$$\frac{\frac{\frac{mac}{s/\diamond(s/.np)}}{s} \quad \frac{\frac{\frac{xmakni}{(s/\times np)/\diamond np} \quad \frac{ix malin}{np}}{s/\times np} > \quad \frac{\frac{\frac{yalni}{(s\backslash\times s)/\diamond np} \quad \frac{naj pel}{np}}{s\backslash\times s} < B_{\times}}{s/\times np} >$$

In fact, the grammar implementation gets both readings, as in

- “Who did Peter say hit Mary?”
- “Peter said ‘Who hit Mary?’”

### 1.3 Issues with the Lexemic Analysis

A lexemic approach such as is posited in §1.2 poses several methodological issues. In practice, the grammar writer must generally specify a larger number of lexical items, and with reduplicated effort comes the increased likelihood of error.

Incorporated pronouns in languages such as Popti’ provide an acute test case of this. Popti’ has two sets of incorporated pronouns, for each the absolutive and ergative verbal arguments. To accommodate this, the lexemic grammar written for the fragment of Popti’ alluded to above specified the following verbal categories:

- (1.15) • StativePredicate  $s/\diamond np$   
 • Subject(-embedded)TransitiveVerb  $s/\times np$   
 • TransitiveVerb  $(s/\times np)/\diamond np$   
 • Subject(-embedded)DitransitiveVerb  $(s/\diamond np)/\times np$

Moreover, aspectual markers and the subject-focus markers were embedded in the lexical items, leading to lexical entries akin to

```
hit:TV { smak; }
hit-sf:TV { xmakni; }
```

In fact, morphologically *smak* is a fairly complex form:

- (1.16)  $x-$        $s-$     *mak*  
 compl e3 hit

where *compl* is completive case and *e3* is the 3rd person ergative marker. A more general morphemic system that can also handle word-internal structure is proposed in the next section.

### 1.4 Morphemic Approach

A morphemic approach (Bozsahin, 2002) differs significantly from a lexemic approach in that it posits categorial types for word-internal

morphemes and combinatory processes for word-formation akin to the syntactic processes themselves. Here, the analysis does not in fact distinguish between morphological and syntactic categories and processes.

A number of approaches were taken to extend the categorial analysis to the word internal morphemes. The basic evaluation of the analyses was the same testing data used to evaluate the lexemic analysis originally. The analysis that had the same or better results as the lexemic grammar for the test sentences is as follows.

Separate entries are made for the ergative marker, ergative incorporated pronouns, aspect markers and subject focus operator. This analysis, moreover, makes greater use of features than the lexemic analysis, the semantics<sup>1</sup> for the relevant entries is specified, super-scripts are used to indicate feature-structure unification and inheritance (indicated with  $\sim$ ), and finally inert slashes  $/^!$  are used to control derivation:

**Aspect**  $(s_{E,\text{fin}}^1 / \times \text{np}^2) / \diamond (s_{\sim\text{fin}}^1 / \times \text{np}^2)$

Examples:

- *x-* @E : ⟨asp⟩ = compl
- *ch-* @E : ⟨asp⟩ = incompl

**Ergative (incorporated) pronouns**  $(s_{E,\sim\text{fin}}^1 / \times \text{np}_{\text{erg}}^2) / * (s^1 / \text{np}^2)$

Ergative pronouns add a semantic argument to the semantics, but otherwise leave the verbal category as is. Examples:

- *hin-* @E : ⟨Act⟩(pro  $\wedge$  ⟨pers⟩1st)
- *ha-* @E : ⟨Act⟩(pro  $\wedge$  ⟨pers⟩2nd)

**Ergative agreement**  $((s_E^1 / \times \text{np}_{\text{erg}}^2) / \text{np}_X) / * (s^1 / \times \text{np}^2)$

Ergative agreement licenses a new grammatical and semantic argument, which is marked ergative and acts as the Actor:

- *s-* @E : ⟨Act⟩(X  $\wedge$  ⟨agr⟩erg)

**Subject focus**  $((s^1 / \times \text{np}_X) / \diamond \text{np}^2) \setminus * (s^1 / \times \text{np}^2)$

Subject focus also licenses a new argument, which allows the subject ( $\text{np}^2$ ) to be extracted out in the manner sketched in §1.2:

- *-ni* @E(⟨Act⟩X)

#### 1.4.1 Results

The morphemic analysis presented here allows for a simplification of how verbal forms are specified in general. Few lexical entries are required for verb forms, and fewer lexical families are required. In fact, the lexical family specifications were cleaned up considerably from what they were in (1.15):

<sup>1</sup>C.f. Baldridge and Kruijff (2002) for the HLDS formalism.

## TWO APPROACHES TO MAYAN GRAMMAR DEVELOPMENT IN CCG / 9

	StativePredicate	$s_{S, \text{fin}} / \diamond \text{np}_X$	$@S : \langle \text{Th} \rangle X$
(1.17)	IntransitiveVerb	$s_{E, -\text{fin}} / \times \text{np}_{X, -\text{erg}}$	$@E : \langle \text{Act} \rangle X$
	TransitiveVerb	$s_{E, -\text{fin}} / \times \text{np}_{X, -\text{erg}}$	$@E : \langle \text{Pat} \rangle X$

This has some interesting consequences for the grammar specification. First of all, it puts the significant distinction between intransitive and transitive verbs into the semantics. Conceivably, coherence constraints could be specified at the semantic level, reminiscent of similar constraints in Lexical-Functional Grammar (Bresnan, 2001). Secondly, and related, it puts greater emphasis on the use of features to enforce constraints on over-generation. Both issues are taken up in §1.5.

The morphemic grammar parses, produces semantics and is able to realize all the interpretations of the sentences the lexemic grammar was. One construction, involving gapping (‘John ate a mango and Mary an orange’ in Popti), eludes both grammars. And in fact, whereas the lexemic grammar was able to parse the focus constructions involving *ha*’ (§1.2.3), but not realize the original form from the semantics, the smaller and simpler morphemic grammar improved on the analysis of *ha*’ and produce the right realizations.

## 1.5 Issues with the Morphemic Approach

Three significant issues arise from this technique:

### 1.5.1 Over-generation

With ergative pronouns as is the grammar allows multiple ergative pronouns to apply to a single constituent. For example, expressions such as

(1.18) *hin- ha- s- mak ...* (‘I you she hit ...’)

are accepted by the morphemic grammar system. Parsing naturally occurring text, of course, clauses such as these do not occur, however though the semantics for (1.18) appears absurd, it is sufficiently well-formed to realize the original form.

One solution is to provide the verbal forms with features specifying whether they have a saturated subject or not, akin to the  $\pm\text{fin}$  finiteness feature, that can be altered by application of ergative morphemes to as to block repeated application. However, initial experiments with this approach results in trouble realizing forms involving ergative morphemes.

### 1.5.2 Efficiency

While the realization coverage goes up with the morphemic system, in most other criteria it appears to take the morphemic system much

more work to do the parsing and realization task. For certain criteria this seems commensurate with the increased number of lexical items in the corresponding clauses, however for certain other criteria, the difference is as much as from  $6.27\times$  (Avg number of rule applications) to  $9.26\times$  (Avg time until stopped/done with realization). Again, more work is necessary to clarify these results.

## 1.6 Conclusions

The morphemic grammar for Popti' has several nice properties with respect to the lexemic alternative, especially with respect to the depth, generality and compositional consistency of its semantic analyses. However, it may not be practical direction to take grammar development, especially with respect to resulting over-generation and decreased efficiency for certain tasks. Perhaps an integrated strategy, the makes reference to a morphemic system at “compile-time” to capture the desired generalities where relevant, but does actual processing and realization with a purely lexemic system, could take advantage of both techniques.

## References

- Baldrige, Jason. 2002. *Lexically Specified Derivational Control in Combinatory Categorical Grammar*. Ph.D. thesis, University of Edinburgh.
- Baldrige, Jason and Geert-Jan Kruijff. 2002. Coupling ccg with hybrid logic dependency semantics. In *Proceedings of ACL 2002*.
- Baldrige, Jason and Geert-Jan M. Kruijff. 2003. Multi-modal categorial grammar. In *Proceedings of EACL '03*. Budapest.
- Bozsahin, Cem. 2002. The combinatory morphemic lexicon. *Computational Linguistics* 28:145–186.
- Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.
- Craig, Colette Grinevald. 1977. *The Structure of Jacaltec*. Austin: UT Press.