

Two Approaches to Mayan Grammar Development in CCG

ELIAS PONVERT

8.1 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 constitute 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 Categorial 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 morphology in computational grammar development.

This paper makes two contributions. First, it presents a CCG for a fragment of the Mayan language Popti’ that has been engineered using OpenCCG (Bozsahin et al., 2006). 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 left-branching phenomena, and there are constraints on what constituents may be raised out of either construction. These constraints are closely related to the CCG slash-modalities introduced by Baldridge (2002) and presented in §8.1.1.

Texas Linguistics Society 10.
Nicholas Gaylord, Stephen Hilderbrand, Heeyoung Lyu, Alexis Palmer and Elias Ponvert eds.
Copyright © 2008, CSLI Publications.
Secondly, it presents an alternate, morphemic presentation of the grammar fragment that analyzes incorporated pronouns and ergative markers directly as syntactic constituents. This simplifies the categorial presentation of verbs in Popti’ considerable, and simplifies and generalizes the semantics of the entities under consideration.

8.1.1 CCG

(Multi-modal) Combinatory Categorial Grammar (CCG) (Baldridge, 2002, Baldridge and Kruijf, 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, ... closed under the the directional slash operators {/,\,...}. Slashes, in turn, are decorated by modalities {∗,⋄,×,...} which dictate the applicability of the rules in the system to the category formed by that slash.

Syntactic analyses are given in the form of derivation trees, such as:

\[
\begin{array}{c}
\text{kisses} \\
\text{Mary} \\
\end{array}
\begin{array}{c}
\text{John} \\
\text{np} \\
\text{s\np} \\
\text{s\np} \\
\text{<} \\
\end{array}
\]

First of all John and Mary have the category np, i.e. they are noun phrases. kisses has the category (s\np)/np, which means that it is basically looking for a constituent with category np on its right to form a constituent of category s\np, the equivalent of an intransitive verb or verb phrase. (There is no distinction between lexical and phrasal categories in Categorial Grammars.) This occurs with an instance of an Application rule, denoted > in the derivation. The category s\np, likewise, is looking for an np on the left to form a full sentence, i.e. a constituent with category s. This occurs with an instance of the corresponding Application rule <. The application rule schemata, then, are:

\[
\begin{align*}
A\backslash B B & \Rightarrow > A \\
B & \Rightarrow < A
\end{align*}
\]

These rules by themselves would account for very little significant linguistic phenomena. The other rule schemata for CCG (at least, relevant to the current analysis) are Composition rules (B) and the Crossed Composition rules (B ×):

\[
\begin{align*}
A\backslash B B\backslash C & \Rightarrow >_B A\backslash C \\
B\backslash C A\backslash B & \Rightarrow <_B A\backslash C \\
A\backslash B B\backslash C & \Rightarrow >_{B×} A\backslash C \\
B\backslash C A\backslash B & \Rightarrow <_{B×} A\backslash C
\end{align*}
\]

See Steedman (2000) for extensive discussion.

Note that the slashes in the rule schemata above are decorated with the modalities *, ⋄ and ×. These control which rules may apply to what classes.
of categories. The relationship between the modalities is hierarchical, as illustrated in Figure 1. Basically, categories whose slash is decorated with ∗ may only enter instances of Application, those decorated with ◦ may enter into instances of Harmonic Composition or Application, those decorated with × may enter into instances of Crossed Composition or Application, and those decorated with · may enter into instances of any of the rules above. As a methodological priority, the fewest number of categories possible are assigned to a given lexical item. Also, for simplicity, the English example above (8.1) was not presented with slash-modalities. In general, the default modality is ◦, i.e. the default assumption is that the Application and Harmonic Composition rules are applicable to most categories.

8.1.2 Grammar Engineering with CCG
The Popti’ grammar fragment discussed in this article was implemented in OpenCCG. The grammar was developed using the CCG grammar specification language developed by Ben Wing1. The fragment is relatively small: 21 CCG categories were specified for 41 lexical items, plus 6 ergative morphemes, described below. The test bed comprises of 23 clauses, many of which are discussed below. The data and examples are taken from a linguistic grammar of Popti’ (Craig, 1977).

8.2 Lexemic Analysis
This section provides a lexemic presentation of a grammar of a fragment of Popti’. In particular, it posits a unified syntactic account of relative clause formation and agent-focus constructions based on derivational control via slash-modalities.

---

1Cf. http://comp.ling.utexas.edu/ccggui/
8.2.1 Basic Word Order

Popti’ features a rich set of determiner-like NP classifiers. Some examples are:

(8.2) naj winaj
    cl  man
    the man

(8.3) ix malin
    cl  Mary
    Mary

(8.4) naj pel
    cl  Peter
    Peter

The generic term cl is used to gloss classifiers in what follows. Classifiers serve both as pronouns (of type np) and as determiners. A possible analysis of the determiner role could be given by a determiner category like np/n, that is a category which provides a noun phrase given a common noun (with category n). However as Popti’ nouns are predicative (8.5), we opt instead for s(np for nouns and np/(s/np) for classifiers. This analysis of predicative nominals follows Baldridge (2002)’s of nominals in Tagalog.

(8.5) winaj hach
       man  you
    You are a man

(8.6) x'apni naj winaj
    arrived  cl  man
    The man arrived

Also, we put forward the straightforward right adjunct category s\s for the adverbial ewi, ‘yesterday’:

(8.7) x'apni naj winaj ewi
    arrived  cl  man  yesterday
    The man arrived yesterday
Relative clauses are formed by appending the clause to the right of the relativized noun directly without a relativizer per se:

(8.8) \( \text{naj winaj } x^{apni} \)

\[ \begin{array}{cccc}
\text{cl} & \text{man} & \text{arrived} \\
\text{the man who arrived} \\
\end{array} \]

(8.9) \( \text{naj winaj } x^{apni} \text{ ewi} \)

\[ \begin{array}{cccc}
\text{cl} & \text{man} & \text{arrived yesterday} \\
\text{the man who arrived yesterday} \\
\end{array} \]

These contrast with (8.6) and (8.7), from which we conjectured that \( x^{apni} \)
has the type \( s/np \). To accommodate both this and (8.8) we posit a type-changing rule of the form

(\(TC\)) \( s/np \Rightarrow (s/np)\backslash(s/np) \)

That is, a rule that changes a predicate into a noun-adjunct\(^2\). This rule is illustrated here with \( \text{naj winaj } x^{apni} \) ‘The man who arrived’:

\[
\begin{array}{ccc}
\text{naj} & \text{winaj} & x^{apni} \\
\text{(s/np)} & \text{s/np} & \text{(s/np)} \backslash (s/np) \\
\text{TC} \end{array}
\]

However, the simple application rules \( >, < \) will not allow the construction in (8.9) to go through, since there is no derivation for \( x^{apni} \text{ ewi} \) as we have it so far (\( *** \) indicates that a derivation cannot go through):

\[
\begin{array}{ccc}
x^{apni} & \text{ewi} & \text{***} \\
s/np & s \backslash s \\
\end{array}
\]

The construction will go through, however, if we allow the Backwards Crossed Composition rules \( B_\times \) to apply, which would suggest that \( x^{apni} \) is minimally \( s/\times np \) and \( \text{ewi} \) minimally \( s \backslash \times s \). Then, the following is a derivation for (8.9):

\[
\begin{array}{ccc}
x^{apni} & \text{ewi} & \text{***} \\
\text{s/np} & s \backslash s \\
\end{array}
\]

\[
\begin{array}{ccc}
\text{naj} & \text{winaj} & x^{apni} \\
\text{(s/np)} & \text{(s/np)} \backslash (s/np) & \text{TC} \\
\text{np} \end{array}
\]

\[
\begin{array}{ccc}
\text{naj} & \text{winaj} & x^{apni} \\
\text{(s/np)} & \text{(s/np)} \backslash (s/np) & \text{TC} \\
\text{np} \end{array}
\]

\( **\)There is more than one way to do it: other analyses can handle this without the type-changing rule, but that does not alter the central argument here.
This provides an elegant distinction between predicative nominals (winaj) and intransitive verbs (x’apni):

(8.10) **Nominals** s/\_np

**Intransitive Verbs** s/\_np

One consequence of this analysis is that the adverb ewi can shift to the left of the subject in finite clauses, as in

(8.11) x’apni ewi naj winaj

arrived yesterday cl/the man

The man arrived yesterday

\[
\begin{array}{c}
\text{x’apni} \\
\text{s/\_np}
\end{array}
\begin{array}{c}
\text{ewi} \\
\text{s/\_np}
\end{array}
\begin{array}{c}
\text{naj} \\
\text{np}
\end{array}
\begin{array}{c}
\text{winaj} \\
\text{s/\_np}
\end{array}
\begin{array}{c}
\text{<B} \\
\text{s}
\end{array}
\]

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

8.2.2 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 \(s/(s/\_np)/np\), as in

(8.12) ha’ naj smak ix

focus him hit she

It’s him who she hit

\[
\begin{array}{c}
\text{ha’} \\
\text{s/(s/\_np)/np}
\end{array}
\begin{array}{c}
\text{naj} \\
\text{np}
\end{array}
\begin{array}{c}
\text{smak} \\
\text{s/\_np}
\end{array}
\begin{array}{c}
\text{ix} \\
\text{np}
\end{array}
\begin{array}{c}
\text{s/(s/\_np)} \\
\text{s/\_np}
\end{array}
\begin{array}{c}
\text{s/\_np}
\end{array}
\]

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 (8.12). In (8.12) 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.

Popti’ has a means to focus the subject of transitives, however, (8.13) highlights the anti-passive -ni form of smak from above. That is, 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:

(8.13) ha’ naj xmakni ix

focus cl/he hit cl/her

It’s he who hit her
This points to a strong generalization about Popti’ verb categories. In general, intransitive and transitive verb categories share a core component of the form $s/\times np$:

(8.14) Intransitive Verbs \( s/\times np_1 \)

Transitive Verbs \( (s/\times np_1)/\times np_2 \)

Note that, given that Popti’ is an ergative/absolutive language, np\(_1\) must always correspond to the the absolutive argument.

### 8.2.3 Complements and quotations

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

(8.15) *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:

\[
\begin{array}{cccc}
\text{xal} & \text{naj} & \text{jet-an} & \text{tato} \\
\frac{(s/\times cp)/\times np}{s/\times cp} & \frac{np}{np} & \frac{B_{\times cp}}{s} & \frac{\times np_2}{cp} \\
\text{jet-an} & \text{tato} & \text{x’apni ya’ cumi} \\
\frac{s/\times s}{s/\times cp} & \frac{s/\times np_2}{s} & \frac{cp}{s/\times np_2} & \frac{s}{s/\times np_2} \\
\end{array}
\]

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

\[
\begin{array}{cccc}
xal (s/\times cp)/\times np \rightarrow yalni (s/\times s)/\times np
\end{array}
\]

For example:

(8.16) x’apni ya’ cumi yalni \( naj\) jet-an

arrived cl/the lady said he to-us

He said the lady arrived
8.2.4 *mac* and long-distance extraction

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

(8.17) *mac* *xul* *ewi*

who arrived yesterday

Who arrived yesterday?

\[
\begin{array}{ccc}
  *mac & *xul & *ewi \\
  s/\phi(s/.np) & s/xnp & s/\lambda s
\end{array}
\]

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

(8.18) *mac* *xawa' ha* *melyu* *tet*

who you-gave your money to

Who did you give your money to

\[
\begin{array}{ccc}
  *mac & *xawa' & *ha \\
  s/\phi(s/.np) & np/\phi(s/.np) & *melyu \\
  s/\phi pp & np & s/\phi np >
\end{array}
\]

However, combining *mac* with the reporting verb *xal* from §8.2.3 does not go through straightforwardly. Crucially, harmonic composition is blocked by the \(\times\) modality on the verbal absolutive nucleus *s/x np*. Consider this attempt for "Who did Peter say hit Mary?":

(8.19) *mac* *xal* *naj pel* *chubil* *(xmakni ix malin)*

who said cl Peter that hit cl Mary

(attempted:) Who did Peter say hit Mary?

\[
\begin{array}{ccc}
  *mac & *xmakni & *ix malin \\
  s/\phi(s/.np) & np/\phi(s/.np) & np
\end{array}
\]

(The *** indicates that there is no rule to carry the derivation forward.) In Craig (1977)'s treatment, (8.19) 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:

(8.20) *mac* *(xmakni ix malin yalni naj pel)*

who hit cl Mary say cl Peter
Who did Peter say hit Mary?

\[
\begin{array}{ccc}
\text{xmakni} & \text{ix malin} & \text{yalni} \\
(s/\times \text{np})/\circ \text{np} & \text{np} & (s/\times \text{np})/\circ \text{np} \\
\text{mac} & > & \text{naj pel} \\
(s/\circ \text{np}) & > & s/\times \text{np} \\
\text{s/}\circ (s/\times \text{np}) & > & s/\times \text{np} \\
\end{array}
\]

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

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

8.2.5 Summary

Summing up the analysis so far, we have to following general category assignments:

<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominals</td>
<td>winnaj s/\circ \text{np}</td>
</tr>
<tr>
<td>Classifiers</td>
<td>ix, naj np/\circ (s/\circ \text{np})</td>
</tr>
<tr>
<td>Intrans. Verbs</td>
<td>x’apni s/\times \text{np}</td>
</tr>
<tr>
<td>Transitive Vbs</td>
<td>smak (s/\times \text{np})/\circ \text{np}</td>
</tr>
<tr>
<td>Temporal Adv.</td>
<td>ewi s/\times \text{s}</td>
</tr>
<tr>
<td>Focus</td>
<td>ha’ (s/\circ (s/\times \text{np})/\circ \text{np})</td>
</tr>
<tr>
<td>Complement Clause</td>
<td>tato cp/\circ s</td>
</tr>
<tr>
<td>Reporting Verbs</td>
<td>xal (s/\times \text{cp})/\circ \text{np}</td>
</tr>
<tr>
<td>Quotative</td>
<td>yalni (s/\circ \times \text{s})/\circ \text{np}</td>
</tr>
</tbody>
</table>

(8.21)

8.3 Issues with the Lexemic Analysis

A lexicographic approach such as is posited in §8.2 poses several methodological issues. In practice, for a language with significant morphology, the grammar writer must generally specify a larger number of lexical items, and with replicated 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 lexicographic grammar written for the fragment of Popti’ alluded to above must specify the redundant Transitive Verb category: ones highlighted in (8.21):

(8.22) • SubjectEmbeddedTransitiveVerb s/\times \text{np}

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

hit: TransitiveVerb { smak; }

hit-subject-focus: TransitiveVerb { xmakni; }

In fact, morphologically smak is a fairly complex form:
where asp is completive aspect and erg is the 3rd person ergative marker. A more general morphemic system that can also handle word-internal structure is proposed in the following section.

### 8.4 Morphemic Approach

A morphemic approach (Bozsahin, 2002) differs significantly from a lexic morphemic 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.

To enable this in part, the analysis below makes use of *inert slashes* /!, \!. A category specified with an inert slash, say, s/!np, may not combine its argument (here, np) as usual, may be regarded as an atomic category decorated with structural and semantic information.

The morphemes under consideration, then, are:

**Aspect** Aspect markers combine with a non-finite, inert VP to the right to produce a finite, aspectually specified VP. Here is the category for the completive aspect marker:

• \( x-: (s/_{x}np)/_s(s/_{x}np) \)

**Ergative (incorporated) pronouns** Incorporated pronouns consume the outer (first) argument of a transitive verb. Leaving out the details, their category is \( (s/_{x}np)/_s((s/_{x}np)/_snp) \). That is, they take a non-finite transitive verb to the immediate right and render an intransitive verb. Specifically:

• \( hin- \) ‘I’, \( ha- \) ‘you’: \( (s/_{x}np)/_s((s/_{x}np)/_snp) \)

**Transitive Verbs** Transitive verbs, without aspect, agreement or embedded pronouns are simply assigned a category identical to that specified before, only now with inert slashes:

• \( il \) ‘to see’, \( mac \) ‘to hit’: \( (s/_{x}np)/_ilnp \)

The aspect and incorporated pronouns enable derivations of sentences that exhibit the shifting arity of verbs at work. In the following is example, the transitive verb \( il \) ‘to see’, is rendered as a pseudo-intransitive \( x-hin-il \), ‘I saw ...’:

(8.24) \( x- \ hin- il \) naj winaj

asp pro see cl man

I saw the man
**Ergative agreement** Ergative agreement takes a transitive verb to the immediate right and basically “turns-on” the verb’s ergative argument:

- $s$: $(s/\overline{\chi} np_1)/\overline{\chi} (s/\overline{\chi} np)/\overline{\chi} np_2$  

This category also checks that the first NP in its delexical sentence is of ergative voice.

By contrast with incorporated pronouns, the 3rd person ergative marker does not change the arity of a transitive verb, as in:

(8.25) $x-s$-mak ix malin naj pel

asp erg hit cl Mary cl Peter

Mary hit Peter

As in this and example (8.2.2) above, the desired category for $x-s$-mak is $(s/\overline{\chi} np)/\overline{\chi} np$. This is given by:

$$
\begin{align*}
\text {Ergative agreement} & \quad \text {Ergative agreement} \\
\frac{x-}{s/np \langle s/\chi np \rangle} & \quad \frac{\text {hin-}}{(s/\chi np)/\overline{\chi} ((s/\chi np)/\overline{\chi} np)} \quad \frac{\text {il}}{(s/\chi np)/\overline{\chi} np} \\
\text {hin-} & \quad s/\chi np \quad \text {il} \quad s/np \\
\text {naj winaj} & \quad np \\
\text {hit} & \\
\text {Mary hit Peter} & \\
\end{align*}
$$

**Subject focus** The anti-passive or subject-focus morpheme -ni takes a transitive verb to the immediate left and, in effect, switches the order of the arguments, so as to make the absolutive argument to the verb accessible to the focus operator $ha^*$, as sketched in §8.2.2:

- $\cdot ni$: $(s/\chi np_1)/\overline{\chi} np_2$  

Note: $np_1$ is shorthand for $[np(1)]$.

With this suffix we can now specify a full analysis of an instance of subject focus with the focus operator $ha^*$:

(8.26) $ha^*$ ix malin $x$-mak -ni naj pel

foc cl Mary asp hit subj-foc cl Peter

It was Mary who hit Peter

The basic derivation is:

$$
\begin{align*}
\text {Subject focus} & \quad \text {Subject focus} \\
\frac{ha^*}{(s/\chi (s/\chi np_2))/\overline{\chi} np_2} & \quad \frac{\text {ix malin}}{(s/\chi np)/\overline{\chi} np_2} \quad \frac{\text {x-mak -ni}}{(s/\chi np)/\overline{\chi} np_1} \quad \frac{naj pel}{np_1} \\
\text {s/\chi (s/\chi np_2)} & \quad np_2 \quad np_2 \quad np_1 \\
\text {s/np_2} & \quad s/\chi np_2 \\
\text {s} & \quad s/np_2 \\
\text {Mary} & \quad Peter \\
\end{align*}
$$
The indices indicate the reversal of the verbal arguments, specific to the analysis of subject extraction above. The complex verbal form $x$-$mac$-$ni$ together with the derivation reversing its arguments is given here:

$$
\begin{array}{c}
\text{mak} \\
(s/x, np_1)/o np_2 \\
\hline
\text{-ni} \\
((s'/x, np_2) / o np_1)((s'/x, np_1) / o np_2) \\
\hline
s/x np_2 /o np_1
\end{array}
$$

$<B$

8.4.1 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.

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’ (§8.2.2), 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.

8.5 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, what it gains in generality and elegance, it loses in efficiency in parsing, as it has more work to do parsing the same material. 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


