UCLA Working Papers in Linguistics
Volume 15, Article 3: 1–18, 2010

Little Tagalog

Free Word Order in Bare Grammar

Meaghan Fowlie

Bare Grammars are a simple and straight-forward model for syntax. Normally, the lexicon in such a grammar is kept very minimal: each lexical item is an ordered pair consisting of the string and its category. However, the simplicity of such a lexicon can make formulating rules that are succinct and that capture grammatical regularities impossible. In this paper I propose that adding more information to the lexicon can make it possible to write rules that are succinct and capture more generalisations. The second part of this paper makes use of this additional information in the lexicon to formulate a grammar for free word order. I propose that it can be accounted for in bare grammars if we allow rules to form not only concatenated strings (which by definition have order) but also sets (which by definition do not).

Keywords free word order, Tagalog, Bare Grammar, sets, multisets

Introduction

This work attempts two feats:

1. To "reify" bare grammar subscripts and other notations that mnemonically relate categories
2. To account for genuinely free word order without appealing to multiple derivations

Bare grammars (Keenan and Stabler 2003) typically define the Lexicon as a set of ordered pairs (string, cat), where *string* is the phonological form of the word and *cat* it the category of the word.

Example. (banana, N)

Sometimes it is convenient to give two separate categories similar names to help the linguist remember that the categories have something important in common. For example, a language like Spanish, which distinguishes masculine and feminine nouns, might have two categories named Nm and Nf, for Noun (masculine) and Noun (feminine). However, these categories will be treated by the derivation as being as different as any other two categories. This means that if we want a rule to apply to both masculine and feminine nouns, we will have to write two separate rules.

Often these separate rules are written as one rule, for readability. In these cases the rules are written as if it is possible, for example, to let a variable range over {m, f}. However, this is really just notational shorthand for two rules.

Merge (notational shorthand): $((s, \mathrm{Ax}), (t, \mathrm{Nx})) \longmapsto (s^\frown t, \mathrm{Nx}),\ x \in \{\mathrm{m}, \mathrm{f}\}$

Merge (full version): $\begin{cases} (\langle s, \mathrm{Am}\rangle, \langle t, \mathrm{Nm}\rangle) \longmapsto \langle s^\frown t, \mathrm{Nm}\rangle \\ (\langle s, \mathrm{Af}\rangle, \langle t, \mathrm{Nf}\rangle) \longmapsto \langle s^\frown t, \mathrm{Nf}\rangle \end{cases}$

This paper proposes that these "fake" variable, like x above, be made real, so that rules that refer to just the N part of Nf and Nm can be written.

The second part of this paper proposes a syntax for free word order phenomena within a Bare Grammar framework. I will apply the first proposal, that subscripts be reified, to this problem. The grammar will consist of a finite vocabulary V, a set *Rule* of functions for combining elements of the vocabulary, a set *Cat* of categories linked to the vocabulary items, and a set θ of theta roles, also linked to the vocabulary items.

We will use a toy grammar modelled on aspects of Tagalog as an example of free word order. Vocabulary items will be linked not only to a category as is done in Keenan and Stabler (2003), but also with a third element that encodes further information about the word, regarding the theta role it plays or requires in the sentence. It is this last set, θ, that makes use of the first proposal. The inclusion of θ is a way of keeping track of the number arguments that will later be linked to the verb.

1 A New Element for the Grammar

1.1 A Problem: Subtypes are not Subtypes

In X-bar and related theories, lexical categories, such as N, are related to the phrases that contain them, such as NP, by projection. When two elements merge, one of the two projects, meaning among other things that the enclosing phrase behaves much like the projecting category in terms of selection. A verb looking for a noun can in fact look for an NP.

However, in bare grammars, projecting categories are not usually related to the lexical categories except through the rules that make phrases from lexical categories. With only two dimensions on which to describe lexical items – string and category – lexical and projecting categories are only related if they are the same; that is, if adding an adjective or article to a noun leaves the category as N, rather than something new, like an N' or an NP.

When, say, a verb looks for a noun phrase, the fact that the noun phrase was built from a noun is obscured. We also have no way to relate nouns that differ in morphological marking, such as case or gender. Once the rule that morphologically marks the nouns has applied, the fact that they are indeed nouns is normally hidden. This means that a verb that can take either a masculine or feminine noun needs two separate rules, forcing conceptually redundant rules and again obscuring the relationship between the gender-marked nouns.

For example, in Little Spanish, given as a model of gender marking in Keenan and Stabler (2003), *Cat* includes Nf (Noun-feminine) and Nm (Noun-masculine), as well as A, Af, and Am (Adjective, Adjective-feminine, Adjective-masculine). Adjective Modification (**AM**) is written with a notational shorthand which refers to the m and f, but m and f are not really part of the grammar. The A in Am and Af has no real meaning; it only serves to remind us, the linguists, which familiar categories we are dealing with.

AM (notational shorthand): $((s,\mathrm{Ax}),(t,\mathrm{Nx})) \longmapsto (s^\frown t,\mathrm{Nx}),\ x \in \{\mathrm{m},\mathrm{f}\}$

AM (full version): $\begin{cases} (\langle s,\mathrm{Am}\rangle,\langle t,\mathrm{Nm}\rangle) \longmapsto \langle s^\frown t,\mathrm{Nm}\rangle \\ (\langle s,\mathrm{Af}\rangle,\langle t,\mathrm{Nf}\rangle) \longmapsto \langle s^\frown t,\mathrm{Nf}\rangle \end{cases}$

If we remove the mnemonic names, and let $\mathrm{Am}=\alpha,\mathrm{Af}=\beta,\mathrm{Nm}=\gamma,\mathrm{Nf}=\kappa$, Merge looks like this:

AM: $\begin{cases} (\langle s,\alpha\rangle,\langle t,\gamma\rangle) \longmapsto \langle s^\frown t,\gamma\rangle \\ (\langle s,\beta\rangle,\langle t,\kappa\rangle) \longmapsto \langle s^\frown t,\kappa\rangle \end{cases}$

Now these four categories are clearly related only by what rules they undergo. While this is, in a sense, exactly what categories are, it does not capture the intuition that the same kinds of things happen to both categories when they undergo the same rule. In these rules, in both cases the first argument's string precedes the second argument's string, and the category of the image is the category of the second argument. The semantics will also be the same: the meaning of both will be whatever the meaning of a modified noun is.

Moreover, despite the intuition that masculine and feminine nouns are both nouns, the lexicon does not relate them at all. The idea is that there are relationships between categories, both intuitive and formal, that are not captured by a lexicon that is a subset of a cross product of only two sets.

Additionally, it is worthwhile to consider what kinds of rules are possible under this system, as well as what kinds of rules might be built in parallel or on analogy with existing rules. As we have it, there are almost no restriction on the possible rules, but one could examine the patterns of existing rules and create restrictions. Simpler here would be to create a rule on analogy with an existing rule, under the supposition that if it is similar enough it will fall under most restrictions one might want to place on *Rule*.

For example, Little Spanish's adjective modification is meant to capture a relationship between two categories, adjectives and nouns, restricted by agreement requirements. However, since there is no such category as "noun" and the adjectives that combine with Nfs and Nms are also no longer adjectives but Afs and Ams, in fact adjective modification combines two unrelated pairs or categories. Why not, then, add a third unrelated set of categories, as would be necessary if there were a third gender? Why not, say, V and A? Such a move would certainly be in keeping with the style of rule, but it would not capture a related modification.

This example also illustrates how unsuccinct this style of rule can be: what if the language has a great number of noun classes, all of which behave the same under adjective modification, i.e. the noun and ajdective must match in gender? There will have to be a great number of subcases of the adjective modification rule.

In order to better capture intuitions about category, and to allow for more succinct rules, I propose that such categories be related in the lexicon. Let us consider what happens if we allow *Lex* to include ordered *triples* as well as ordered pairs. The third element essentially represents useful subscripts, such as case (a, n), gender (f, m), or predicate arity (0, 1, 2).

In the example of gender, not only can the rule now have only one case, but there is a very simple way to describe the gender-agreement: the third element of the noun must be identical with the third element of the adjective.

AM (modified): $(\langle s, \mathrm{A}, x\rangle, \langle t, \mathrm{N}, x\rangle) \longmapsto \langle s^\frown t, \mathrm{N}, x\rangle, \forall x \in \{f, m\}$

Notice this is almost exactly the notational shorthand for the rule given in Keenan and Stabler (2003).

Suppose, by way of illustration, we let this third element simply be a number representing case or predicate arity. Then we can match the verbs with the right number of arguments with the right case-marking. Consider the following simplified model.

1.1.1 Verbs

Verbs come inherently intransitive, transitive, or ditranstitive. Rather than calling these P1, P2 and P3, I propose separating the number from the P, allowing math to be done on it, for example $\langle \mathit{sneezed}, \mathrm{P}, 1\rangle$, $\langle \mathit{admired}, \mathrm{P}, 2\rangle$, and $\langle \mathit{gave}, \mathrm{P}, 3\rangle$.

Now the arity of the predicate – the number of arguments required by the predicate – can be accessed formally. We will see below, for example, that the arity of the verb will be reduced as arguments are merged, and the case of the arguments will be selected accurately because the case number will match the gradually lowering arity number, which is exactly the intuition to be captured.

1.1.2 Case

Nouns are generally listed in Lex_0 as ordered pairs, and when case-marked also take a number as a third element. Suppose case-markers in *Lex* look like this: $\langle$-NOM, K, 1$\rangle$, $\langle$-ACC, K, 2$\rangle$ and Case Marking is a rule that takes a K and an N and yields a case-marked noun.[1]

CM: $(\langle s, \mathrm{N}\rangle, \langle t, \mathrm{K}, n\rangle) \longmapsto \langle s^\frown t, \mathrm{N}, n\rangle$

Example. $\langle \mathit{man}, \mathrm{N}\rangle, \langle$-NOM, K, 1$\rangle \longmapsto \langle \mathit{man}$-NOM, N, 1$\rangle$

With the case type *1* separated from the categories K and N, each part of the image under CM is taken directly from a part of the preimage. The string is concatenated

[1] This is not the only possible way to relate un-case-marked and case-marked nouns, of course, but this works, and does have a certain logic: both are nouns in that they have a second element *N*, but they differ in that case-marked nouns carry more information, instantiated in their having a third element.

from the strings of the input, the category is one of the categories, and the case number is the third element of one of the pairs in the input. Nothing new need be introduced in the output: everything was already present in the input. This means that the grammar is manipulating only what it is given. Nothing is being pulled out of the aether.

CM: $(V^* \times Cat) \times (V^* \times Cat \times \mathbb{N}) \to (V^* \times Cat \times \mathbb{N})$,
$\text{CM}(A,B) = (\langle [A]_1, [B]_1 \rangle, [A]_2, [B]_3)$

That is, the image of the pair (A,B) under CM is the ordered triple consisting of 1) the sequence of the first elements of A and B, 2) the second element of A, and 3) the third element of B.

This is, if nothing else, more intuitive than a grammar in which new things are introduced in the output. Although the definition of a function is not constrained at all by such considerations, such functions are arguably more elegant and more intuitive. A possible restriction on *Rule* presents itself. Just as the rules of Little Spanish never have a string in the output that was not present in the input, perhaps bare grammars should not have rules with categories in the output that were not present in the input. Such a restriction would not be possible without separating case or gender from noun.

By way of comparison, suppose we defined an operation CM2 without the ordered triples. Let further $\text{N} = \alpha$, $N(\textsc{nom}) = \beta$, $N(\textsc{acc}) = \gamma$, Case marker($\textsc{nom}$) $= \kappa$, Case marker($\textsc{acc}$) $= \delta$ to make the separateness of the categories clear.

CM2: $\begin{cases} (\langle s, \alpha \rangle, \langle t, \kappa \rangle) \longmapsto \langle s^\frown t, \beta \rangle \\ (\langle s, \alpha \rangle, \langle t, \delta \rangle) \longmapsto \langle s^\frown t, \gamma \rangle \end{cases}$

Not only are there now two quite different lines to our definition of CM, but new elements are introduced in the image that were not present in any part of the preimage. Only the first element, the sequence of the first elements of the input, can be copied from the input.

CM2: $(V^* \times Cat) \times (V^* \times Cat) \to (V^* \times Cat)$,
$$\text{CM2}(A,B) = \begin{cases} \langle [A]_1, [B]_1, \beta \rangle \text{ if } [B]_2 = \kappa \\ \langle [A]_1, [B]_1, \gamma \rangle \text{ if } [B]_2 = \delta \end{cases}$$

1.1.3 Verbs and Arguments

Consider a Little English with such elements as $\langle admired, \text{P}, 2 \rangle$, $\langle him, \text{N}, 2 \rangle$, and $\langle she, \text{N}, 1 \rangle$. Then we can write a Predicate-Argument function (**PA**) like this:

PA: $(\langle x, \text{P}, n \rangle, \langle y, \text{N}, n \rangle) \longmapsto \langle x^\frown y, \text{P}, n-1 \rangle \qquad n \in \mathbb{N}$

Notice that the third elements of the domain pair must match, and the image's third element, while not copied from the preimage, is calculated from the third elements in the preimage using simple natural number arithmetic. The match between the predicate arity and the case marker is the reason I defined both in numbers. Each argument reduces the arity of the predicate until it is a P0, a sentence. We want the

object to have the same type as a transitive verb because it will combine with the verb to form a P1, which is looking only for a nominative. A nominative is of type *1*.

Now we can see in the simple case that a P1 takes as argument a nominative-marked noun phrase. We have defined NP-NOM as an ordered triple $\langle s, \mathrm{N}, 1\rangle$.

Example. $(\langle \textit{sneezed}, \mathrm{P}, 1\rangle, \langle \textit{he}, \mathrm{N}, 1\rangle) \longmapsto \langle \textit{he sneezed}, \mathrm{P}, 0\rangle$

Similarly, *She admired him* can be built up using this one formulation of the PA rule, modulo word order.[2]

(1) a. $(\langle \textit{admired}, \mathrm{P}, 2\rangle, \langle \textit{him}, \mathrm{N}, 2\rangle) \longmapsto \langle \textit{admired him}, \mathrm{P}, 1\rangle$
b. $(\langle \textit{admired him}, \mathrm{P}, 1\rangle, \langle \textit{she}, \mathrm{N}, 1\rangle \longmapsto \langle \textit{She admired him}, \mathrm{P}, 0\rangle$

Adding a third element to the Lexicon makes *Rule* much more succinctly and mathematically definable, and captures intuitions not otherwise captured.

2 Tagalog

Tagalog has free word order in a number of constituents. I will be using multisets in the free word order proposal. Following is a brief introduction thereto.

2.1 Multisets

Intuitively, a multiset is a set in which elements can be repeated. In a regular set, $\{1,2,3\} = \{1,1,1,1,1,2,2,3\}$. If these were multisets, they would not be the same set.

Since sets are defined by their elements, technically a multiset cannot be thought of as a set per se, although we will treat them as though they are for simplicity of notation. Instead, a multiset is perhaps best thought of as a function.

Definition 1 (Multiset)**.** A *multiset of a set* S is a map $m\colon S \to \mathbb{N}$ from S to $\mathbb{N} = \{0,1,2,\ldots\}$

The intuition is that x is "in" $m(x)$ iff $m(x) > 0$. Moreover, the number of x's in the multiset is the value of m at x.

I will notate multisets just as I would real sets, with the understanding that if an element is repeated in the list notation, the number of times it appears is its image under m. In other words, I will notate multisets intuitively.

Here are some definitions of set theoretic notions for multisets. For any multisets M, P of S:

Membership: $s \in M$ iff $M(s) > 0$

Equality: $M = P$ iff $\forall s \in S, M(s) = P(s)$

Union: $(M \cup P)\colon S \to \mathbb{N}$ s.t. $(M \cup P)(s) = M(s) + P(s)$

[2] The word order problem I don't care about right now, as I'm working on Tagalog! But certainly it is an issue.

Intersection: $(M \cap P)\colon S \to \mathbb{N}$ s.t. $(M \cap P)(s) = M(s) \wedge P(s)$

Difference: $(M - P)\colon S \to \mathbb{N}$ s.t. $(M - P)(s) = M(s) - (M(s) \wedge P(s))$

Subset: $M \subseteq P$ iff $\forall s \in S, M(s) \leq P(s)$

Let us now turn to Tagalog.

2.2 Tagalog Data

Tagalog shows free word order in most of the sentence, and does not seem to have any preferred order or change in meaning. Generally, V comes first, followed by everything else, in any order (data from Kroeger 1993).

(2) *Nagbigay ng-libro sa-babae ang-lalaki*
gave GEN-book DAT-woman NOM-man
'The man gave the woman a book'
Nagbigay ng-libro ang-lalaki sa-babae
Nagbigay sa-babae ng-libro ang-lalaki
Nagbigay sa-babae ang-lalaki ng-libro
Nagbigay ang-lalaki sa-babae ng-libro
Nagbigay ang-lalaki ng-libro sa-babae

When an adjective is added, the case marker is consistently in first position within the DP, but the adjective and noun can be in either order with the same meaning.

(3) a. *ng libro-ng malaki*
GEN book-LK big
'the big book'
b. *ng malaki-ng libro*
GEN big-LK book

When the DP *the big book* is ordered as in (3-a), the three DPs yield the usual six orders as in (2). Similarly, when *the big book* is ordered as in (3-b), these DPs can be arranged in six possible orders, yielding a total of twelve possible word-orders.

2.3 Proposal: Two Combining Operations

I propose that in Tagalog there are two ways to form expressions: concatenation and multiset formation. When word-order matters, elements are concatenated. When it is free, they form multisets. (They must be multisets rather than regular sets given that we don't want two identical strings to be treated as one: *Some rabbit killed some rabbit* ≠ *Some rabbit killed*.)

Because sets, unlike sequences (i.e. concatenation), have no inherent order defined on them, I propose that free word order arises when merge forms sets rather than concatenating. The unordered set elements can appear in any order, but only one derivation will be required for all possible orders.

In order for these multisets to be concatenated with other elements, including other multisets, Tagalog's version of concatenation must be *multiset* concatenation. Since concatenation is just sequence formation, we need only say that the elements of the sequence are multisets. Since both multiset concatenation and multiset formation need to be able to apply to any element of the language, everything must be multiset formation in some sense. Therefore multiset concatenation rules will have as output a *multiset* containing a sequence of multisets. This may not always be made explicit as we go, as the brackets get quite messy, but keep in mind that every element of the language is a multiset.

Ordered: multiset concatenation

Free: multiset formation

I also claim that Tagalog Vocabulary consists of unit multisets of words rather than words. This allows a single formulation of each $f \in Rule$, without having to reformulate it when the arguments are outputs of other rules. We will see later on why this is desireable.

3 Verbs and Arguments in Tagalog

3.1 V-Initial

Before we go on to the actual grammar and examples, we must look at the V-initial nature of Tagalog. Normally, one would probably say that V starts lower in the sentence, and moves up to C or T, leaving behind its arguments to scramble. Since we lack a theory of movment, I reformulated it under the assumption that the verb really does combine late. Therefore rather than the usual cumulative gathering of arguments by the verb, I propose that the arguments gather themselves into a multiset. Call the Rule *Argument Linking* (AL).

Now, the required noun phrases of the verb are determined by a combination of case marking and theta role. There is not a one-to-one relationship between case marking and theta roles in Tagalog, which presents a difficult and interesting problem.

3.2 The ang *NP*

Every sentence of Little Tagalog, and indeed, nearly every sentence of real Tagalog, has one NP which is marked with the case marker *ang*. There is debate about what exactly the *ang*-marked element is (whether it be a topic, "subject", or something else entirely (see Kroeger 1993 for discussion), but whatever it is, it can be any NP, argument or adjunct, agent or object. The verb is voice-marked to indicate the theta-role of the *ang*-marked NP. The remaining NPs will be marked with *sa*, which Kroeger calls genitive, or *ng*, which he calls dative. The case markers are divided up by theta role. Some sentences have NPs marked with the same case.

I propose that the verb starts with a theta grid, which determines the case of the NP(s) it needs. Moving away from numbers, I propose that the theta roles are

grammatical primitives, i.e. $\theta :=$ {actor, theme, goal, recipient, locative, instrumental, benefactive, possessor}[3], which I abbreviate as $\theta := \{a, t, g, r, l, i, b, p\}$. These theta roles will need to be grouped into sets, so let $\mathbb{K} = \wp(\theta)$. Now every verb can be given not only an arity but a theta grid, for example $\langle gave, \mathrm{P}, \{a, t, r\}\rangle$, $\langle cooked, \mathrm{P}, \{a, t\}\rangle$, $\langle sneezed, \mathrm{P}, \{t\}\rangle$.

Voice-marking the verb maps a theta-role to a case-marker, specifically to whatever *ang* is. Following Kroeger I will call it NOM. Any theta role can be nominative, but the other two case markers partition θ. Let $\textsc{gen} = \{a, t, p, i, b\}$, $\textsc{dat} = \{l, g, r\}$, and $\textsc{nom} = \theta$.

Definition 2 (f_θ)**.** Let $f_\theta \colon \wp(\theta) \to \{\textsc{gen}, \textsc{dat}, \emptyset\}$ be a partial function defined as follows:

$$f_\theta(X) = \begin{cases} \emptyset & \text{iff } X = \emptyset \\ Y & \text{iff } X \subseteq Y \text{ and } X \neq \emptyset \end{cases}$$

.

That is, f_θ takes sets of theta roles and maps them to their associated case.

We lift f_θ to g_θ which maps sets of theta-roles to multisets of their associated case.

Definition 3 (g_θ)**.**

$$g_\theta(X) = \{f_\theta(\{x\}) | \{x\} \subseteq X\}$$

Schematically, f_θ and g_θ are as follows:

$$f_\theta(\{\theta_1, ..., \theta_n\}) = Y \iff \{\theta_1, ..., \theta_n\} \subseteq Y$$

$$g_\theta(\{\theta_1, ..., \theta_n\}) = \{f_\theta(\{\theta_1\}), ..., f_\theta(\{\theta_n\})\}$$

Note that f_θ and g_θ are single-valued (and therefore functions) since GEN and DAT partition $\wp(\theta)$. Note also that no set containing $\emptyset$ is in the range of g_θ since $\emptyset = f_\theta(\emptyset)$, and $\emptyset$ is not a singleton set as required by g_θ.

3.3 Voice-Marking

I now add to the grammar category Voi and lexical items voice-marking affixes.[4] The third coordinate is the theta role (or set of theta roles) that will be *ang*-marked.

$$\langle \mathrm{AV}, \mathrm{Voi}, \{a\}\rangle \quad \langle \mathrm{TV}, \mathrm{Voi}, \{t\}\rangle \quad \langle \mathrm{IV}, \mathrm{Voi}, \{i\}\rangle \quad \langle \mathrm{BV}, \mathrm{Voi}, \{b\}\rangle \quad \langle \mathrm{DV}, \mathrm{Voi}, \textsc{dat}\rangle$$

The verb can now be instructed to look for appropriately case-marked nouns to take as arguments. Here is the rule *Voice Mark* (VM), with its subfunction $v_S(T)$ defined below.

[3] Simplification: real Tagalog allows themes to be -DAT or -GEN, interpreting one as definite and the other as indefinite.

[4] In real Tagalog, these are prefixes, suffixes and infixes. To simplify, I am treating them as suffixes here, as they need to stick to the verb.

VM: $(\langle x, \text{P}, T\rangle, \langle y, \text{Voi}, S\rangle) \longmapsto \langle x^\frown y, \text{P}, v_S(T)\rangle$

Since voice-marking already maps a theta role to case, let us define it to map all the verb's theta roles to their appropriate case. For this, we must define a function, v_S for each $S \in \mathbb{K}$.

$$v_S(T) = (g_\theta(T) - \{f_\theta(T \cap S)\}) \cup \{\text{NOM}\}$$

The function v_S maps the set of theta roles of the verb to the multiset of their associated cases, if necessary subtracts the case associated with any argument theta role it is voice-marked for, and adds NOM (*ang*). If the voice-marker marks the verb for an argument of the verb, $T \cap S \neq \emptyset$. Then the difference operator removes that case from the set the verb is looking for and replaces it with NOM.

If the voice marker is for what would normally be an adjunct, $T \cap S = \emptyset$ so it is just added into the set of required NPs for the verb as a NOM.

Looking more closely at v_S, consider the case in which the verb is voice-marked for an *argument*. Recall that T is the theta grid of the verb and S the set of theta roles associated with the voice marker.

1. The set of theta roles T in the theta grid of the verb are mapped to their corresponding cases by g_θ.
2. $T \cap S$ is the set of theta-roles the verb and voice-marker have in common. Normally this is a singleton set, unless the voice marker is dative, in which case it will be some subset of DAT.
3. $f_\theta(T \cap S)$ is the case-marker associated with the voice-marker, since $T \cap S \subseteq S$ and f_θ is defined interms of subsets.
4. The case for the voice marker is subtracted from the set of cases the verb is seeking.
5. The subtracted case is replaced by NOM.

Now the verb is looking for the same number of arguments, but it knows which cases it needs, including one NOM, which replaced one of the original cases the un-voice-marked verb was seeking.

Now consider the case when the verb is voice-marked for a *non-argument* of the verb. $T \cap S = \emptyset$ so v_S just maps the theta grid to the corresponding cases, and adds NOM.

1. The set of theta roles in the theta grid of the verb are mapped to their corresponding cases by g_θ.
2. $\{f_\theta(\emptyset)\} = \{\emptyset\}$. $\emptyset$ is never a member of $g_\theta(T)$ so the difference operation changes nothing.
3. NOM is added.

Result of v_S: a multiset of the cases of the nouns the verb will look for.

Example. Below we see an instrumental-marked verb. The instrumental theta role is not part of the theta grid of the verb. The usual theta roles are retained, but the verb is now looking for four, not three, NPs, as one has been added by the voice marker.

$$\text{VM}(\langle \textit{gave}, \text{P}, \{a, t, r\}\rangle, \langle \text{IV}, \text{Voi}, \{i\}\rangle) = \langle \textit{gave-IV}, \text{P}, \{\text{GEN}, \text{NOM}, \text{GEN}, \text{DAT}\}\rangle$$

Here $v_{\{i\}}(\{a,t,r\})$ is calculated as follows. We have $\{i\} \cap \{a,t,r\} = \emptyset$, so

$$\begin{aligned} v_{\{i\}}(\{a,t,r\}) &= (g_\theta(\{a,t,r\}) - f_\theta(\emptyset)) \cup \{\textsc{nom}\} \\ &= g_\theta(\{a,t,r\} \cup \{\textsc{nom}\} \\ &= \{\textsc{gen}, \textsc{gen}, \textsc{dat}\} \cup \{\textsc{nom}\} \\ &= \{\textsc{gen}, \textsc{gen}, \textsc{dat}, \textsc{nom}\} \end{aligned}$$

Example. The verb below is dative-marked. The theta role which would normally be dative is now nominative. This is acheived by subtracting DAT from the set of cases sought and replacing it with NOM.

$$\text{VM}(\langle \textit{gave}, \text{P}, \{a,t,r\} \rangle, \langle \text{DV}, \text{Voi}, \textsc{dat} \rangle) = \langle \textit{gave-DV}, \text{P}, \{\textsc{gen}, \textsc{nom}, \textsc{gen}\} \rangle$$

Recall that $\textsc{dat} = \{l, g, r\}$.

$$\begin{aligned} v_{\textsc{dat}}(\{a,t,r\}) &= (g_\theta(\{a,t,r\}) - f_\theta(\{l,g,r\} \cap \{a,t,r\})) \cup \{\textsc{nom}\} \\ &= (g_\theta(\{a,t,r\}) - f_\theta(\{r\})) \cup \{\textsc{nom}\} \\ &= (\{\textsc{gen}, \textsc{gen}, \textsc{dat}\} - \{\textsc{dat}\}) \cup \{\textsc{nom}\} \\ &= (\{\textsc{gen}, \textsc{gen}\}) \cup \{\textsc{nom}\} \\ &= \{\textsc{gen}, \textsc{gen}, \textsc{nom}\} \end{aligned}$$

Note that the difference between a voice-marked verb and an un-voice-marked verb is the third coordinate. Before a verb is voice-marked, its third coordinate is an element of $\mathbb{K}$. A voice-marked verb's third coordinate is an element of $\wp(\mathbb{K})$. The verb is now set up to seek appropriately case-marked NPs. Let us now turn to the nouns.

3.4 Nouns

Nouns are listed in the lexicon as simple ordered pairs $\langle s, N \rangle$. Case markers are ordered triples which have as their third coordinate a subset of $\mathbb{K}$. We add three items to the lexicon: $\langle \textit{ang}, \text{K}, \textsc{nom} \rangle$, $\langle \textit{sa}, \text{K}, \textsc{gen} \rangle$, $\langle \textit{ng}, \text{K}, \textsc{dat} \rangle$. And we add CM as defined in Sec. 1.1.2 to *Rule*.

As mentioned in section 2.3 above, in order for the free-order operation (multiset formation) and our ordered operation (multiset concatenation) to work together, everything must be multisets. Lexical items must in fact be singleton multisets. Multi-set concatenation must form multisets consisting of sequences of multisets. This is of particular importance in the next rule, *Argument Linking*.

Since Little Tagalog is verb-initial, the set of nouns must be linked together before the verb selects them. This means that the verb will select the whole group of nouns together, rather than picking them up one at a time. Most of the constraints on this process will actually be on the result of the rule PA that links the nouns to the verb. Nouns will be allowed to combine quite freely, but only certain sets of them will be able to be taken as an argument set by the verb. The only restriction here is that there may not be more than one nominative (*ang*-marked) NP.

AL: $(\langle x, \text{N}, k \rangle, \langle y, \text{N}, l \rangle) \longmapsto \langle x \cup y, \text{N}, k \cup l \rangle$ if $(k \cup l)(\textsc{nom}) \leq 1$

AL creates a multiset of nouns, accompanied by a multiset of their case-markers.

3.5 Putting it all Together: Predicate-Argument Operation

Finally, the voice-marked V must select its arguments.

PA: $(\langle x, \mathrm{P}, K\rangle, \langle y, \mathrm{N}, L\rangle) \longmapsto \langle x^\frown y, \mathrm{P}, K - L\rangle$ for $K, L \in \wp(\mathbb{K})$

The idea is that if L contains all the cases the verb is looking for, the result is a P0. Otherwise it is not a fully saturated verb, and cannot be a licit sentence. Any non-nominative adjuncts in L are ignored, since the difference operator does nothing with elements of L not also in K.

Example. Figures 3–6 depict how the sentence meaning *The man gave the woman the book* with the agent *man* as the *ang*-phrase – i.e. example (2) on page 7 – is built from the argument set {{ng} ⌢{woman}, {sa} ⌢{book}, {ang} ⌢{man}}, N, {NOM, GEN, DAT}⟩.

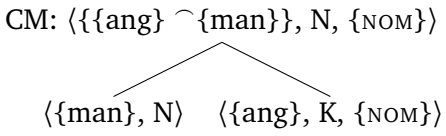

Figure 1: Case-marking *man*

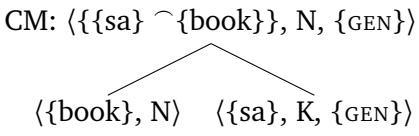

Figure 2: Case-marking *book*

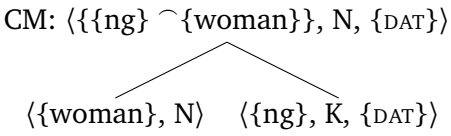

Figure 3: Case-marking *woman*

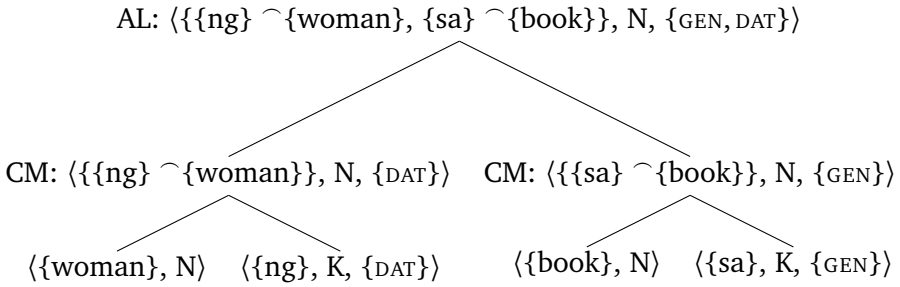

Figure 4: Linking the arguments *woman, book*

(Continued in figures 5 and 6.)

3.6 Adjectives

Recall that Tagalog can freely order adjectives with nouns, but adjectives and nouns must stay together. Please ignore the linker *-ng*, which will be excluded from Little Tagalog (data from Raphael Marcado, p.c. 2007).

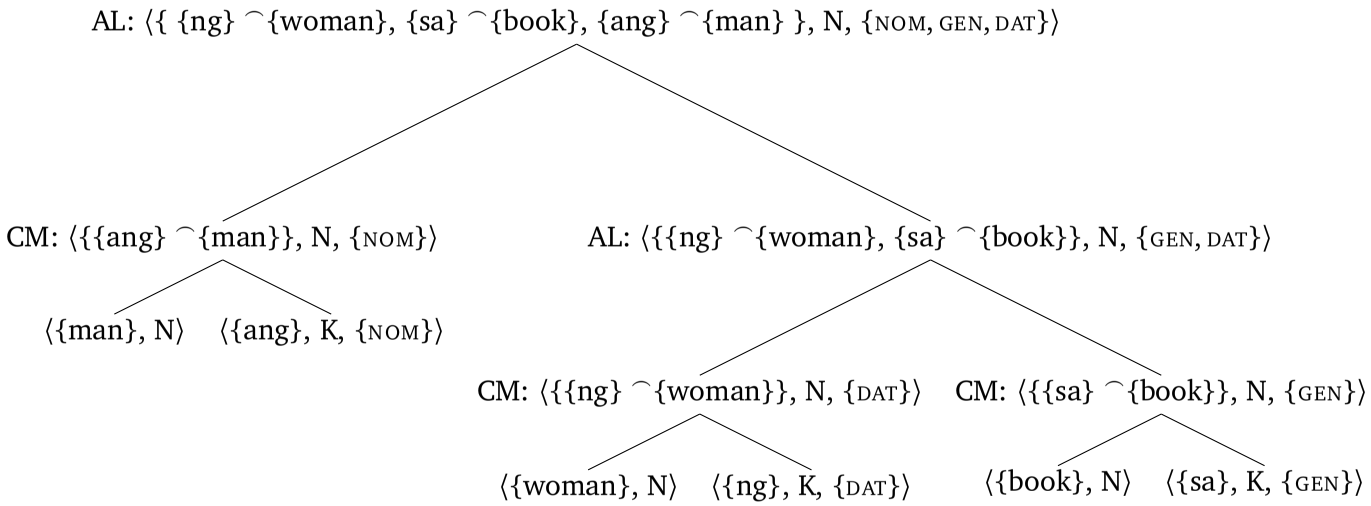

Figure 5: *man, woman, book* argument-linked

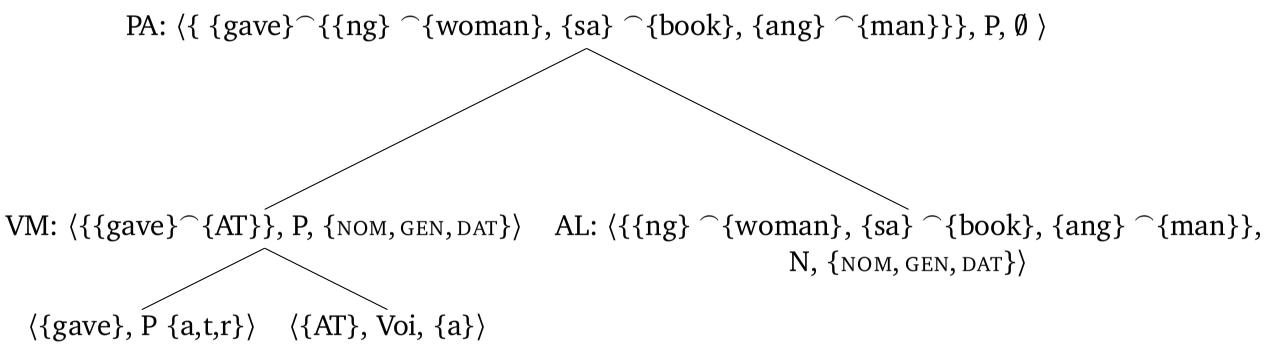

Figure 6: Adding the verb to build the full sentence *the man gave the woman the book*

(4) a. *ng libro-ng malaki*
GEN book-LK big
'the big book'
b. *ng malaki-ng libro*
GEN big-LK book

I propose a rule of Adjective Modification, thus:

AM: $(\langle x, \mathrm{N}\rangle, \langle y, \mathrm{A}\rangle) \longmapsto \langle x \cup y, \mathrm{N}\rangle$

The words are combined with multiset formation, so they are freely ordered. Notice that the category is the same as for the noun, so it can still be combined with the case marker. Because CM combines using multiset concatenation, we get the case-marker followed by the noun and adjective, in either order. An example is given in Fig. 7.

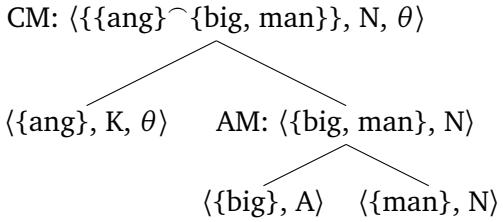

Figure 7: Assembling *ang big man*

3.7 Coordination

$$\textbf{COORD:} \begin{cases} (\langle and, \mathrm{Cj}\rangle, \langle x, \mathrm{C}\rangle, \langle y, \mathrm{C}\rangle) \longmapsto \langle x^\frown and^\frown y, \mathrm{C}\rangle \\ (\langle and, \mathrm{Cj}\rangle, \langle x, \mathrm{C}, X\rangle, \langle y, \mathrm{C}, X\rangle) \longmapsto \langle x^\frown and^\frown y, \mathrm{C}, X\rangle \end{cases}$$
$$\forall C \in Cat, X \in \mathbb{K} \cup \wp(\mathbb{K})$$

Let us see how COORD interacts with AM and CM. Suppose we want to coordinate nouns. COORD predicts that we can do so before or after CM.

(5) a. COORD
$(\langle and, \mathrm{Cj}\rangle, \langle man, \mathrm{N}\rangle, \langle woman, \mathrm{N}\rangle) \longmapsto \langle \{man^\frown and^\frown woman\}, \mathrm{N}\rangle$
CM
$(\langle \{man^\frown and^\frown woman\}, \mathrm{N}\rangle, \langle \{ang\}, \mathrm{K}, \theta\rangle) \longmapsto$
$\langle \{ang\}^\frown \{man^\frown and^\frown woman\}, \mathrm{N}, \theta\rangle$

b. CM
$(\langle \{man\}, \mathrm{N}\rangle, \langle \{ang\}, \mathrm{K}, \theta\rangle) \longmapsto \langle \{\{ang^\frown man\}\}\rangle$
$(\langle \{woman\}, \mathrm{N}\rangle, \langle \{ang\}, \mathrm{K}, \theta\rangle) \longmapsto \langle \{\{ang^\frown woman\}\}\rangle$
COORD
$(\langle and, \mathrm{Cj}\rangle, \langle \{\{ang\}^\frown \{man\}\}\rangle, \langle \{\{ang\}^\frown \{woman\}\}\rangle)$
$\longmapsto \langle \{\{ang\}^\frown \{man\}\}^\frown \{and\}^\frown \{\{ang\}^\frown \{woman\}\}, \mathrm{N}, \theta\rangle$

Tagalog does in fact allow coordination of both case-marked and un-case-marked nouns, so Little Tagalog generates correctly here. More interesting is that Little Tagalog generates correctly for modified nouns.

(6) COORD
$(\langle and, \mathrm{Cj}\rangle, \langle man, \mathrm{N}\rangle, \langle woman, \mathrm{N}\rangle) \longmapsto \langle \{man^\frown and^\frown woman\}, \mathrm{N}\rangle$
AM
$(\langle \{man^\frown and^\frown woman\}, \mathrm{N}\rangle, \langle \{big\}, \mathrm{A}\rangle) \longmapsto \langle \{big, man^\frown and^\frown woman\}, \mathrm{N}\rangle$
CM
$(\langle \{big, man^\frown and^\frown woman\}, \mathrm{N}\rangle, \langle \{ang\}, \mathrm{K}, \theta\rangle$
$\longmapsto \langle \{\{ang\}^\frown \{big, man^\frown and^\frown woman\}\}, \mathrm{N}, \theta\rangle$

The word orders here are as follows (data from Nerissa Black, p.c. 2009).

(7) a. ang big man and woman
b. ang man and woman big

The string in example (7-a) can also mean that the man, but not the woman, is big. (7-b) can also mean that the woman but not the man is big. This would be generated by modifying only one of the nouns instead of the coordination of the nouns:

(8) a. [ang [[big man] and woman]]
b. [ang [man and [woman big]]]

Both can mean that both the man and the woman are big, though to be unambiguous one could modify both with the adjective. Clearly, our grammar generates these too.

(9) a. [ang [[big man] and [big woman]]]
b. [ang [[man big] and [woman big]]]

I don't yet know whether *ang big man and woman big* and *ang man big and big woman* are grammatical in real Tagalog. I predict them to be.

To see the ambiguity of (7-a), consider the trees in Fig. 8 and 9. The tree in Fig. 8 is the tree for the derivation decribed above. The tree in Fig. 9 is a different derivation, wherein only *man* is modified by the adjective. The word-orders for this tree are:

(10) a. ang big man and woman
b. ang man big and woman

Despite the different derivation tree, (10-a) here is identical to (7-a), explaining the ambiguity.

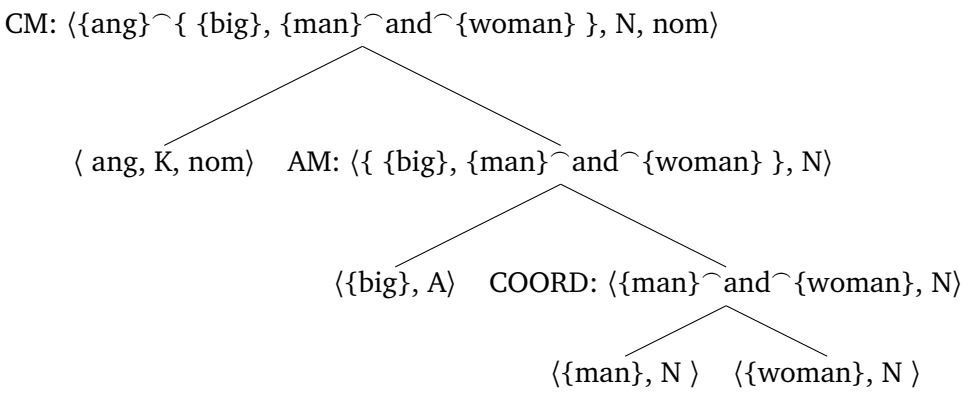

Figure 8: Tree with *man and woman* modified by the adjective

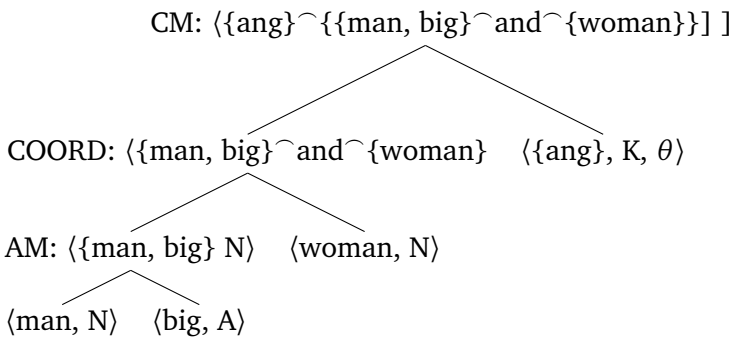

Figure 9: Tree with only *man* modified by the adjective

I also predict correctly that *ang man big and woman* can only mean the man is big. This is because in the case where both the man and woman are big, *man and woman* must form a concatenated constituent by the action of COORD. Then *big* merges with *man and woman* and therefore cannot appear between them in this case.

Conclusion

By adding more elements to the tuples in the lexicon, the grammar can be expressed more succinctly and can capture more thoroughly the relationships between categories. We have seen how a third element can be used to calculate case marking and the fulfillment of theta grids.

Truly free word order can be accounted for if we allow the derivation of concatenated (multi-)sets, rather than just concatenated words. The elements of a multiset are unordered, and arbitrary orders can therefore be defined from a single derivation.

Appendix: Little Tagalog Grammar

$L_T = \langle V, Cat, Rule, Lex, \theta \rangle$

- $Cat := \{\text{P}, \text{N}, \text{A}, \text{Voi}, \text{CJ}\}$

- $\theta := \{a, t, g, r, l, i, b, p\}$
- $\mathbb{K} = \wp\theta$
- $Lex \subset (V^* \times Cat) \cup (V^* \times Cat \times \mathbb{K}) \cup (V* \times Cat \times \wp(\mathbb{K}))$:

$\langle gave, \text{P}, \{a, t, r\}\rangle$	$\langle ang, \text{K}, \text{NOM}\rangle$	$\langle man, \text{N}\rangle$
$\langle cooked, \text{P}, \{a, t\}\rangle$	$\langle sa, \text{K}, \text{GEN}\rangle$	$\langle woman, \text{N}\rangle$
$\langle sneezed, \text{P}, \{a\}\rangle$	$\langle ng, \text{K}, \text{DAT}\rangle$	$\langle book, \text{N}\rangle$
$\langle \text{AV}, \text{Voi}, \{a\}\rangle$	$\langle big, \text{A}\rangle$	$\langle and, \text{CJ}\rangle$
$\langle \text{TV}, \text{Voi}, \{t\}\rangle$		
$\langle \text{IV}, \text{Voi}, \{i\}\rangle$		
$\langle \text{BV}, \text{Voi}, \{b\}\rangle$		
$\langle \text{DV}, \text{Voi}, \text{DAT}\rangle$		

- *Rule*

VM $((x, \text{P}, T), (y, \text{Voi}, S)) \longmapsto (x^\frown y, \text{P}, v_S(T))$

CM $(\langle x, \text{N}\rangle, \langle y, \text{K}, k\rangle) \longmapsto \langle y^\frown x, \text{N}, k\rangle$

AL $(\langle x, \text{N}, k\rangle, \langle y, \text{N}, l\rangle) \longmapsto \langle x \cup y, \text{N}, k \cup l\rangle$ if $(k \cup l)(\text{NOM} \leq 1)$

PA $(\langle x, \text{P}, K\rangle, \langle y, \text{N}, L\rangle) \longmapsto \langle x^\frown y, \text{P}, \emptyset\rangle$ where $K, L \in \wp(\mathbb{K})$.

AM $(\langle x, \text{N}\rangle, \langle y, \text{A}\rangle) \longmapsto \langle x \cup y, \text{N}\rangle$

Coord
$$\begin{cases} (\langle and, \text{Cj}\rangle, \langle x, \text{C}\rangle, \langle y, \text{C}\rangle) \longmapsto \langle x^\frown and^\frown y, \text{C}\rangle \\ (\langle and, \text{Cj}\rangle, \langle x, \text{C}, X\rangle, \langle y, \text{C}, X\rangle) \longmapsto \langle x^\frown and^\frown y, \text{C}, X\rangle \end{cases}$$

Acknowledgements

This paper was written as part of a Fall 2009 course at UCLA taught by Ed Stabler and Ed Keenan. I would also like to thank Raphael Mercado and Nerissa Black for their help as Tagalog consultants and Lisa deMena Travis for her guidance in formulating my initial analysis of Tagalog.

References

Keenan, Edward L., and Edward P. Stabler. 2003. *Bare grammar*. Stanford: CSLI Publications.

Kroeger, Paul. 1993. *Phrase structure and grammatical relations in Tagalog*. Stanford: CSLI Publications.

Affiliation

Meaghan Fowlie
University of California, Los Angeles
mfowlie@ucla.edu