Phonological factors in Tagalog adjective-noun word order variation

Stephanie S Shih  
*Stanford University/UC Berkeley*  
stephsus@stanford.edu

Kie Zuraw  
*UCLA*  
kie@ucla.edu

1 INTRODUCTION TAGALOG ADJECTIVE/NOUN PAIRS

- Both orders are possible and claimed to be semantically interchangeable = “without any apparent difference in meaning” (Schachter and Otanes 1972: 122):

\[(1) \text{magandá-ng babáe} \sim \text{babáe-ng magandá} \quad (\text{Kroeger 1998})\]

- Elsewhere, it’s been shown that word order variation can be conditioned by numerous factors – phonological, syntactic, semantic, usage-based, psycholinguistic, sociolinguistic (e.g., McDonald et al. 1993; Wasow 2002; Wright, Hay & Bent 2005; Benor & Levy 2006; Bresnan et al. 2010; and references therein).

1.1 THIS TALK

- Large-scale corpus study investigating factors affecting adjective/noun order.
- As far as we know, this is the first systematic study of Tagalog adjective/noun order variation.
- In particular, we focus on phonological factors that may affect order, including segmental and prosodic well-formedness conditions.

---

* Acknowledgments to RAs Seth Ronquillo, Arienne Filio, and Kimberly Nachor for data checking; Vera Gribanova, Laura McPherson, and participants in the UCLA Phonology Seminar for comments. Ivan Tam developed the software that created the Tagalog corpus.

Handout available [here](http://stanford.edu/~stephsus/ShihZuraw-LSA2014.pdf)
Results

- In addition to semantic and usage-based factors, phonological conditions influence adjective/noun word order in Tagalog.
- Word order is conditioned in part by the phonological surface form of the linker particle.
  
  - Word order can optimize phonological well-formedness (see also Schlüter 2005; Shih, in prep).
  - Phonological (surface) information is available at the point of word order choices.
  
- Theoretical implications of our results:
  - for understanding the role of phonology in determining word order, and
  - for considering the design of the interface between phonology, morphology and syntax in both formal grammatical models and psycholinguistic models of language production.

2 DATA

2.1 Tagalog Web Text Corpus from (Zuraw 2006)

- Web text from 2004, variety of genres
- 47,144,971 word tokens, 105,464 word types

2.2 Adjective/Noun Pair Extraction

- Nouns and adjectives automatically and manually extracted from part-of-speech tags on the SEAsite online Tagalog-English dictionary (SEAsite 2001).
- Searched corpus for all possible noun-linker-adjective and adjective-linker-noun sequences.
- Automatic exclusion of pairs in the following circumstances:
  - instances wherein the second word potentially had a ng linker
  - instances with punctuation occurring between the two words
  - instances where a noun or adjective must bear a certain affix according to the Seasite Dictionary but does not
  - instances with the following tokens: alám, am, dápat, habang, hanggáng, lamáng, said, sayá (ng), tápos, tódo(ng), úpang

- Adjective/noun pairs hand-checked by three Tagalog-English bilinguals:
  - 1,205 noun/adjective pairs (types) were selected, to include the nouns and adjectives occurring most frequently in the set, as well as the most-frequent pairs.
  - 11 words identified as problematic: adverbs rather than adjectives; ambiguities caused by linker (e.g., noóng could be noó-ng ‘forehead’ + linker, or noóng ‘when-past’).
  - Adjective/noun pairs containing any problematic words were excluded.
- Words’ stress patterns manually obtained from a paper dictionary (English 1986).

1 Am (e.g., versus PM) and said were excluded as English words. The rest are more commonly used as verbs (alam, dapat), prepositions (habang, hanggang, tapos, upang), quantifiers (todo), or enclitics (lamang, saya).
- 149,689 adjective/noun pair tokens, 14,591 types
- 1,708 noun types, 587 adjective types

- Some caveats = potential sources of noise:
  - $X$-linker-$Y$ might not form a constituent: $Y$ may be part of a complex modifier.

  (3) a. **táo-ng armádo** ‘person-LINK armed’
  b. **táo-ng armádo sa kanilang ika-apat na henerasyon mobile na aparato**
     ‘person- LINK armed with their fourth LINK generation mobile LINK device’

- We don’t know which pairs are predicates and which are not (claimed to be important by Schachter and Otanes 1972).
- We don’t know if either of the words is focused or represents new information, etc. (though we hope to add givenness information in future versions of the corpus)

3 METHODOLOGY

- Mixed-effects regression model using **glmer()** from **lme4** R package (Bates et al. 2013; R Core development team):
  - Multivariate analysis controls for numerous predictors at once.
  - Data: unique adjective/noun combinations ($n = 14,591$)
  - Dependent variable: rate of noun-adjective order
    = (Instances in noun-adjective order) / (Total instances in both orders)
  - Predictors were centered; numerical predictors also standardized following (Gelman & Hill 2007; Gelman 2008).
  - Random intercepts: noun, adjective

- Independent variables, culled from previous studies of word order variation, descriptions of Tagalog adjective/noun word order, Tagalog phonology, cross-linguistically common phonological behaviors (see §5).
  - prosodic phonological predictors
  - segmental phonological predictors
  - non-phonological predictors
4 General Results

- Adjective-noun order is overwhelmingly preferred = “default” order?
  - Linker preference: -ng linker preferred to na.
    - (Schachter and Otanes 1972): If one order requires -ng linker and the other requires na linker, order that results in -ng linker is preferred.
    - Especially true for -ng linker adjectives.

\[\begin{align*}
\text{(4) } & \text{ áso-} \text{ng ulól is more frequent than ulól na áso} \\
& \text{dog-LINK mad} \quad \text{mad LINK dog} \quad \text{‘mad dog’}
\end{align*}\]

\[\begin{align*}
\text{b. } & \text{ bágo-} \text{ng titser > titser na bágo} \\
& \text{new-LINK teacher teacher LINK new} \quad \text{‘new teacher’}
\end{align*}\]

- Regression coefficient estimates:
  - Adjective linker = ng
    \[\beta = -0.02065, \ SE = 0.004365, \ t = -4.73\]
  - Noun linker = ng
    \[\beta = 0.00839, \ SE = 0.001621, \ t = 5.18\]

(7) | Phonological predictors tested | Effect? |
--- | --- | --- |
| a. Phonotactic, Obligatory Contour Principle | ![Checkmark] | ![Cross] |
| *[nasal] + [nasal] | ![Checkmark] | ![Cross] |
| *[velar] + [velar] | ![Checkmark] | ![Cross] |
| b. Contextual markedness (e.g., *NC) | ![Checkmark] | |
| c. Long-distance phonotactics (e.g., *nasal onset + nasal onset) | ![Cross] | |
| d. Phonological faithfulness (e.g., avoid replacing [?] with nasal) | ![Checkmark] | |
| e. Hiatus avoidance (e.g., *V+V) and syllable structure optimization | ![Checkmark] | |
f. Phonological and morphological alignment
  ✔
g. Stress lapse avoidance
   ✔
h. Stress clash avoidance
   ✔
i. Length in segments of adjective and noun
   ✔
j. Linker surface form (-ng versus na)
   ✔

(8) Non-phonological predictors tested

<table>
<thead>
<tr>
<th>Effect?</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔</td>
</tr>
<tr>
<td>✔</td>
</tr>
<tr>
<td>✔</td>
</tr>
<tr>
<td>✔</td>
</tr>
</tbody>
</table>

Non-phonological predictors tested

- Adjective prototypicality (*ma*- prefix)
- Quantifier/limiter adjectives
- Total length in segments
- Log frequency: word, pair (calculated from Tagalog Web Text Corpus (Zuraw 2006))

(9) Regression coefficients of reliable and trending predictors (see details in §5)
5 RESULTS: PREDICTORS OF ADJ/NOUN WORD ORDER

5.1 SEGMENTAL PHONOLOGY RESULTS

5.1.1 Obligatory Contour Principle effects

- Avoidance of adjacent similar segments affects genitive construction choice in English: adjacent sibilants are avoided (Menn & MacWhinney 1984; Zwicky 1987; a.o.).

(10) Genitive alternation (Hinrichs & Szmrecsányi 2007; a.o.)

the descendants of the veterans > the veterans + ’s descendants

- Root-externally in Tagalog, nasal-nasal sequences are underrepresented:
  - In a database of 4,617 native, disyllabic roots (from English 1986), of which 1,579 have a medial cluster, we expect 51 nasal-nasal medial clusters, but find only 8.

  (11) tsinélas na itím > itím na tsinélas

slippers LINK black black LINK slippers ‘black slippers’

5.1.2 Contextual markedness: *NC̆

- Cross-linguistically, NC̆ clusters can trigger a number of phonological repairs (e.g., deletion, assimilation, fusion; Pater 1996, 2001).
- The *NC̆ constraint (Pater 1996, 2001; Hayes and Stivers 1996) is active at Tagalog prefix-stem boundaries, as in many related languages (see Zuraw 2010 for more):

(12) A stem-initial voiceless obstruent usually fuses with a preceding nasal, but a voiced obstruent usually does not.

a. /ma-pañ-kanmám/ → [ma-pañ-amkám] ‘rapacious’
b. / pañ-diníg/ → [pan-diníg] ‘sense of hearing’

- Tagalog word order: Avoid linker -ng + voiceless-initial noun.

(β = -0.006461, SE = 0.002127, t = -3.04)

(13) péra-ng nakalaán > nakalaá-ng péra

money-LINK dedicated dedicated-LINK money ‘dedicated money’
5.1.3 Long-distance phonotactics

- Non-local onset similarity avoidance is perhaps a bit more unusual, but repeated similar or identical sequences are often avoided (Stemberger 1981; Menn & MacWhinney 1984; Yip 1998; Frisch et al. 2004; Löfstedt 2010).

(14) Center-embedding of possessive NPs in Ancient Greek blocked when a sequence of adjacent identical articles would be formed (Golston 1995: 353).

\[ \text{[h-ee tólm-a] [t-óon leg-ónt-oon]} \]
\[ \text{the-N:F courage-N:F the-G:M:P speak-ing-G:M:P} \]
\[ \text{‘the courage of those speaking’} \]

(15) Tagalog word order: trend of onset similarity avoidance between \textit{na} linker and following N-initial adjective.

\( \beta = -0.005072, SE = 0.002816, t = -1.8 \)

(b) \[ \text{[h-ee t-óon leg-ónt-oon] tólm-a} \]
\[ \text{the-N:F the-G:M:P courage-N:F speak-ing-G:M:P} \]
\[ \text{‘the courage of those speaking’} \]

(c) \[ \text{[[t-óon oikeí-oon] tin-ás] [t-óon ekeín-oon]} \]
\[ \text{‘some of the slaves of those [people]’} \]

(d) \*\[ \text{[[t-óon ekeí-ooon] oikeí-oon] tin-ás] \]
\[ \text{‘some of the slaves of those [people]’} \]

5.1.4 Syllable structure effects

- Hiatus avoidance and \textbf{ONSET} (Prince & Smolensky 1993:17) affects binomial ordering of name pairs in English (Wright et al. 2005; Benor and Levy 2006; a.o.)

(16) John and Yoko \( \rightarrow \) Yoko and John

(17) Tagalog word order: trend of hiatus (and resulting [ʔ] epenthesis) avoidance in noun-adjective order.

\( \beta = -0.007616, SE = 0.004399, t = -1.73 \)

(b) \text{espesyál na bágay > bágay na espesyál}
\text{special LINK thing thing LINK ordinary ‘special thing’}
5.1.5 Phonological and morphological alignment

- Morpheme boundaries prefer to coincide with syllable boundaries.
  - e.g., (McCarthy & Prince 1993) ALIGN(Stem, Right; Syllable, Right) for Axininca Campa, Lardil, Hebrew, Bedouin Arabic, and Kamaiurá
  - Alignment between syllable boundaries and higher-level boundaries can affect ordering of syllables in these cases as well: e.g., choice between [na.ta] and *[ta.na] in Axininca Camp depends on ALIGN-L.

- Tagalog word order: V-initial words prefer to be initial, avoiding resyllabification or glottal stop insertion.
  
  $\leftarrow$ Avoid V-initial adjectives in second position.
  \[(\beta = -0.036161, SE = 0.010007, t = -3.61)\]

  \[\text{(18)} \quad \text{ordináryo-ng táo} \quad \text{> táo-ng ordináryo} \quad \text{ordinary-LINK person} \quad \text{person-LINK ordinary} \quad \text{‘ordinary person’}\]

  $\leftarrow$ Avoid ng linker + V-initial noun.
  \[(\beta = -0.008055; SE = 0.003014, t = -2.67)\]

  \[\text{(19)} \quad \text{ítlóg na pulá} \quad \text{> pulá-ng ítlóg} \quad \text{egg LINK red} \quad \text{red-LINK egg} \quad \text{‘red [brined] egg’}\]

5.2 PROSODIC PHONOLOGY RESULTS

5.2.1 Rhythm

- Stress clashes are avoided via alternative word order in English (see also Temperley 2009):

  (20) Clash avoidance leads to avoidance of $a$-adjectives in prenominal positions in diachronic change in English (Schlüter 2005).

  a. $\sigma \sigma \sigma \sigma \sigma \sigma \sigma \sigma \sigma$  
  the pérsont who was awáre $\rightarrow$ ??the awáre pérsont

  b. $\sigma \sigma$  
  $\rightarrow$ ??the asléep pérsont
Stress lapses are avoided via alternative word order in English:

(21) Genitive alternation (e.g., Shih et al. to appear; Shih in prep; see also binomial pair ordering: McDonald et al. 1993; Wright et al. 2005; Benor & Levy 2006)

\[
\begin{align*}
\sigma & \sigma \sigma \sigma \sigma \\
\text{the system's benefits} & \quad > \quad \sigma \sigma \sigma \sigma \sigma \\
& \text{the benefits of the system}
\end{align*}
\]

- Background: Tagalog primary stress is contrastive: can be final, penultimate, or in some loans, antepenultimate.
- Stressed non-final vowels are notably longer than unstressed.
- Secondary stress is not fully understood for Tagalog (French 1988; French 1991; a.o.).
  - Usually not included in dictionary; we did not attempt to annotate for it

- Tagalog word order: trends in data towards clash and lapse avoidance in adjective-noun word order. The trends are evidenced in a subset of the data controlling for word length (since rhythmic measures are correlated with word length).

- Tagalog word order: primary stress clashes caused by adjective-noun order are avoided.
  (in data subset with bi- and tri-syllabic words only: \( \beta = 0.002865, SE = 0.002732, t = 1.049 \))

\[
\begin{align*}
\sigma \sigma \sigma \sigma & \quad > \quad \sigma \sigma \sigma \sigma \\
\text{báya-ng sawí}^p & \quad > \quad \text{sawí-ng báyan} \quad \text{‘unfortunate country-country/people’}
\end{align*}
\]

- Tagalog word order: long primary stress lapses (3 or more unstressed syllables) are avoided in adjective-noun order.
  (in data subset with bi- and tri-syllabic words only: \( \beta = -0.004239, SE = 0.002431, t = -1.744 \))

\[
\begin{align*}
\sigma \sigma \sigma \sigma \sigma & \quad > \quad \sigma \sigma \sigma \sigma \sigma \\
espesyál na áraw & \quad > \quad \text{áraw na espesyál} \quad \text{‘special day’}
\end{align*}
\]

5.2.2 Weight/length

- Heavier/longer constituents tend to come at the peripheries (Behagel 1909; Quirk et al. 1985; Hawkins 1994; Wasow 2002; a.o.).
  - In English, heavier/longer constituents come last. In V-final languages like Japanese, heavier-longer constituents tend to come first (e.g., Hawkins 1994: 426).
  - Numerous proposals for the underlying cause of heavy-to-edge phenomena include processing-based (e.g., Gibson 2000; Temperley 2006), syntactic (e.g., Wasow 2002), and prosodic explanations (Zec and Inkelas 1990; Zubizarreta 1998; Anttila et al. 2010; a.o.) (see Szmrecsányi 2004; Grafmiller & Shih 2011 for empirical comparisons of these proposals).
“Heavy-last” in genitive syntactic variation in English (e.g., Rosenbach 2005; Hinrichs and Szmrecsányi 2007; Grafmiller and Shih 2011; Shih et al., to appear):

a.  \([\text{the attitude}]_{\text{NP}}\) of \([\text{people who are really into classical music and feel that if it’s not seventy-five years old, it hasn’t stood the test of time}]_{\text{NP}}\)  
>  

b.  \([\text{people who are really into classical music and feel that if it’s not seventy-five years old, it hasn’t stood the test of time}]_{\text{NP}}\)’s \([\text{attitude}]_{\text{NP}}\)

- Tagalog word order is usually VSO and, like English, heavy-last is expected (cf. Japanese).
  - e.g., Relative clause order is variable, but Schachter and Otanes (1972: 123) note a “tendency to prefer the order head-linker-modifier when the modifying phrase is long.”

\[(25)\]
\[
\text{ang pagkái-ng nilúto mo} > \text{ang nilúto mo-ng pagkái} \quad (\text{S&O 1972: 123})
\]
\[
\text{DET food-LINK cooked you} \quad \text{DET cooked you-LINK food} \quad \text{‘the food you cooked’}
\]

- Tagalog word order: Results here, however, demonstrate a different effect: long words (with more segments) tend to prefer canonical, “default”, adjective-noun order\(^2\). (Also tested syllable count with the same result.)

\[\leftarrow\]  
Longer nouns prefer default second position.  
\[
(\beta = -0.001574; \ SE = 0.000445, \ t = -3.54)
\]

\[\leftarrow\]  
Longer adjectives prefer default first position.  
\[
(\beta = -0.005807; \ SE = 0.00179, \ t = -3.23)
\]

\[\leftarrow\]  
Interaction for long nouns and long adjectives, even more “default” effect: longer nouns and adjectives prefer adjective-noun order.  
\[
(\beta = -0.000381, \ SE = 0.000209, \ t = -1.82)
\]

\[\leftarrow\]  
Possible explanation (for future inquiry):
  - In order to ease processing of heavy constituents, default order is used?

---

\(^2\) cf. Donohue (2007:360, fn 11), who observes that “relative length” affects order.
5.3 Non-phonological results

5.3.1 Prototypical adjectives

- Prototypical items are more easily accessed, and tend to come first in word order (Kelly et al. 1986).
- In Tagalog, adjectives are often formed from *ma-* prefix (e.g., *ma-baho* ‘malodorous’ from *baho* ‘bad smell’).

[Tagalog word order: *Ma-* initial adjectives—i.e., “prototypical” adjectives—prefer the predominant adjective-noun order.]

\[
(\beta = -0.012272, SE = 0.005007, t = -2.45)
\]

\[(29) \text{mabáho-ng utót} > \text{utót na mabáho}\]

5.3.2 Limiters

- Schacter and Otanes (1972:121): “Limiter adjectives” (e.g., cardinal numbers, ordinal numbers, quantifiers) prefer first position.

\[
(\beta = -0.039683, SE = 0.018935, t = -2.1)
\]

\[(30) \text{lahát na táo} > \text{táo-ng láhát}\]

all LINK person person- LINK all ‘all people’

5.3.3 Frequency

- More frequent nouns prefer initial position. \((\beta = 0.006111, SE = 0.000772, t = 7.92)\).
- Frequency effect for adjectives not reliable (likely because they already prefer first position; also possible ceiling effect of frequency)\(^3\).
- More frequent adjective/noun pairs prefer canonical, adjective-noun order\(^4\).

\[(\beta = 0.051286, SE = 0.01049, t = 4.89)\]

---

\(^3\) We also found a puzzling interaction between adjective frequency and pair: more frequent adjectives that appear in more frequent adjective/noun pairs (regardless of noun frequency) have a tendency to occur in noun-adjective order \((\beta = 0.007679, SE = 0.003173, t = 2.42)\)

\(^4\) Pair frequency (how often a given adjective and noun combination occurs) was residualized on both log(adjective frequency) and log(noun frequency).
6 DISCUSSION

- Our study results show that, controlling for non-phonological predictors, phonological conditions influence adjective/noun word order.
  - Word order alternations can act as repairs to satisfy phonological conditions (see also Schütter 2005; Shih, in prep; a.o.).
- Similar effects of phonology influencing higher-level operations such as word order and construction choices have been demonstrated in other languages.
  - However, large-scale quantitative studies of these types of effects have mostly been limited to English.

6.1 PHONOLOGICAL EFFECTS ON WORD ORDER IDENTIFIED

- What type of phonological conditions can affect word order?
- Hypothesis (see also e.g., Shih, in prep): Phonological conditions that tend to affect word order are ones that hold over syntagmatic configurations—i.e., likely to trigger repairs when elements (e.g., words and phrases) (linearly) combine.
  - Previously identified phonological effects on word order
    - Phonotactic, OCP
    - Syllable structure optimization (e.g., hiatus avoidance)
    - Prosody: rhythm, weight
  - Newly identified phonological effects on word order
    - Contextual markedness (e.g., *NC̥)
    - Long-distance phonotactics
    - Phonological and morphological alignment

  (Syntagmatic) phonological and morphological conditions, which are more familiarly satisfied via phonological optimization, can also be satisfied via word order.

6.2 DISCUSSION: THEORETICAL IMPLICATIONS

- Syntax → Phonology generally have a feed-forward relationship in standard formal models of grammar (Zwicky & Pullum 1986) and psycholinguistic models of language production (Levelt 1989; Ferreira & Slevc 2007; cf. Vigliocco & Hartsuiker 2002)
- A more concrete example: in Distributed Morphology (DM), phonological exponence is not available to syntactic operations (late-insertion theory).
- Current question in DM: What is the ordering of linearization and vocabulary insertion, which provides phonological exponents? (see e.g., Embick 2007).
  - linearization prior to phonological information? (Embick & Noyer 2001; Pak 2008)
  - prosodic structure-building interleaved with linearization? (e.g., Tucker 2011)
  - more cycles through PF? (e.g., Embick 2007).
Even if prosodic structure is allowed to interact with linearization, the general assumption is that segmental phonological information in the phonological form is largely unavailable and irrelevant for linearization purposes.

- **Open question:** How much phonological information is available at the point of linearization? How much phonological encoding occurs before/during grammatical encoding?

- The current study: looks at what happens when there are additional morphophonological processes that compound the problem of available phonological information during linearization.

- Phonologically-conditioned surface form of the linker particle must be available at the time of choice:
  - -ng particle preferred, and
  - segmental constraints (e.g., hiatus avoidance, OCP effects) make reference to the surface form.

- Suggests that effects of word order are in part dependent on knowledge and availability of surface segmental phonological information (not just prosodic structure).

- A possible alternative counter-hypothesis would be that the morphosyntax generates both options equally (i.e., both are viable grammatical alternatives) and then phonology filters.

- **However,** both orders do not appear to be equal options:
  - Adjective-first order is overwhelmingly preferred, even with controls for word, frequency, etc.
  - Phonological structure of adjective-noun order is more closely regulated: a greater number of robust constraints in the model penalize poor phonological structure in adjective-noun order than in noun-adjective order.
7 Conclusion

- Presented quantitative study of phonological and (some) non-phonological conditions on adjective/noun word order variation in Tagalog.

**Results**

- Though variable, adjective/noun pairs exhibit a canonical order preference for adjectives followed by nouns.
- Phonological conditions—both segmental and prosodic—influence adjective/noun word order preference.
- The phonologically-determined surface allomorph of the linker particle affects adjective/noun word order choice.

- Word order variation can be used to optimize phonological well-formedness, suggesting that phonological information factors into considerations of linearization, word order, and/or grammatical encoding.

**Next directions**

- Variable versus frozen adjective/noun pairs
  - Mollin (2012): Obeying well-formedness preferences that condition variable word order can cause word orders to become frozen → frozen word orders obey well-formedness preferences to a greater extent than their variable counterparts.
  - Do adjective/noun pairs in Tagalog that exhibit less reversibility demonstrate more adherence to the conditions on word order presented here?
  - May shed light on how phonological effects on word order may become grammaticalized over time.

- Variation with linker particle:
  - Variation in linker realization: bare adjective/noun pairs without linker are rare but do occur.
  - Variation in linker surface form: e.g., Richards (1999: 307) reports that some speakers accept linker *na* when *-ng* is called for phonologically, but this only occurs in noun-adjective order.
  - Open questions: What conditions such linker variation? and how does it interact with word order variation?

8 References


Hayes, Bruce & Tanya Stivers. 1996. The phonetics of post-nasal voicing.


http://www.seasite.niu.edu/Tagalog/on_site_dictionary.htm.


