

Sub-extraction in Tagalog: An experimental investigation

Overview. We investigated sub-extraction in Tagalog and the factors that can affect its acceptability. We considered the alienability of the possession relation in experiment 1 and the case marking of the DP out of which sub-extraction occurs in experiment 2. **At the time of writing, we are still collecting data (5% complete). We intend to have data collection be done by early to mid February and data analysis by late February to early March.**

Background. It is well-known that A'-extraction processes such as topicalization or relativization in Tagalog are subject to many restrictions. Research in this area tends to focus on the extraction behavior of clausal arguments. However, scholars have also observed that A'-extraction may target noun phrases (e.g., possessors) contained **within** these clausal arguments (i.e., sub-extraction) [1-3]. To our knowledge, previous works all claim that sub-extraction must happen out of the nominative- or *ang*-marked pivot argument of a clause, and some point out that factors such as alienability affect acceptability. However, such claims have often been demonstrated with only a handful of examples, so it is difficult to ascertain how strong the effect of these factors is. Anecdotally, there also appears to be inter-speaker variation regarding how acceptable sub-extraction is in the first place, making traditional elicitation ineffective for addressing these issues. We therefore conducted two experiments to elucidate when sub-extraction is acceptable and when it is not.

Exp1: Alienability of the possession relation. Sub-extraction seems to be most productive when the possession relation is inalienable [1]. To examine the extent to which alienability affected the acceptability of possessor extraction, we constructed 24 items by crossing whether the possessor relation was **alienable** or not (ALIEN: AL, INAL) and whether the possessor was **sub-extracted** or not (SUBEX: +, -). A sample item is provided below. Participants were asked to rate the acceptability of the sentence using a 7-point scale.

ALIEN	SUBEX	Sample item					
INAL	-	<i>Ang</i> NOM	<i>sungay</i> horn	<i>ng</i> GEN	<i>kalabaw</i> water buffalo	<i>ay</i> AY	<i>nabali</i> broke
INAL	+	<i>Ang</i> NOM	<i>kalabaw</i> water buffalo	<i>ay</i> AY	<i>nabali</i> broke	<i>ang</i> NOM	<i>sungay</i> horn
~ The horn of the water buffalo broke							
AL	-	<i>Ang</i> NOM	<i>araro</i> plow	<i>ng</i> GEN	<i>kalabaw</i> water buffalo	<i>ay</i> AY	<i>nabali</i> broke
AL	+	<i>Ang</i> NOM	<i>kalabaw</i> water buffalo	<i>ay</i> AY	<i>nabali</i> broke	<i>ang</i> NOM	<i>sungay</i> horn
~ The plow of the water buffalo broke							

In lieu of discussing our results thus far, we talk about our predictions. Sentences that do not involve sub-extraction (i.e. -SUBEX) will serve as the baseline for how participants rate sentences with extraction of the entire DP. If alienability affects the acceptability of sub-extraction, then the difference between -SUBEX and +SUBEX in the INAL-condition should be smaller than the difference in the AL-condition. In other words, the interaction will be statistically significant. On the other hand, if alienability has little to no effect on sub-extraction, then we expect little to no difference between -SUBEX and +SUBEX in the INAL- and AL-conditions. In other words, there will be a non-significant interaction and a main effect of SUBEX, such that +SUBEX will be rated lower than -SUBEX.

Exp2: Case marking of the possessum. It is claimed that extraction of possessors must occur out of *ang*-marked DPs [1-3]. To examine the extent to which this holds, we asked participants to rate how good a

sentence that did not involve sub-extraction (S2) is as a paraphrase of a sentence that involved sub-extraction (S1) using a 7-point scale. We constructed 12 items by manipulating where the fronted possessor of S2 is attached to in S1 (i.e., whether it is attached to DP marked with *ang* or *ng*). A sample item is provided below.

S1	<i>Si</i> NOM	<i>Juan</i> PN	<i>sinampal</i> slapped	<i>ng</i> GEN	<i>nars</i> nurse	<i>yung</i> NOM	<i>doktor</i> doctor
~ The nurse slapped the doctor of Juan (per the observations of [1-3])							
S2 (<i>ng</i>)	<i>Sinampal</i> slapped	<i>ng</i> GEN	<i>nars</i> nurse	<i>ni</i> GEN	<i>Juan</i> PN	<i>yung</i> NOM	<i>doktor</i> doctor
~ The nurse of Juan slapped the doctor							
S2 (<i>ang</i>)	<i>Sinampal</i> slapped	<i>ng</i> GEN	<i>nars</i> nurse	<i>yung</i> NOM	<i>doktor</i> doctor	<i>ni</i> GEN	<i>Juan</i> PN
~ The nurse slapped the doctor of Juan							

If the case marking of the DP out of which sub-extraction occurs affects acceptability, then we expect the paraphrase of S1 where the possessor is attached to the *ng*-DP to be rated lower than when it is attached to the *ang*-DP. On the other hand, if it does not affect acceptability, then we expect very little to no difference between sentences where the fronted possessor in S1 was attached to the *ng*-DP and to the *ang*-DP in S2.

Details of our statistical analyses. All of the statistical analyses will be implemented in R [4] using the package `brms` [5] for Bayesian estimation. In experiment 1, we will fit an ordinal mixed effects regression model with a probit link function. In experiment 2, we will fit a logistic mixed effects regression model. We will use uninformative priors that do not place strong constraints on the model’s predictions, and incorporate very little knowledge about what makes a plausible response distribution. We will use LKJ(2) as the prior for our correlation matrix. For each model, we will run four Monte Carlo Markov chains in parallel, with 40,000 samples each. The first 8,000 will always be discarded as part of warm-up. To quantify how much evidence we have for an effect, given our data, we will be conducting Bayes factor analyses. Because Bayes factor analyses are highly sensitive to prior specifications [6], we will also conduct a sensitivity analysis.

Theoretical implications. Despite the somewhat uncertain status of sub-extraction in Tagalog, several works over the years have made claims about the syntax of A’-extraction (both within Tagalog and more generally) based on this phenomenon. Ascertaining its status is thus important for determining the validity of existing theoretical claims. For example, a weak effect of Case marking may suggest that the mechanisms responsible for regular extraction and sub-extraction should not be as unified as in previous proposals [7–8]. Furthermore, a strong effect of alienability suggests the possibility that sub-extraction be derived through a mechanism entirely separate from A’-extraction. On the other hand, our data will allow us to better understand the nature of inter-speaker variation and variability in acceptability. For example, we may find that sub-extraction is indeed categorical for certain speakers, in line with what has been previously assumed.

Selected References: [1] Kroeger, P. (1993). *Phrase structure and grammatical relations in Tagalog*. [2] Ceña (1979). Tagalog counterexamples to the Accessibility Hierarchy. *Studies in Philippine Linguistics*. [3] Nakamura (1996). *Economy of chain formation*. [4] R Core Team (2024). *R: A language and environment for statistical computing*. [5] Bürkner, P-C (2017). `brms`: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*. [6] Nicenboim, B., Vasishth, S., & Rösler, F. (2020). Are words pre-activated probabilistically during sentence comprehension? evidence from new data and a Bayesian random-effects meta-analysis using publicly available data *Neuropsychologia*. [7] Branam, K. (2018). Attraction at a Distance: A’-movement and Case. *Linguistic Inquiry*. [8] Hsieh, H. (2020). *Beyond nominative: A broader view of A’-dependencies in Tagalog*.