Pizarro-Guevara, Jed Sam & Wagers, Matthew. 2024. A tale of two Tagalogs. *Glossa: a journal of general linguistics* 9(1). pp. 1–53. DOI: https://doi.org/10.16995/glossa.11032

Open Library of Humanities

A tale of two Tagalogs

Jed Sam Pizarro-Guevara, University of Massachusetts, Amherst, US, jpguevara@umass.edu

Matthew Wagers, University of California, Santa Cruz, US, mwagers@ucsc.edu

A well-received generalization in Tagalog is that only the argument that is cross-referenced by voice is eligible for A-bar extraction. However, recent work has shown that agents that are not cross-referenced by voice are also eligible. We provide naturally occurring data, along with experimental evidence, consistent with this more permissive picture. Further, we present computational evidence that participants were treating agent-extractions not cross-referenced by voice categorically, that is, they were either accepting or rejecting them in any given trial. Thus, we identify a piece of grammatical knowledge (i.e., extraction) that is systematic within an individual speaker but varies unpredictably across a population of Tagalog speakers. In other words, our data reveal two separable types of Tagalog speakers vis-à-vis extraction. We propose that this is a form of grammar competition that arises via the idea that the agent-first bias affects how child learners parse input strings under noisy conditions during acquisition.

Glossa: a journal of general linguistics is a peer-reviewed open access journal published by the Open Library of Humanities. © 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

OPEN ACCESS

1 Introduction

Austronesian-type voice has been central to many investigations because it interacts with case marking, verbal morphology, and certain syntactic operations, like relativization and *wh*-question formation, to name a few. It is characterized by the following: (i) verbal morphology that co-varies with a privileged argument; (ii) the privileged argument receives a particular morphological form, irrespective of its thematic role; and (iii) the eligibility to participate in said syntactic operations is often limited to this privileged argument (Foley 2008; Erlewine et al. 2017; Chen & McDonnell 2019). Throughout, we refer to these syntactic operations as A-bar extraction.

The present study looks at Tagalog, an Austronesian language spoken in the Philippines by approximately 22.5 million native speakers (Philippine Statistics Authority 2021).[1] Tagalog voice morphology and its interactions with the other parts of the language's grammatical system are well-studied in the Austronesian literature. At first blush, the language exhibits all of the hallmark properties of the Austronesian voice system that Foley, Erlewine et al., and Chen & McDonnell discuss. Here we re-examine the nature of the interaction between the privileged argument and A-bar extraction in the language. In short, we find that the eligibility to undergo A-bar extraction is *not* limited to only the privileged argument. For some speakers, even when agents are not cross-referenced by voice morphology, they are still extractable.

In this section, we provide an overview of the Tagalog voice system, and the interaction between A-bar extraction and the argument singled out by voice. We situate our investigation in the context of the received view of this interaction. Finally, we describe the questions of our experiments, and preview our results and contributions.

1.1 Voice cross-references an argument and interacts with A-bar extraction

Verbs in Tagalog typically carry what has been called VOICE MORPHOLOGY, which always cross-references the argument of the clause marked by *ang* [aŋ] (Schachter & Otanes 1983). For example, the sentences in (1) involve a man, a fish, and a store in a buying event, and the thematic relations assigned by the verb to its arguments remain constant throughout. Notice that the verb form changes and the case markers of the arguments vary depending on the verb form. In (1a), the verb exhibits AGENT VOICE (AV). The agent *lalaki* 'man' is marked nominative; the

[1] This estimate increases dramatically if speakers of other Philippine languages are included since most of them are early sequential bilinguals. These speakers learn the language of the region/province where they are from, and upon entry into the educational system, they learn Tagalog and English (Galang 1988; 2001). Even though the Constitution of 1987 declares Filipino as the national language, the Tagalog/Filipino distinction is a non-issue for us. For present purposes, we treat them as one and the same: Filipino is the standardized form of Tagalog that borrows from other languages, like Cebuano and English, for example.

patient *isda* 'fish', genitive, with *ng* [naŋ]; and the location *tindahan* 'store', dative, with *sa*.[2] In (1b), the verb exhibits PATIENT VOICE (PV). The patient is now marked nominative; the agent, now genitive; and the location, still dative. In (1c), the verb exhibits LOCATIVE VOICE (LV).[3] The location 'store' is now marked nominative, and the other arguments, genitive. Throughout, the argument cross-referenced by the verb is in bold; and the agent, underlined.

(1) Adapted from Foley & Van Valin, Jr. (1984: 135)

a. Bumili ang **<u>lalaki</u>** ng isda sa tindahan
bought.AV NOM man GEN fish DAT store
'The man bought fish at the store.' *Agent voice (AV)*

b. Binili ng <u>lalaki</u> ang **isda** sa tindahan
bought.PV GEN man NOM fish DAT store
'The man bought the fish at the store.' *Patient voice (PV)*

c. Binilhan ng <u>lalaki</u> ng isda ang **tindahan**
bought.APPL GEN man GEN fish NOM store
'The man bought fish at the store.' *Locative voice (LV)*

As in many other Austronesian languages, the argument cross-referenced by voice interacts with A-bar extraction in Tagalog. This interaction has been called many names in the syntactic literature, from "subject-only restriction" (Keenan & Comrie 1977) to "absolutive restriction on A-bar extraction" (Aldridge 2008). We refer to this interaction throughout as the "extraction restriction." In (2) is the received generalization that emerges upon surveying the Tagalog syntax literature. We refer to the received view as the STRICT version of the extraction restriction—to contrast with what we will be referring to as the LESS STRICT version presented in section 1.2.

(2) The *strict* extraction restriction in Tagalog
Only the *ang*-marked argument (*i.e.,* the argument cross-referenced by voice) is eligible for A-bar extraction (Schachter 1977; Ceña 1979; Aldridge 2002; Rackowski & Richards 2005; Kaufman 2009; Law 2016; Aldridge 2017; 2018).

What (2) says is that an argument that is extracted must match the argument that is cross-referenced by voice (i.e., the argument that is marked with *ang*). In (3), we show this interaction. When the verb exhibits AV, the agent *babae* 'woman,' which is cross-referenced by AV, is

[2] As one of the reviewers pointed out, the way in which we glossed the case markers *ang* and *ng* (i.e., as nominative and genitive, respectively) presume a certain view of the voice/case system. Alternatively, these can also be glossed as "pivot" and "non-pivot", respectively. Nothing crucial hinges on our choice of labels. We remain agnostic about the debate in generative syntax/syntactic typology concerning the morphological alignment of Philippine-type languages. Refer to Latrouite (2011) for a summary of the different ways scholars have viewed the interaction between voice and case marking.

[3] Following Rackowski (2002) and Aldridge (2004), we assume that the other voices in Tagalog are different flavors of high applicatives. See Chen (2017) for arguments against an applicative analysis.

extractable and can serve as the pivot of a cleft, as in (3a). Meanwhile, the patient *baro* 'dress,' which is not cross-referenced by AV, is not extractable and cannot serve as the pivot of a cleft, as in (3b). In contrast, when the verb exhibits PV, the patient 'dress,' which is cross-referenced by PV, is extractable and can serve as the pivot of a cleft, as in (3c). The agent 'woman,' which is not cross-referenced by PV, is not extractable and cannot serve as the pivot of a cleft, as in (3d).

(3) Adapted from Schachter (1977: 286)

a. Iyon ang **babae**-ng bumili ng baro
that NOM woman-LNK bought.AV GEN dress
'That's the woman who bought a dress.' *AV, Agent-extraction*

b. *Iyon ang baro-ng bumili ang **babae**
that NOM dress-LNK bought.AV NOM woman
Intended: That's the dress that the woman bought. *AV, Patient-extraction*

c. Iyon ang **baro**-ng binili ng babae
that NOM dress-LNK bought.PV GEN woman
'That's the dress that a/the woman bought.' *PV, Patient-extraction*

d. *Iyon ang babae-ng binili ang **baro**
that NOM woman-LNK bought.PV GEN dress
Intended: That's the woman who bought the dress. *PV, Agent-extraction*

Throughout, we use the term VOICE-MATCH to describe extractions where the extracted argument is cross-referenced by voice, as in sentences like (3a) and (3c). We use the term VOICE-MISMATCH to describe extractions where the extracted argument is *not* cross-referenced by voice, as in sentences like (3b) and (3d). In (4) we reframe the strict version of the extraction restriction in terms of voice-match.

(4) The *strict* extraction restriction in Tagalog (reframed)
Only voice-match extractions are allowed.

For ease of exposition, we exemplified the voice-match restriction using AV and PV. The restriction is reported to extend to the other voices (Rackowski & Richards 2005).

1.2 The puzzle: Voice-mismatch is allowed if the agent is extracted

Contra the generalization in (2)/(4), others have observed that the interaction is not as clear-cut. For some speakers, voice-mismatch is allowed if the extracted argument is the agent (Ceña & Nolasco 2011; 2012; Tanaka 2016; Hsieh 2019; Bondoc 2021).[4] In (5), we provide the less strict version. Note that (4) is a proper subset of (5).

[4] Even though Hsieh (2019) treats agent-extractions under PV as ungrammatical, he notes in footnote 10 that some speakers do not judge agent-extractions under PV to be ungrammatical. In other words, sentences like (3d) do not have the same status as sentences like (3b), which are uncontroversially ungrammatical.

(5) The *less strict* extraction restriction in Tagalog
Voice-match extractions are allowed; voice-mismatches are allowed only with agent-extractions.

Ceña & Nolasco (2011: 181) observed the following contrasts in Tagalog. In (6), the verb exhibits AV. The agent *bantay*, the argument cross-referenced by voice, can be relativized, as in (6a). The patient *basi* cannot be relativized, as in (6b). The source *tao* also cannot be relativized, as in (6c). These patterns are consistent with (2)/(4) thus far.

(6) When the verb exhibits AV, only the agent can be relativized
a. **bantay** na bumili ng basi sa tao
guard LNK bought.AV GEN rice wine DAT person
'guard that bought rice wine from the person'
b. *basi na bumili ang **bantay** sa tao
rice wine LNK bought.AV NOM guard DAT person
Intended: rice wine that the guard bought from the person
c. *tao na bumili ng basi ang **bantay**
person LNK bought.AV GEN rice wine NOM guard
Intended: person that the guard bought rice wine from

The data reported by Ceña & Nolasco diverge from the generalization in (2)/(4) when the verb exhibits non-AV voices. In (7), the verb exhibits PV. The patient *tao*, the argument cross-referenced by voice, can be relativized, as in (7a). The location *palengke* cannot be relativized, as in (7b)—at least, not using the same relativization strategy (Hsieh 2018). These patterns are consistent with (2)/(4) thus far. However, (7c) is inconsistent with it: the agent *bantay* can also be relativized.

(7) When the verb exhibits PV, the patient and the agent can be relativized
a. **tao** na niloko ng bantay sa palengke
person LNK deceive.PV GEN guard DAT market
'person that the guard deceived at the market'
b. *palengke na niloko ng bantay ang **tao**
market LNK deceive.PV GEN guard NOM person
Intended: market where the guard deceived the person
c. bantay na niloko ang **tao** sa palengke
guard LNK deceive.PV NOM person DAT market
'guard that deceived the person at the market'

In (8), when the verb exhibits applicative morphology, the recipient *tao*, the argument cross-referenced by voice, can be relativized, as in (8a). The patient *pitaka* cannot be relativized, as in (8b). These patterns are consistent with (2)/(4) thus far. However, (8c) is inconsistent with it: the agent *bantay* can also be relativized.

(8) When the verb exhibits applicative morphology, the applied argument and the agent can be relativized

a. **tao** na dinukutan ng bantay ng pitaka
person LNK steal.APPL GEN guard GEN wallet
'person from whom the guard stole the wallet'

b. *pitaka na dinukutan ng bantay ang **tao**
wallet LNK steal.APPL GEN guard NOM person
Intended: wallet that the guard stole from the person

c. bantay na dinukutan ng pitaka ang **tao**
guard LNK steal.APPL GEN wallet NOM person
'guard that stole the wallet from the person'

Table 1 summarizes the empirical picture by the thematic role of the extracted argument and the version of the extraction restriction. A '✓' means that the extraction is licit; a '(✓)', licit for *some* speakers; and '×', not licit.

Thematic role of extracted argument	**Strict version**			**Less strict version**		
	AV	**PV**	**APPL**	**AV**	**PV**	**APPL**
Agent	✓	×	×	✓	(✓)	(✓)
Patient	×	✓	×	×	✓	×
Other	×	×	✓	×	×	✓

Table 1: An overview of the empirical picture by the thematic role of the extracted argument and the version of the extraction restriction.

Before proceeding, we make fully available Ceña & Nolasco's description of the extraction restriction. Ceña & Nolasco also observed that those who do allow the agent to be relativized in voice-mismatch contexts only do so in head-initial relative clauses. This is shown in (7c), repeated below as (9a). Voice-mismatches are never allowed in head-final relative clauses, as shown in (9b).

(9) In head-initial RCs, the agent may be relativized in voice-mismatch contexts; in head-final RCs, the agent cannot be relativized in said contexts

a. ang bantay na niloko ang **tao** sa palengke
NOM guard LNK deceive.PV NOM person DAT market
'the guard that deceived the person at the market'

b. *ang niloko ang **tao** sa palengke na bantay
NOM deceive.PV NOM person DAT market LNK guard
Intended: the guard that deceived the person at the market

In the judgment studies reported in sections 4.2 and 4.3 (and in our own work with speakers), we replicate the contrasts between agent-extraction in head-initial and head-final RCs, as shown in (9). We leave it to future work to see why these RCs behave differently. One possible reason for their difference in behavior vis-à-vis extraction could be due to structural differences. However, it is an open question as to whether these RCs have different derivations (Aldridge 2017) or share a single one (Law 2016).

1.3 The current study

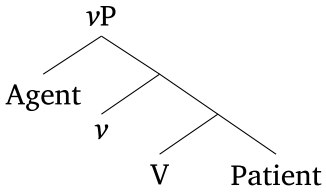
We re-examine the nature of the interaction between voice and A-bar extraction in Tagalog. We have outlined the two competing generalizations of the restriction above, and we repeat these in (10). We framed this interaction in terms of voice-(mis)match to preview the terminologies that we will be using in the experiments that follow.

(10) The two competing generalizations of the extraction restriction
   a. Strict: Only voice-match extractions are allowed
   b. Less strict: Voice-match extractions are allowed; voice-mismatches are allowed only with agent-extractions

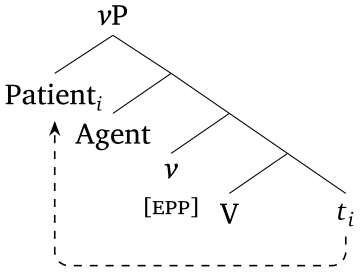
The present study asks two questions. First, we ask whether voice-mismatch extractions are indeed allowed—in particular, those with agent-extractions. Under the strict version of the restriction, only voice-match extractions are allowed; voice-mismatch extractions are not allowed. Thus, the MISMATCH PENALTY, the contrast between voice-match and voice-mismatch extractions should not vary by voice. On the other hand, under the less strict version, the mismatch penalty incurred in AV should be greater than that in non-AV voices because of the acceptability of agent-extractions even in voice-mismatch contexts. Second, if they are allowed when the extracted argument is the agent, we ask whether this permissiveness is a general property shared by all A-bar dependencies or whether this is specific to relative clauses. This is an empirical question because the studies in section 1.2 that observe the less strict version mostly only involved relative clauses.

The present study makes two main contributions. Our first contribution is empirico-theoretical. We show that voice-mismatch agent-extractions do exist “in the wild” and are indeed accepted by some speakers in an experimental setting. These types of extractions are important to formal syntacticians because they challenge current theories of extraction that rely on the patient/applied argument moving to a structural position higher than the agent to determine voice, as in (11) and in (12), and Attract Closest to derive the extraction restriction (Aldridge 2002; Rackowski & Richards 2005; Aldridge 2012). These proposals undergenerate; they cannot generate agent-extractions under non-AV voices for those who do allow them.

(11) Partial derivation of AV



(12) Partial derivation of PV



These types of extractions are important to (developmental) psycholinguists because they show how two types of Tagalog speakers can emerge despite being exposed to the same or similar linguistic input. Ultimately, we argue that what underlies this variation is grammar competition à la Han et al. (2007) that arises from the agent-first bias (Ferreira 2003; Bornkessel & Schlesewsky 2006) interacting with reanalysis (Frazier & Rayner 1982; 1987), and a noisy channel assumption in acquisition (Perkins et al. 2022).

Our second contribution is methodological. We applied the type of computational modeling that Dillon et al. (2017) employed to model English speakers' behavior in agreement attraction, a completely unrelated phenomenon in an unrelated language. We constructed quantitative models that demonstrate that there are two separable types of Tagalog speakers with respect to extraction. This serves as further proof of concept for how a distributional analysis of judgment ratings can provide evidence for multiple types of speakers in a language community. Quantitative models that estimate participants' judgment processes are particularly useful to researchers who argue for the existence of two varieties to account for variation in behavior with respect to a particular linguistic phenomenon. For example, this type of modeling could also be applied to the scopal variation in Korean with respect to negation and quantified objects. Even though Han et al. (2007) and Zeijlstra (2024) burden different parts of the Korean grammar, both proposals account for scopal variation by positing two varieties. The methodology employed in the present study can provide evidence for or against the tenability of having two varieties of Korean, and it can also help adjudicate between the two proposals if we model the speakers' behavior in corners of the grammar where the predictions of the two proposals diverge.

The remainder of the paper is structured as follows. Section 2 presents naturally occurring examples of agent-extractions under non-AV voices and reviews previous experimental findings by Pizarro-Guevara & Wagers (2020). Both are inconsistent with a strict view of the restriction and is consistent with a less strict view. Section 3 reanalyzes experiment 1 of Pizarro-Guevara & Wagers using a Bayesian approach to statistical inference in order to incorporate their insights into the statistical analyses of the judgment studies presented after. Section 4 presents a series of acceptability judgment experiments, which were designed to be a replication and an extension of Pizarro-Guevara & Wagers, and updates the empirical landscape of the extraction restriction. Section 5 presents computational evidence that there are two separable types of Tagalog speakers vis-à-vis extraction. Section 6 concludes by proposing a way in which we can think of how the two types of Tagalog speakers emerge despite being exposed to the same or similar linguistic input. We view these two types of speakers as a form of grammar competition (Han et al. 2007) that arises via the idea that the agent-first bias could affect how child learners parse input strings under noisy conditions during acquisition.

2 Evidence for agent-extractability in voice-mismatch contexts

In this section, we offer two types of evidence consistent with the less strict version of the restriction. First, we provide naturally occurring examples where the agent is extracted even in voice-mismatch contexts. Second, we review previous experimental evidence in the literature that serves as the basis of our current study.

2.1 Voice-mismatches "in the wild"

One of our main contributions to the discussion on the nature of the extraction restriction is to provide examples of these voice-mismatch extractions "in the wild." The examples we provide below are naturally occurring and are from various sources online, ranging from newspapers and tabloids, from Wikipedia entries to the Bible.

The examples in (13) involve agent-extraction when the verb exhibits PV. The examples in (13a)–(13b) involve relative clauses; those in (13c)–(13d), topicalizations; and those in (13e)–(13f), *ay*-inversions. The examples in (13c)–(13f) suggest that the acceptability of agent-extractions under PV is not specific to relative clauses.

(13) Extraction of the agent under PV

a. dyirap na kinain ang **buwan**
giraffe LNK ate.PV NOM moon
'giraffe that ate the moon'[5]

[5] A title of a children's book sold by Amazon: https://web.archive.org/web/20240423221216/https://www.amazon.com/Giraffe-That-Ate-Moon-Childrens/dp/1515014258.

b. lalaki-ng alam ang **kanya-ng hangganan**
man-LNK know.PV NOM 3SG.DAT-LNK limit
'man that knows his limit'[6]

c. Hotshots ginulpi ang **Picanto**
PN beat up.PV NOM PN
'Hotshots [a basketball team], they beat Picanto with a landslide victory.'[7]

d. Lalaki, kinagat **ang buwaya na sumakmal sa**
man bit.PV NOM crocodile LNK snatched.AV (with mouth) DAT
kaniya sa Indonesia
3SG.DAT DAT PN
'The man, he bit the crocodile that snatched him using its mouth in Indonesia'[8]

e. Ang mga ama ay kinain **ang maasim na ubas** at ang mga
NOM PL father AY ate.PV NOM sour LNK grapes and NOM PL
ngipin ng anak ay tumalas
tooth GEN child AY sharpened.av
'As for the fathers, they ate sour grapes and as for teeth of the child, they sharpened'[9]

f. Ang mga epidemya-ng ito na may dala-ng mga salot ay pinatay
NOM PL epidemic-LNK DEM LNK EXIST carry-LNK PL plague AY killed.PV
ang **halos 25 milyon-g mga Tsino at iba pa-ng mga Asyano**
NOM almost 25 million-LNK PL Chinese and other also-LNK PL Asian
'As for these epidemics that carried the plague, they killed almost 25 million Chinese and other Asians.'[10]

The examples in (14) involve agent-extraction when the verb exhibits applicative morphology. The example in (14a) involves a relative clause; that in (14b), topicalization; and those in (14c)–(14d), *ay*-inversion. The examples in (14b)–(14d) suggest that acceptability of subject-extractions under applicatives is not specific to relative clauses.

---

[6] A movie teaser for *Goyo: Ang batang heneral* on YouTube (May 5, 2018): https://web.archive.org/web/20240423222353/https://www.youtube.com/watch?v=gT6tXw2OWs&t=75s.

[7] A sports article in PhilStar (January 11, 2018): https://web.archive.org/web/20220503124157/https://www.philstar.com/pilipino-star-ngayon/palaro/2018/01/11/1776517/hotshots-ginulpi-ang-picanto.

[8] GMA's Balitambayan headline (16 January 2024): https://web.archive.org/web/20240424141451/https://www.gmanetwork.com/news/balitambayan/umg/894347/lalaki-kinagat-ang-buwaya-na-sumakmal-sa-kaniya-sa-indonesia/story/.

[9] A Facebook post by the St. Mary & St. Mark Coptic Orthodox Church in the Philippines: https://web.archive.org/web/20240424140117/https://www.facebook.com/StMaryStMarkCopticOrthodoxChurchPhilippines/posts/katanunganang-kasamaan-ba-ng-mga-ama-ay-mapupunta-sa-mga-anak-tulad-ng-nabanggit/1529639590725532/.

[10] The Wikipedia entry on *Salot na itim* (Black Death): https://web.archive.org/web/20240424133951/https://tl.wikipedia.org/wiki/Salot_na_Itim.

(14) Extraction of the agent under applicative morphology

a. lalaki-ng kinunan ng cellphone video ang **babae-ng** **nagsusukat**
man-LNK take.APPL GEN NOM woman-LNK trying on.AV
ng **swimwear**
GEN
'man who took a cellphone video of the woman who was trying on a swimsuit'[11]

b. Andeng, dinalhan ng pagkain si **Marco**
PN bring.APPL GEN food NOM PN
'Andeng, she brought food to Marco'[12]

c. Ang babae po kasi na ibinigay n'yo sa akin ay
NOM woman HON PART LNK gave.PV 2PL.GEN DAT 1.SG.DAT AY
binigyan **ako** ng bunga
gave.APPL 1SG.NOM GEN fruit
'As for the woman who You [God] gave to me, she gave me a fruit'[13]

d. Ang Komisyon sa Wikang Filipino (KWF) ay masusi-ng pinagaralan
NOM Commission on language.LNK Filipino AY thorough-LNK studied.APP
ang **mga** **nagdaa-ng** **ortograpiya-ng** **Filipino** …
NOM PL past-LNK orthography-LNK PN
'As for the Commission on the Philippine language, they carefully examined the previous Filipino orthographies…'[14]

2.2 Experimental evidence from judgment studies

Pizarro-Guevara & Wagers (2020) investigated the extraction restriction experimentally by looking at whether and how comprehenders used voice when processing A-bar dependencies in real-time. They hypothesized that voice allows comprehenders to commit to an interpretation as early as the verb. They reasoned that at this point in the utterance, comprehenders already have information about the event structure via the lexical semantics of the verb, and about the thematic role of the moved element via voice. Thus, at the verb, comprehenders should already have enough information to evaluate the thematic fit of the moved argument.

Their hypothesis crucially hinged on the assumption that there was a one-to-one mapping between voice and the identity of the moved argument. In other words, their hypothesis assumed

[11] GMA News headline (November 4, 2019): https://web.archive.org/web/20240424145628/https://www.youtube.com/watch?v=9fWP9RWj1gY.

[12] Episode recap of TV show *Sandugo*: https://web.archive.org/web/20240424144825/https://www.youtube.com/watch?v=ƀQcVlASPDU.

[13] *Ang Salita ng Dios* version of the Tagalog Bible (Genesis 3:12): https://web.archive.org/web/20240424144117/https://www.bible.com/bible/1264/GEN.3.6-19.ASND.

[14] Department of Education on National Orthography (14 August 2013): https://web.archive.org/web/20240424143805/https://www.deped.gov.ph/wp-content/uploads/2013/08/DO's2013'034.pdf.

the strict version of the restriction. To evaluate whether this assumption was tenable, they conducted a series of acceptability judgment studies and tested three types of dependencies: *ay*-inverted sentences, *wh*-questions, and relative clauses. They manipulated whether the verb exhibited AV or PV (VOICE: AV, PV), and whether the extracted argument was cross-referenced by voice or not (MATCH: ±). The logic of their design is as follows. By comparing the participants' ratings of agent-extractions in AV (+MATCH), as in (15a), against patient-extractions in AV (–MATCH), as in (15b), we can estimate the mismatch penalty incurred when extracting patients not cross-referenced by voice. By comparing their ratings of patient-extractions in PV (+MATCH), as in (15c), against agent-extractions in PV (–MATCH), as in (15d), we can estimate the mismatch penalty incurred when extracting agents not cross-referenced by voice.

(15) a. Ang mga lasinggero ay kumakanta lagi ng mga kundiman…
NOM PL drunkard AY sing.AV always GEN PL love song
'As for the drunkards, they always sing love songs…' [AV, +MATCH]

b. Ang mga kundiman ay kumakanta lagi ang mga lasinggero…
NOM PL love song AY sing.AV always NOM PL drunkard
'As for the love songs, drunkards always sing them…' [AV, –MATCH]

c. Ang mga kundiman ay kinakanta lagi ng mga lasinggero…
NOM PL love song AY sing.PV always NOM PL drunkard
'As for the love songs, drunkards always sing them…' [PV, +MATCH]

d. Ang mga lasinggero ay kinakanta lagi ang mga kundiman…
NOM PL drunkard AY sing.PV always NOM PL love song
'As for the drunkards, they always sing love songs…' [PV, –MATCH]

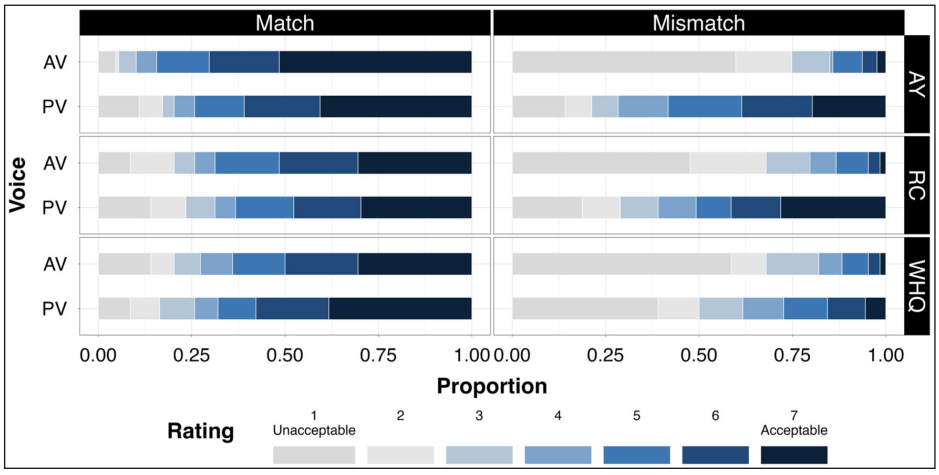


Figure 1: Distribution of ratings in exp. 1 of Pizarro-Guevara & Wagers (2020) by VOICE, MATCH and dependency type. The left panel corresponds to voice-match extractions; the right panel corresponds to voice-mismatch extractions. Blue corresponds to a rating of 7 'acceptable', while gray corresponds to a rating of 1 'unacceptable.'

Figure 1[15] visualizes the distribution of their ratings by VOICE, MATCH, and dependency type. They had three main findings. First, across dependency type, participants gave higher ratings to sentences in voice-match contexts, that is, when the extracted argument is cross-referenced by the voice morphology on the verb. We see this in the left panel: there is a higher proportion of blue and bluish bars compared to the gray(ish) bars. This finding is consistent with both versions of the extraction restriction.

Second, across dependency type, participants gave lower ratings to AV-sentences in voice-mismatch contexts, as in (15b). We see this in the right panel: The top bar for each dependency type has a higher proportion of gray(ish) bars compared to the blue and bluish bars. This finding is again consistent with both versions of the restriction.

Third, participants were more variable with how they rated PV-sentences in voice-mismatch contexts, as in (15d). We see this in the right panel: The bottom bars for the AY and RC panels have a higher proportion of blue and bluish bars compared to gray(ish) bars—relative to the second finding. These proportions are not quite as high as the first finding, however. We see this pattern in the WHQ panel as well, but to a lesser extent. This finding is only consistent with the less strict version.

3 A Bayesian reanalysis of Pizarro-Guevara & Wagers 2020

We first reanalyze experiment 1 of Pizarro-Guevara & Wagers (2020) using a Bayesian approach to statistical inference. Our main goal is to estimate the magnitude of mismatch penalties in AV and PV, so we can incorporate their findings as informative priors in our own judgment studies. Our secondary goal is to quantify how much evidence there was, given their data, for a larger mismatch penalty in AV than in PV.

Recall that they employed a 2×2 design crossing whether the verb exhibited AV or PV (VOICE: AV, PV) and whether the extracted argument was cross-referenced by voice or not (MATCH: $\pm$). They tested three types of dependencies: *ay*-inverted sentences, *wh*-questions, and relative clauses. See (15) for a sample item involving *ay*-inversion.

3.1 Analysis

We fitted Bayesian ordinal mixed effects regression models with a probit link function in R (R Core Team 2020) using brms (Bürkner 2017), a front end of the Stan language for Bayesian estimation of model parameters (Gelman et al. 2015). For each dependency, we used the participants' ratings as the dependent measure and fitted two types of models: crossed

[15] This is a revisualization of **Figure 2** of Pizarro-Guevara & Wagers (2020). They had plotted mean ratings by VOICE, MATCH and dependency type. Their visualization implies that their ratings data are numeric. We opted for a type of visualization that respects the ordinal nature of their ratings data.

and nested. In the crossed model, we included the main effects of VOICE, MATCH, and their interaction as fixed effects. In the nested model, we separated the fixed effects parameters for MATCH in the AV and PV conditions. The random effects structure included random intercepts for participants and items, and VOICE, MATCH, and their interaction as random slopes for participants and items. **Table 2** provides the contrast coding for all the fixed effects in both models for all the dependencies.

		Crossed model			**Nested model**		
VOICE	**MATCH**	**VOICE**	**MATCH**	**V × M**	**VOICE**	**M in AV**	**M in PV**
AV	+MATCH	–.5	–.5	.25	–.5	–.5	0
AV	–MATCH	–.5	.5	–.25	–.5	.5	0
PV	+MATCH	.5	–.5	–.25	.5	0	–.5
PV	–MATCH	.5	.5	.25	.5	0	.5

Table 2: Table of contrast coding coefficients used for the crossed and nested statistical models described in the text. Experimental conditions are given in rows, model contrasts given in columns.

As a starting point, we used Normal(0,5)[16] as the priors for all the fixed effects and the intercept. This means that the parameter has a mean of 0 and a standard deviation of 5. These are uninformative priors as they do not place strong constraints on the model's predictions, and incorporate very little knowledge about what makes a plausible ratings distribution. We used LKJ(2) as the prior for our correlation matrix. For each model, we ran four Monte Carlo Markov chains in parallel, with 40,000 samples each. The first 8,000 were always discarded as part of warm-up. For all the models reported below, the R-hat statistic was at 1.0. No divergences were observed.

In addition, we quantified how much evidence we have for an effect by conducting Bayes factor analyses. We compared the marginal likelihood of the full model against the marginal likelihood of a null model. In the crossed models, the full model included VOICE, MATCH and their interaction, while the null model excluded the interaction. Comparing the marginal likelihoods of these two models quantifies how much evidence we have in favor of/against the mismatch penalty being larger in AV than in PV. In the nested models, the full model included VOICE, MATCH in AV, and MATCH in PV. There are two null models: one that excluded MATCH

[16] The notation Normal(X, Y) means that the parameter has a mean of X, a standard deviation of Y, and is normally distributed.

in AV and one that excluded MATCH in PV. Comparing the marginal likelihoods of the full model and the null model that excludes MATCH in AV quantifies how much evidence we have in favor of/against the presence of a mismatch penalty in AV. Similarly, comparing the marginal likelihoods of the full model and the null model that excludes MATCH in PV quantifies how much evidence we have in favor of/against the presence of a mismatch penalty in PV.

Because BF analyses are highly sensitive to prior specifications (Nicenboim et al. 2020), we also conducted a sensitivity analysis by considering a range of priors, Normal(0,3) and Normal(0,1), for the interaction term in the crossed models and the separated MATCH parameter in the nested models. We used these increasingly informative priors for the parameters of interest, but kept the rest of the priors as originally described above.

3.2 Results

See our OSF page for two tables: (i) the mean posterior distribution, along with the 95% Bayesian credible intervals (CrI), which indicate where the plausible parameter values for these fixed effects parameters lie, given the data and the priors; and (ii) the sensitivity analysis, where we calculated the Bayes factor 100 times and report the median over these values, along with a 95% interval representing the 2.5 and 97.5 percentiles. For present purposes, what is important is that a Bayes factor smaller than 1 favors the null model, indicating evidence that the effect of interest is absent. A value larger than 1 favors the full model, indicating evidence that the effect of interest is present. A value larger than 10 is considered to be strong evidence (Jeffreys 1998).

Ay-inverted sentences. There is strong evidence for a mismatch penalty in AV. The evidence is between 2.58×10^9 and 1.26×10^{10}, suggesting that when the verb has AV, agent-extractions (+MATCH) were rated higher than patient-extractions (–MATCH). There is inconclusive evidence for a mismatch penalty in PV, ranging from anecdotal evidence for the null model to anecdotal evidence for the full model. The evidence is between .23 and 2.00, suggesting that when the verb has PV, we do not have enough evidence, given the data, to say whether patient-extractions (+MATCH) were rated higher than agent-extractions (–MATCH) or whether these were not rated differently from each other. There is strong evidence for a larger mismatch penalty in AV than in PV. The evidence is between 2.62×10^4 and 8.96×10^4.

Head-initial RCs. There is strong evidence for a mismatch penalty in AV. The evidence is between 8.81×10^3 and 3.57×10^4, suggesting that when the verb has AV, agent-extractions (+MATCH) were rated higher than patient-extractions (–MATCH). There is moderate evidence for no mismatch penalty in PV. The evidence is between .09 and .41, suggesting that when the verb has PV, patient-extractions (+MATCH) were not rated differently from agent-extractions (–MATCH). There is strong evidence for a larger mismatch penalty in AV than in PV. The evidence is between 11.22 and 15.90.

Wh-questions. There is strong evidence for a mismatch penalty in AV. The evidence is between 3.50×10^4 and 9.33×10^4, suggesting that when the verb has AV, agent-extractions (+MATCH) were rated higher than patient-extractions (–MATCH). There is also strong evidence for a mismatch penalty in PV. The evidence is between 771 and 2088, suggesting that when the verb has PV, patient-extractions (+MATCH) were rated higher than agent-extractions (–MATCH). There is anecdotal to moderate evidence for the mismatch penalties in AV and PV not being reliably different. The evidence is between .28 and .91.

3.3 Discussion

We reanalyzed experiment 1 of Pizarro-Guevara & Wagers (2020) using a Bayesian approach to statistical inference to estimate the magnitude of a mismatch penalty in AV and PV. We will use these values as informative priors for the parameters of interest in our replication, which will be presented in the next section. To estimate the mean of the interaction, we took the average of the mean point estimates of the interaction across the different prior specifications. For example, we estimated the mean of the interaction for head-initial RCs to be 1.58, which is the average of 1.74, 1.69, and 1.31. To estimate the standard deviation of the interaction, we took the average of the errors (i.e. the width of 95% CrI, divided by four). For example, we estimated the standard deviation of the interaction for head-initial RCs to be .56, which is the average of .59, .58, and .50. We followed the same procedure for calculating the prior specifications for the MATCH in AV and MATCH in PV of the nested models.

In our replication, we will be incorporating their insights in our models as informative priors. For head-initial RCs, we will use Normal(1.58,. 56) for the interaction in the crossed model, and Normal(–1.90,. 09) and Normal(–.17,. 10) for MATCH in AV and MATCH in PV in the nested models, respectively. For *wh*-questions, we will use Normal(.64,. 10) for for the interaction in the crossed model, and Normal(–2.26,. 10) and Normal(–1.64,. 08) for MATCH in AV and MATCH in PV in the nested models, respectively.

Our reanalysis also quantified how much evidence there was in their experiment for a larger mismatch penalty in AV than in PV in order to see whether their empirical generalizations were warranted, given the data. Our reanalysis is congruent with their generalizations in broad strokes. We found strong evidence for a mismatch penalty in AV across dependency types. We also found that there was variable evidence for a mismatch penalty in PV. However, the finer details varied.

In *ay*-inverted sentences, the mismatch penalty in AV was larger than in PV. What we are uncertain about is whether it was a smaller mismatch penalty in PV or whether it was absent. In relative clauses, there was no mismatch penalty in PV and thus, the mismatch penalty in AV was larger than in PV. In *wh*-questions, there was evidence that the mismatch penalties in AV and PV did not differ.

Given their data, the empirical terrain that emerges is the following: only the extraction behavior in *wh*-questions is more consistent with the strict version of the extraction restriction. Their extraction behavior in *ay*-inverted sentences and head-initial RCs is more consistent with the less strict version of the restriction.

Their findings raise three important questions. First, are these findings replicable? Second, if they are, how can the less strict extraction restriction be modeled? Lastly, what does this tell us about the landscape of A-bar extraction in general? To answer the first question, we conducted our own judgment studies that sought to replicate and extend their findings. We present our studies comparing the extraction behavior of Tagalog speakers under AV and PV in section 4.2, and their extraction behavior under applicatives in section 4.3. To answer the second question, in lieu of presenting a formal analysis for the less strict version of the extraction restriction, we suggest a way to think about this variation as a form of grammar competition in learning, in the sense of Han et al. (2007), that arises via the idea that an agent-first bias could filter the linguistic input or otherwise affect how child learners parse input strings under noisy conditions during acquisition. While all three questions are equally important, we leave it to future work to develop a formal analysis for the second question and to answer the last question. We invite readers interested in a formal analysis of voice-mismatch extractions to see Hsieh (2023) for one way to model the phenomenon.

4 The current study: A replication and an extension

This section presents our investigation of the extraction restriction. We conducted a series of acceptability judgment studies, designed to replicate (experiment 1) and extend (experiment 2) the findings of Pizarro-Guevara & Wagers (2020). In the first half, we present our studies comparing voice-mismatch extractions in AV and PV; in the second half, we present our studies comparing voice-mismatch extraction in applicatives. The main question we address here is whether speakers of Tagalog find it more acceptable to extract the agent in voice-mismatch contexts.

We chose to investigate the extraction behavior of Tagalog speakers in head-initial RCs and *wh*-questions. In our Bayesian reanalysis of Pizarro-Guevara & Wagers' data, we found that head-initial RCs provided the most convincing evidence consistent with the less strict version of the restriction, while *wh*-questions provided evidence consistent with the strict version. By choosing these two dependencies, we investigated environments where each view of the restriction had the best chance to succeed. We also chose to investigate the extraction behavior of Tagalog speakers in head-final RCs. As discussed in Section 1.2, the speakers who do allow voice-mismatch agent-extractions in head-initial RCs do not allow them in head-final RCs (Ceña & Nolasco 2011).

4.1 Participants and procedure

We recruited 30 participants from the University of the Philippines, Diliman and surrounding communities. Participants needed to (i) be 18 years or older; (ii) live in or around Metro Manila at time of testing; and (iii) use Tagalog every day.[17] Prior to any analysis, we excluded one participant due to internet connection issues. The remaining 29 participants ranged from 18 to 42 years of age ($M = 25$, $SD = 5$). They received either a gift certificate from a local coffee shop or a bookstore, valued at 300 Philippine Pesos, for participating.

The same speakers participated in experiments 1 and 2, which were combined in one session and were randomized along with items involving extraction out of islands and complement clauses. The judgment studies were administered online via Ibex (Drummond 2016). Participants were asked to rate the acceptability of sentences using a 7-point scale, with 1 being *talagang hindi katanggap-tanggap* 'really unacceptable' and 7 being *talagang katanggap-tanggap* 'really acceptable.' They were explicitly instructed that a sentence is *katanggap-tanggap* if (i) as a speaker of Tagalog, they can imagine themselves saying or writing this; and (ii) they can imagine other Tagalog speakers saying or writing this, as well. On average, each session took about 30 to 40 minutes to complete.

4.2 Experiment 1: Comparing voice-mismatch extractions in AV and PV

We replicated the finding that there is a consistently large mismatch penalty in AV, irrespective of the dependency involved. We did not replicate the variable mismatch penalty in PV. Instead, we found a consistent mismatch penalty in PV, albeit a smaller one, in head-initial RCs and *wh*-questions. In other words, agents enjoyed a certain degree of extractability, congruent with the less strict version of the extraction restriction. This did not extend to head-final RCs, however. This is consistent with the observation made by Ceña & Nolasco (2011) and is more congruent with the strict version.

4.2.1 Materials

We followed the design of the original study and employed a 2×2 factorial design, crossing whether the verb exhibited AV or PV (VOICE: AV, PV), and whether the extracted argument is cross-referenced by voice or not (MATCH: +MATCH, –MATCH).

[17] We acknowledge that the recruitment criteria were fairly lax. It should be noted that people from all over the Philippines live in Metro Manila and most have some proficiency in Tagalog. The laxness did lead to the inclusion of participants who were also exposed to or possibly even speak other Philippine languages. While some would argue that this introduced significant variance in the design of the study, this inclusion does not pose any serious concerns. Most Filipinos whose native language is not Tagalog are early sequential bilinguals. They first learn the language of the region or province where they live, and upon entry into the educational system (around age 5), they learn Tagalog and English (Galang 1988; 2001). That they were in Metro Manila at the time of testing ensured the active use of Tagalog in their daily lives. We view this inclusion positively because it reflected the social reality of who uses Tagalog in the Philippines, thereby increasing the study's ecological validity.

We created three sets of 24 items, one with *wh*-questions, one with head-initial relative clauses, and one with head-final relative clauses (TYPE: HI, WHQ, HF), with each item distributed evenly across 4 lists via Latin square design. These were randomized by the three sets of 24 items used in experiment 2, as well as 48 distractor items. These distractor items examined Tagalog speakers' extraction behavior out of *wh*-islands and complex NP-islands, and will become relevant in Section 5. See our OSF page for the supplementary material associated with these distractor items.

We provide a sample head-initial RC item, a sample *wh*-question item, and a sample head-final RC item in (16)–(18), respectively. Items (a) and (b) involve AV, while items (c) and (d) involve PV. Items (a) and (c) involve voice-match extractions, while (b) and (d) involve voice-mismatch extractions.

(16) Medyo baduy ang…
Somewhat tacky NOM
The … is somewhat tacky. [HI]

a. manliligaw na kumakanta ng kundiman para sa nililigawan.
suitor LNK sings.AV GEN love song for DAT wooee
'suitor that sings a love song…' [+MATCH, AV]

b. kundiman na kumakanta ang manliligaw para sa nililigawan.
love song LNK sings.AV NOM suitor for DAT wooee
'love song that the suitor sings…' [–MATCH, AV]

c. kundiman na kinakanta ng manliligaw para sa nililigawan.
love song LNK sings.PV GEN suitor for DAT wooee
'love song that the suitor sings …' [+MATCH, PV]

d. manliligaw na kinakanta ang kundiman para sa nililigawan
suitor LNK sings.PV NOM love song for DAT wooee
'suitor that sings the love song …' [–MATCH, PV]

(17) Sinabi ni Juan kung…
said.PV GEN PN $comp_{[+Q]}$
'Juan said…' [WHQ]

a. sino ang bumibili ng mangga para sa ulila buwan-buwan.
who NOM buys.AV GEN mango for DAT orphan every month
'… who buys mango for the orphan every month.' [+MATCH, AV]

b. ano ang bumibili ang magsasaka para sa ulila buwan-buwan.
what NOM buys.AV NOM farmer for DAT orphan every month
'… what the farmer buys for the orphan…' [–MATCH, AV]

c. ano ang binibili ng magsasaka para sa ulila buwan-buwan.
what NOM buys.PV GEN farmer for DAT orphan every month
'… what the farmer buys for the orphan every month.' [+MATCH, PV]

d. sino ang binibili ang mangga para sa ulila buwan-buwan.
who NOM buys.PV NOM mango for DAT orphan every month
'… who buys the mango for the orphan…' [–MATCH, PV]

(18) Talaga-ng pangit ang…
really-LNK ugly NOM
The … is really ugly. [HF]

a. nagtahi ng damit para sa babae na lalaki
sewed.AV GEN dress for DAT woman LNK man
'man who sewed a dress for the woman' [+MATCH, AV]

b. nagtahi ang lalaki para sa babae na damit.
sewed.AV NOM man for DAT woman LNK dress
'dress that the man sewed for the woman' [–MATCH, AV]

c. tinahi ng lalaki para sa babae na damit
sewed.PV textscgen man for DAT woman LNK dress
'dress that the man sewed for the woman' [+MATCH, PV]

d. tinahi ang damit para sa babae na lalaki
sewed.PV NOM dress for DAT woman LNK man
'man who sewed the dress for the woman' [–MATCH, PV]

4.2.2 Analysis

We fitted the same type of models and used the same model structure as in our reanalysis. We also used the same contrasts from **Table 2**. We used priors informed by our reanalysis of Pizarro-Guevara & Wagers (2020) for *wh*-questions and head-initial RCs. In head-initial RCs, the prior that we used for the interaction in the crossed model is Normal(1.58,. 56). The priors we used in the nested model for MATCH in AV and MATCH in PV are Normal(–1.90,. 09) and Normal(–.17,. 10), respectively. In *wh*-questions, the prior we used for the interaction in the crossed model is Normal(.64, .10). The priors in the nested model for MATCH in AV and MATCH in PV are Normal(–2.26,. 10) and Normal(–1.64,. 08), respectively. Because we did not have any previously collected ratings data to inform the priors of the interaction for head-final RCs, we used Normal(0,5) as a starting point. This is a fairly uninformative prior, which means that it does not place strong constraints on the model's predictions and incorporate very little knowledge about what makes a plausible ratings distribution.

We used the same model parameters as in our reanalysis in section 3. We used LKJ(2) as the prior for our correlation matrix. For each model, we ran four Monte Carlo Markov chains in parallel, with 40,000 samples each. The first 8,000 were always discarded as part of warm-up. For all the models reported below, the R-hat statistic was at 1.0. No divergences were observed. We also conducted a sensitivity analysis by considering a range of priors, Normal(0,3) and

Normal(0,1), for the interaction term in the crossed models and the separated MATCH parameter in the nested models. We used these slightly less informative priors for the parameters of interest, but kept the rest of the priors as Normal(0,5).

4.2.3 Results and discussion

In **Figure 2**, we visualize the distribution of participant ratings by VOICE, MATCH, and dependency type. See our OSF page for the following: (i) the mean posterior distribution, along with the 95% Bayesian CrI; and (ii) the sensitivity analysis, where we report the median Bayes factor and a 95% interval representing the 2.5 and 97.5 percentiles, calculated over 100 estimates of the Bayes factor. What is important for present purposes is that a value larger than 1 favors the full model, indicating evidence that the effect of interest is present. A value larger than 10 is considered to be strong evidence.

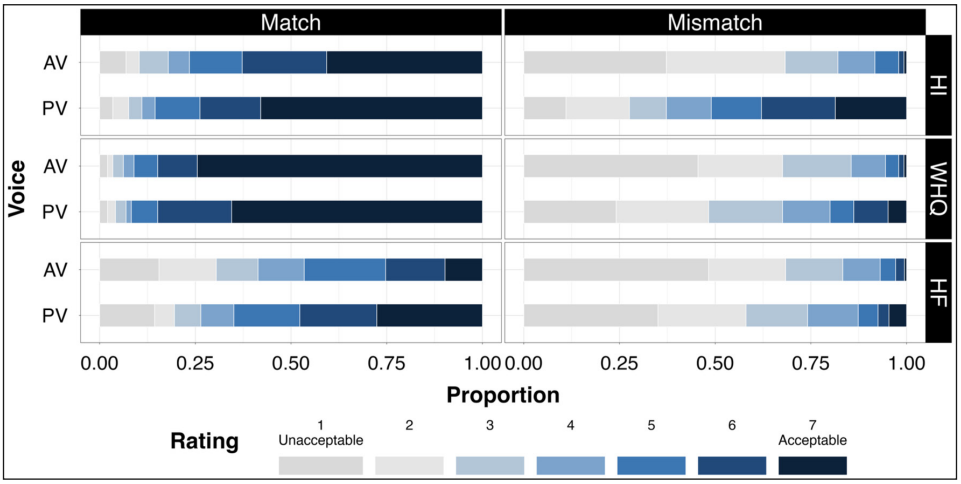


Figure 2: Distribution of ratings in experiment 1 by VOICE, MATCH and dependency type. The left panel corresponds to voice-match extractions; the right panel corresponds to voice-mismatch extractions. Blue corresponds to a rating of 7 'acceptable', while gray corresponds to a rating of 1 'unacceptable.'

Head-initial RCs. There is strong evidence for a mismatch penalty in AV. The evidence is between 1.21×10^6 and 4.46×10^6, suggesting that when the verb has AV, agent-extractions (+MATCH) were rated higher than patient-extractions (–MATCH). There is also strong evidence for a mismatch penalty in PV. The evidence is between 81 and 296, suggesting that when the verb has PV, patient-extractions (+MATCH) were also rated higher than agent-extractions (–MATCH). At best, there is anecdotal evidence for a larger mismatch penalty in AV than in PV. The evidence is between .87 and 2.68.

Wh-questions. There is strong evidence for a mismatch penalty in AV. The evidence is between 4.71×10^5 and 1.68×10^9, suggesting that when the verb has AV, agent-extractions

(+MATCH) were rated higher than patient-extractions (–MATCH). There is also strong evidence for a mismatch penalty in PV. The evidence is between 6.02×10^5 and 2.16×10^9, suggesting that when the verb has PV, patient-extractions (+MATCH) were also rated higher than agent-extractions (–MATCH). There is strong evidence for a larger mismatch penalty in AV than in PV. The evidence is between 145 and 393.

Head-final RCs. There is strong evidence for a mismatch penalty in AV. The evidence is between 1.66×10^5 and 6.55×10^5, suggesting that when the verb has AV, agent-extractions (+MATCH) were rated higher than patient-extractions (–MATCH). There is also strong evidence for a mismatch penalty in PV. The evidence is between 5.61×10^4 and 2.20×10^5, suggesting that when the verb has PV, patient-extractions (+MATCH) were also rated higher than agent-extractions (–MATCH). There is moderate evidence for no difference in the mismatch penalties between AV and PV. The evidence is between .07 and .32.

Here our main goal was to see whether we can replicate the findings of Pizarro-Guevara & Wagers (2020). Let's first focus on 2 of the 3 dependencies they examined, head-initial RCs and *wh*-questions. Like them, we found that voice-mismatch extractions in AV consistently incurred a large mismatch penalty, irrespective of the dependency involved. Unlike them, we found that voice-mismatch extractions also consistently incurred a mismatch penalty, albeit a smaller one, compared to voice-mismatch extractions in AV.

In head-initial RCs, our reanalysis and our replication found evidence for a larger mismatch penalty in AV than in PV. Where they vary is in terms of how large the mismatch penalty is in PV. Our reanalysis found moderate evidence for there being no mismatch penalty in PV, while our replication found strong evidence for a mismatch penalty in PV, albeit a smaller one compared to that in AV. One potential reason for the difference in effect sizes could be the priors used. Our reanalysis used uninformative priors, while our replication used more informative priors from our reanalysis. When the effect size is small, uninformative priors tend to favor the null model (Nicenboim et al. 2020). The estimated effect size of the mismatch penalty in PV in the original study is –.17 (c.f. the mismatch penalty in AV is between –1.98 and –1.76). Meanwhile, with more informative priors, the estimated effect size of the mismatch penalty in PV in our replication is between –1.43 and –.95 (c.f. the mismatch penalty in AV is between –2.40 and –2.21). Abstracting away from the underlying cause of the difference, the results of the original and of our replication are inconsistent with the strict restriction. Crucially, both are congruent with the less strict version.

In *wh*-questions, our reanalysis and our replication found strong evidence for a mismatch penalty in PV. Where they vary, however, is whether the size of this penalty is smaller or comparable to the size of the mismatch penalty in AV. Our reanalysis found no reliable difference between the mismatch penalties incurred in AV and in PV. Our replication, on the other hand, found strong evidence for a larger mismatch penalty in AV than in PV. One potential reason is the

priors used. As discussed above, uninformative priors tend to favor the null model, which could be the reason why our reanalysis did not detect any reliable difference between the mismatch penalties incurred in AV and in PV. The difference could also be due to the differences in the shape of the experimental items. Our experimental items differed along two dimensions. First, Pizarro-Guevara & Wagers used matrix *wh*-questions, while ours used embedded *wh*-questions. Second, they used *d*-linked interrogatives (i.e., akin to English *which X*), while ours used bare interrogatives. We have not done any systematic investigations on the effect of embedding context and *d*-linking on extractability. We leave these questions open for now. What is important for present purposes is that in *wh*-questions, there is some evidence consistent with the strong version and some evidence consistent with the less strict version.

A secondary goal was to determine the extent to which agent-extractability in voice-mismatch contexts varies in head-initial and head-final RCs. Recall that Ceña & Nolasco (2011) observed that agent-extractions under PV are permissible in head-initial RCs by some speakers, but never in head-final RCs. This generalization is consistent with our data. In head-initial RCs, the mismatch penalty incurred in PV is smaller than that in AV. However, in head-final RCs, there is moderate to strong evidence to believe that the mismatch penalties in AV and PV are not reliably different.

4.3 Experiment 2: Comparing voice-mismatch extractions in applicatives

We found that agent-extractions incurred a mismatch penalty, albeit a smaller one relative to patient-extractions, when the verb has applicatives in head-initial RCs. These results are more congruent with the less strict version of the extraction restriction. It is unclear, however, if this permissibility extended to *wh*-questions and head-final RCs. At best, given the data, we only have anecdotal evidence that agent-extractions and patient-extractions under applicative incurred comparable mismatch penalties in *wh*-questions, and inconclusive evidence in head-final RCs.

4.3.1 Materials

The experiment employed a 2×2 factorial design, crossing whether the extracted argument was cross-referenced by voice or not (MATCH: +MATCH, –MATCH) and whether extracted argument is the agent or not the agent (AGEX: AG, NAG). The design of experiment 2 is quite different from that of experiment 1, but the basic idea of calculating mismatch penalties remains the same. We make explicit the comparisons that we are making first in prose and then we summarize them in **Table 3**.

When a verb has applicative morphology, there are three logical possibilities for which argument can be extracted: (i) the applied object, which is cross-referenced by the applicative morphology; (ii) the agent; and (ii) the patient. Because we were interested in seeing whether agents can be extracted under applicatives, we needed a baseline for extractions under

applicatives, one where the extracted argument is cross-referenced by applicative morphology. Thus, one of the conditions of the experiment looked at the extraction of applied objects under applicatives (i.e., [+MATCH, NAG]). We also needed a baseline for agent-extraction, one where the extracted agent is cross-referenced by voice morphology. Thus, one of the conditions of the experiments looked at agent-extraction under AV [+MATCH, AG]). Now that these baselines are established, we can compare them to the actual condition of interest, that is, to agent-extraction when it not voice-matched [–MATCH, AG]). The remaining condition is when it is not voice-matched and it is not agent-extraction: Patient-extraction under applicative [–MATCH, NAG]).

	MATCH	
AGEX	**+MATCH**	**–MATCH**
AG	Agent-extraction under AV	Agent-extraction under applicative
NAG	Applied object-extraction under applicative	Patient-extraction under applicative

Table 3: Overview of how the experimental conditions in experiment 2 mapped onto the types of extractions.

The logic of the design is as follows. By comparing the participants' ratings of agent-extractions under applicative against those of agent-extractions under AV, we get an estimate of the mismatch penalty incurred when extracting the agent not cross-referenced by voice. By comparing their ratings of patient-extractions under applicative against those of applied object-extractions under applicative, we get an estimate of the mismatch penalty incurred when extracting non-agents not cross-referenced by voice.

We created three sets of 24 items, one with head-initial RCs, one with *wh*-questions, and one with head-final RCs (TYPE: WHQ, HI, HF), with each item distributed evenly across 4 lists via Latin square design. These were randomized by the three sets of 24 items used in experiment 1, as well as 48 distractor items involving extraction out of *wh*-islands and complex NP-islands. We provide a sample item involving head-initial RCs, *wh*-questions, and head-final RCs in (19)–(21), respectively. Items (a) and (b) involve agent-extractions, while items (c) and (d) involve non-agent extractions. Items (a) and (c) involve voice-match extractions, while (b) and (d) involve voice-mismatch extractions.

(19) Talagang nakakatuwa …
Really funny
'The… is really funny.' [HI]

a. tatang na nagturo ng matematika sa batang yagit
old man LNK taught.AV GEN math DAT street urchin
'…old man that taught math to the street urchin…' [+MATCH, AG]

b. tatang na tinuruan ang batang yagit ng matematika
old man LNK taught.APPL NOM street urchin GEN math
'…old man that taught math to the street urchin…' [–MATCH, AG]

c. batang yagit na tinuruan ng tatang ng matematika
street urchin LNK taught.APPL GEN old man GEN math
'… street urchin that the old man taught math to…' [+MATCH, NAG]

d. matematika na tinuruan ng tatang ang batang yagit
math LNK taught.APPL GEN old man NOM street urchin
'…math that the old man taught to the street urchin…' [–MATCH, NAG]

(20) Sinabi ni Inday kung…
said.PV GEN PN $COMP_{[+Q]}$
'Inday said…' [WHQ]

a. sino ang bumibili ng mangga para sa ulila buwan-buwan.
who NOM buys.AV GEN mango for DAT orphan every month
'… who buys a mango for the orphan …' [+MATCH, AG]

b. sino ang binibilhan ang ulila ng mangga buwan-buwan.
who NOM buys.APPL NOM orphan GEN mango every month
'…who bought mango for the orphan…' [–MATCH, AG]

c. sino ang binibilhan ng magsasaka ng mangga buwan-buwan.
who NOM buys.APPL GEN farmer GEN mango every month
'… who the farmer buys a mango for …' [+MATCH, NAG]

d. ano ang binibilhan ng magsasaka ang ulila buwan-buwan.
what NOM buys.APPL GEN farmer NOM orphan every month
'… what the farmer bought for the orphan…' [–MATCH, NAG]

(21) Nakakahighblood ang…
hypertension-inducing NOM
'The … is hypertension-inducing (literal)/aggravating (figurative).' [HF]

a. nagluto ng bulalo para sa pamilya na kuya
cooked.AV GEN a type of soup for DAT family LNK big brother
'…big brother that cooked bulalo for the family…' [+MATCH, AG]

b. nilutuan ang pamilya ng bulalo na kuya
cooked.APPL NOM family GEN soup LNK big brother
'…big brother that cooked bulalo for the family…' [–MATCH, AG]

c. nilutuan ng kuya ng bulalo na pamilya
cooked.APPL GEN big brother GEN soup LNK family
'…family who the big brother cooked bulalo for…' [+MATCH, NAG]

d. nilutuan ng kuya ang pamilya na bulalo
cooked.APPL GEN big brother NOM family LNK soup
'…bulalo that the big brother cooked for the family…' [–MATCH, NAG]

4.3.2 Analysis

We fitted Bayesian ordinal mixed effects regression models in R using brms. For each dependency, we used the participants' ratings as the dependent measure and fitted crossed and nested models. In the crossed model, we included the main effects of AGEX, MATCH, and their interaction as fixed effects. In the nested model, we separated the fixed effects parameters for MATCH in the AG and NAG conditions. The random effects structure included random intercepts for participants and items, and AGEX, MATCH, and their interaction as random slopes for participants and items.

Because this is the first experimental investigation of how voice-mismatches affect extraction possibilities when the verb exhibits applicative morphology, we used uninformative priors. We used Normal(0,5) as the priors for all the fixed effects and the intercept as a starting point. We used LKJ(2) as the prior for our correlation matrix. We also conducted a sensitivity analysis by considering a range of priors—Normal(0,3) and Normal(0,1)—for the interaction term in the crossed models and the separated MATCH parameter in the nested models.

4.3.3 Results and discussion

In **Figure 3**, we visualize the distribution of participant ratings by AGEX, MATCH, and dependency type. See our OSF page for the following: (i) the mean posterior distribution, along with the 95% Bayesian CrI; and (ii) the sensitivity analysis, where we report the median Bayes factor and a 95% interval representing the 2.5 and 97.5 percentiles, calculated over 100 estimates of the Bayes factor. What is important for present purposes is that a value larger than 1 favors the full model, indicating evidence that the effect of interest is present. A value larger than 10 is considered to be strong evidence.

Head-initial RCs. There is strong evidence for a voice-mismatch penalty when extracting the agent. The evidence is between 18.12 and 88.54, suggesting that extracting the agent when the verb had AV (+MATCH) was rated higher than when the verb had applicative morphology (–MATCH). There is also strong evidence for a voice-mismatch penalty when extracting the patient. The evidence is between 6.38×10^4 and 3.13×10^5, suggesting that when the verb had applicative morphology, extracting the applied object (+MATCH) was rated higher than extracting the patient (–MATCH). There is anecdotal to moderate evidence for a smaller mismatch penalty in agent-extractions under applicative than in patient-extractions. The evidence is between 1.94 and 6.17.

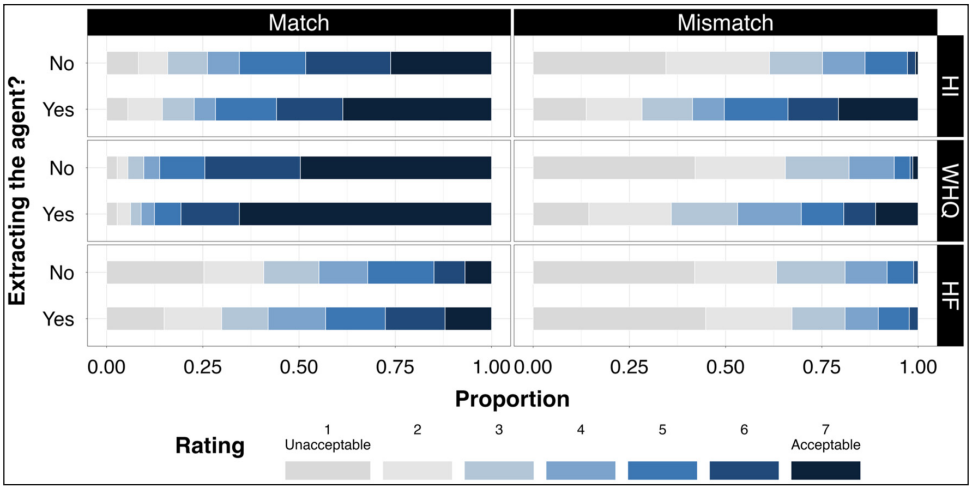


Figure 3: Distribution of ratings in experiment 2 by AGEX, MATCH and dependency type. The left panel corresponds to voice-match extractions; the right panel corresponds to voice-mismatch extractions. Blue corresponds to a rating of 7 'acceptable', while gray corresponds to a rating of 1 'unacceptable.'

Wh-questions. There is strong evidence for a voice-mismatch penalty when extracting the agent. The evidence is between 2.34×10^4 and 7.19×10^5, suggesting that extracting the agent when the verb had AV (+MATCH) was rated higher than when the verb had applicative morphology (–MATCH). There is also strong evidence for a voice-mismatch penalty when extracting the patient. The evidence is between 5.68×10^6 and 1.75×10^8, suggesting when the verb had applicative morphology, extracting the applied object (+MATCH) was rated higher than extracting the patient (–MATCH). At best, there is anecdotal evidence for a smaller mismatch penalty penalty in agent-extractions under applicative than in patient-extractions. At worst, there is anecdotal evidence that their mismatch penalties are not different. The evidence is between .30 and 1.07. In other words, given the data, we do not know whether extracting the agent incurred a smaller mismatch penalty than extracting the patient when the verb had applicative morphology or whether there is no difference between their mismatch penalties.

Head-final RCs. There is strong evidence for a voice-mismatch penalty when extracting the agent. The evidence is between 1.21×10^4 and 8.70×10^4, suggesting that extracting the agent when the verb had AV (+MATCH) was rated higher than when the verb had applicative morphology (–MATCH). There is also strong evidence for a voice-mismatch penalty when extracting the patient. The evidence is between 257 and 1857, suggesting that when the verb had applicative morphology, extracting the applied object (+MATCH) was rated higher than extracting the patient (–MATCH). At best, there is anecdotal evidence for a smaller mismatch penalty penalty in agent-extractions under applicative than in patient-extractions. At worst, there is anecdotal evidence that their mismatch penalties are not different. The evidence is between

.45 and 1.95. In other words, given the data, we do not know whether extracting the agent incurred a smaller mismatch penalty than extracting the patient when the verb had applicative morphology or whether there is no difference between their mismatch penalties.

Here our main goal was to examine the extraction behavior of Tagalog speakers when the verb exhibits applicative morphology. We found that agent-extractions incurred a mismatch penalty, albeit a smaller one relative to patient-extractions, when the verb has applicatives in head-initial RCs. This echoes the findings in experiment 1. It is unclear, however, if this permissibility extended to *wh*-questions and head-final RCs.

4.4 The empirical terrain: Summary of Experiments 1 and 2

We presented a series of acceptability judgment experiments, designed to replicate and extend the findings of Pizarro-Guevara & Wagers (2020). We compared the extraction behavior of Tagalog speakers when the verb exhibits AV and PV (experiment 1), and when the verb exhibits applicative morphology (experiment 2).

In experiment 1, head-initial RCs and *wh*-questions had comparable extraction profiles. The ratings distributions of voice-match extractions were more concentrated toward the higher values of the Likert scale. The ratings distribution of voice-mismatch extractions when the verb had AV (i.e., patient-extractions under AV), toward the lower values. The ratings distribution of voice-mismatch extractions when the verb had PV (i.e., agent-extractions under PV) was in a sense more intermediate. It was not quite like the distributions of the other two. This empirical terrain is more congruent with the less strict version of the extraction restriction. However, head-final RCs had a different extraction profile. The ratings distributions of voice-match extractions were more concentrated toward the higher values of the Likert scale, while the ratings distributions of voice-mismatch extractions were more concentrated toward the lower values. This empirical terrain is more congruent with the strong version of the extraction restriction.

In experiment 2, head-initial RCs replicated the extraction profile exhibited in experiment 1. The ratings distributions of voice-mismatch extractions were more concentrated toward the higher values; that of voice-mismatch patient-extractions, toward the lower values; and that of voice-match agent-extractions, more intermediate. *Wh*-questions and head-final RCs replicated almost all of the extraction profile they exhibited in experiment 1. What is unclear is their behavior in voice-mismatch agent-extractions. At best, there is anecdotal evidence for a smaller mismatch penalty penalty in agent-extractions under applicative than in patient-extractions. At worst, there is anecdotal evidence that their mismatch penalties are not different.

It is crucial to understand the nature of the ratings distribution that underlies the intermediate ratings distribution of most of the voice-mismatch agent-extractions. Intermediate mean ratings are compatible with (at least) two types of response distributions. In one scenario, participants classified these extractions categorically. The intermediate ratings distribution

stemmed from a mixture of acceptable and unacceptable responses, creating an "illusory" intermediate ratings distribution. In another scenario, participants perceived these extractions as somewhere between acceptable and unacceptable. The intermediate ratings distribution therefore reflected an overall upward shift in acceptability relative to unacceptable extractions or an overall downward shift in acceptability relative to acceptable extractions, creating a true intermediate ratings distribution. The next section develops quantitative models to estimate the participants' judgment process to gain a more nuanced understanding of the underlying distributions of their ratings.

5 Quantitative models of the participants' judgment process

In order to estimate the underlying distribution of the participants' ratings of agent-extractions when the verb exhibited non-AV morphology, as reported in Sections 4.2 and 4.3, we develop quantitative models using the framework developed by Dillon et al. (2017). The distributions of voice-mismatch agent-extractions exhibited are in a sense intermediate in that they are not quite like the distributions of voice-match extractions (the left panels in **Figures 1, 2** and **3**) and not quite like the distributions of voice-mismatch non-agent extractions (top bars of the right panel for each dependency type). This intermediate ratings distribution is compatible with (at least) two types of distributions: (i) a mixture of categorical responses, and (ii) an overall shift in acceptability. These two possible scenarios are illustrated in **Figure 4**. Following Dillon et al., we refer to the first scenario as having a DISCRETE distribution, and the second scenario, as having a GRADIENT distribution.

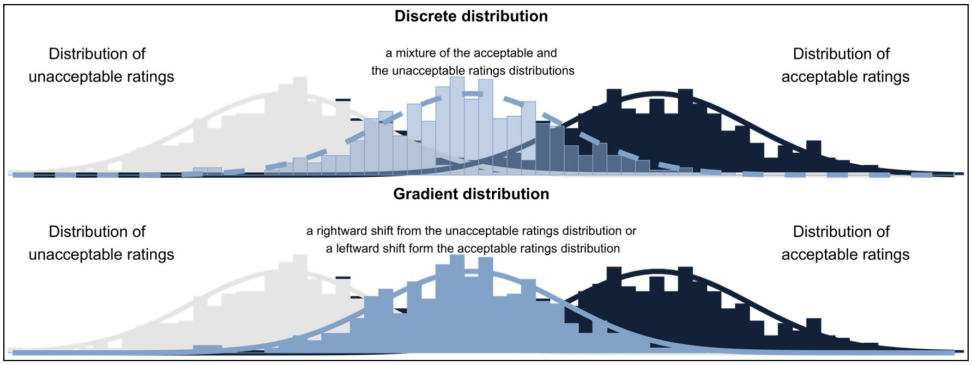


Figure 4: Two underlying distributions consistent with an intermediate ratings distribution. *Top panel:* A discrete model of the participants' judgment process, where intermediate ratings are a result of a mixture of categorical responses. *Bottom panel:* A gradient model of participants' judgment process, where intermediate ratings reflect an overall shift in acceptability. The light gray curves represent the unacceptable distribution. The dark blue curves, the acceptable distribution. The dotted light blue curve, the "illusory" intermediate ratings distribution under a discrete model. The solid light blue curve, a third distribution under a gradient model.

In short, we found that for voice-mismatch extractions involving the agent, the discrete models provided a closer fit to the empirical response distribution (i.e., the actual distribution of the acceptability ratings in experiments 1 and 2) compared to the gradient model. We take these results as evidence that in any given trial, participants either accepted or rejected agent-extractions when the verb exhibited non-AV morphology. We take these findings from our simulations, along with the descriptions made by Ceña & Nolasco (2011; 2012) and others, as evidence consistent with there being two types of Tagalog extractors: one that is consistent with the strict version of the extraction restriction in disallowing agent-extractions under NAV, and one that is consistent with a restriction that allows voice-mismatch extractions under NAV, with the proviso that the extracted argument is the agent.

5.1 Method

To determine which model best captured the response behavior of the participants, three types of distributions are needed as ingredients for the simulations: (i) a test distribution, which is the distribution of the linguistic object that we want to model; (ii) an acceptable reference distribution that gives us an estimate of the sampling space when a linguistic object is perceived as uncontroversially acceptable; and (iii) an unacceptable reference distribution that gives us an estimate of the sampling space when a linguistic object is perceived as uncontroversially unacceptable.

Recall that for our judgment studies we had 48 distractor items. The goal of these items was to compare Tagalog speakers' extraction behavior out of two syntactic islands (i.e., whether-islands and complex NP-islands). We crossed whether the verb exhibited AV or PV (VOICE: AV, PV) and whether the extraction site was the complement clause or the island (SITE: COMPLEMENT, ISLAND). This is, to our knowledge, the first experimental investigation of island effects in Tagalog. In **Figure 5**, we visualize the distribution of participant ratings by whether the extraction was out of an island. What is relevant is that extractions out of complement clauses were rated high, while extractions out of an island were rated low. For the purposes of modeling the participants' judgment process, we used the ratings for extractions out of complement clauses as the grammatical reference distribution. We then used the ratings for extractions out of islands as the ungrammatical reference distributions. See our OSF page for the supplementary material with a description and the Bayesian analyses of these experiments.

We followed closely how Dillon et al. generated the distributional predictions of these response models. See our OSF page for more information about how these distributions were generated under the two response models. As a metric for evaluating how well a given model fits the empirical distribution, we calculated the difference in the models' Bayesian Information Criterion (BIC). Negative BIC difference scores indicate an advantage for the discrete model over the gradient model, while positive BIC difference scores indicate an advantage for the gradient model over the discrete model.

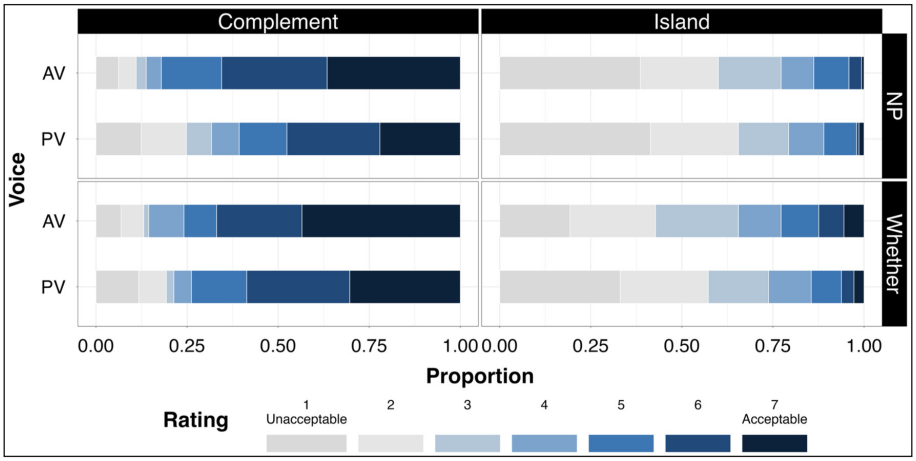


Figure 5: Distribution of ratings in experiment 2 by VOICE, SITE and island type. The left panel corresponds to extractions out of a complement; the right panel corresponds to voice-mismatch extractions. Blue corresponds to a rating of 7 'acceptable', while gray corresponds to a rating of 1 'unacceptable.'

5.2 Judgment process in experiment 1

We pooled the observations across participants within conditions. We fitted two types of discrete and gradient models for each dependency: (i) patient-extractions when the verb had AV; and (ii) agent-extractions when the verb had PV. We ran 40,000 Monte Carlo simulations of the models. Below we report the mean BIC, and in square brackets, the minimum and maximum BIC values.

Head-initial RCs. When the verb had AV, 30,918 of the 40,000 simulations had a positive BIC (Mean = 4.27 [–129.3, 38.71]). This suggests that the gradient model provided a better fit than the discrete model for the empirical distribution about 77% of the time. When the verb had PV, 40,000 of the 40,000 simulations had a negative BIC (Mean = –108.49 [–212.94,–53.97]). This suggests that the discrete model provided a better fit than the gradient model for the empirical distribution 100% of the time.

Wh-questions. When the verb had AV, 30,176 of the 40,000 simulations had a positive BIC (Mean = 4.57 [–165.6, 36.95]). This suggests that the gradient model provided a better fit than the discrete model for the empirical distribution about 75% of the time. When the verb had PV, 39,999 of the 40,000 simulations had a negative BIC (Mean = –38.95 [–90.79, 3.34]). This suggests that the discrete model provided a better fit than the gradient model for the empirical distribution a little less than 100% of the time.

Head-final RCs. When the verb had AV, 28,980 of the 40,000 simulations had a positive BIC (Mean = 4.77 [–183.32, 45.26]). This suggests that the gradient model provided a better fit than the discrete model for the empirical distribution about 72% of the time. When the verb had PV, 20,747 of the 40,000 simulations had a negative BIC (Mean = –0.5 [–183.32, 45.26]). This suggests that the discrete model provided a better fit than the gradient model for the empirical distribution about 52% of the time.

To summarize, gradient distributions best characterize the mismatch penalty in AV. This suggests that patient-extractions are better thought of as an overall decrease in acceptability from the grammatical reference distribution/an overall increase in acceptability from the ungrammatical reference distribution. Meanwhile, discrete distributions best characterize the mismatch penalty in PV—at least in head-initial RCs and *wh*-questions. This suggests that agent-extractions are better thought of as a mixture of categorical responses, sampling from both acceptable and unacceptable reference distributions.

5.3 Judgment process in experiment 2

We pooled the observations across participants within conditions. We fitted two types of discrete and gradient models for each dependency: patient-extractions and agent-extractions when the verb had applicative morphology. We ran 40,000 Monte Carlo simulations of the models. Below we report the mean BIC, and in square brackets, the minimum and maximum BIC values.

Head-initial RCs. When the patient was extracted, 31,386 of the 40,000 simulations had a positive BIC (Mean = 3.34 [–98.84, 28.97]). This suggests that the gradient model provided a better fit than the discrete model for the empirical distribution about 78% of the time. When the agent was extracted, 40,000 of the 40,000 simulations had a negative BIC (Mean = –128.83 [–245.13, –74.56]). This suggests that the discrete model provided a better fit than the gradient model for the empirical distribution 100% of the time.

Wh-questions. When the patient was extracted, 30,627 of the 40,000 simulations had a positive BIC (Mean = 4.17 [–125.48, 34.36]). This suggests that the gradient model provided a better fit than the discrete model for the empirical distribution about 77% of the time. When the agent was extracted, 40,000 of the 40,000 simulations had a negative BIC (Mean = –65.44 [–137.48, –18.74]). This suggests that the discrete model provided a better fit than the gradient model for the empirical distribution 100% of the time.

Head-final RCs. When the patient was extracted, 32,360 of the 40,000 simulations had a positive BIC (Mean = 5.95 [–171.84, 41.23]). This suggests that the gradient model provided a better fit than the discrete model for the empirical distribution about 81% of the time. When the agent was extracted, 30,757 of the 40,000 simulations had a positive BIC (Mean = 4.97 [–211.36, 40.1]). This suggests that the gradient model provided a better fit than the discrete model for the empirical distribution about 77% of the time.

To summarize, when the verb has applicative morphology, gradient distributions best characterize the mismatch penalty involving patient-extractions. This suggests that patient-extractions are better thought of as an overall decrease in acceptability from the grammatical reference distribution/an overall increase in acceptability from the ungrammatical reference distribution. Meanwhile, discrete distributions best characterize the mismatch penalty involving

agent-extractions—at least in head-initial RCs and *wh*-questions. This suggests that agent-extractions in head-initial RCs and *wh*-questions are better thought of as a mixture of categorical responses, sampling from both acceptable and unacceptable reference distributions.

5.4 Discussion

In this section, we developed quantitative models of the participants' judgment process to estimate the underlying distribution of their voice-mismatch distribution ratings across experiments 1 and 2. In **Table 4** we provide a summary of our simulations.[18] The percentages reported indicate which model provided a better fit for the empirical distribution. When it is inconclusive, it is framed in terms of the discrete model.

	Dependency	**Agent-extraction**	**Non-agent extraction**
Exp. 1	WHQ	Discrete (100%)	Gradient (77%)
	HI	Discrete (~100%)	Gradient (75%)
	HF	Inconclusive (52%)	Gradient (72%)
Exp. 2	WHQ	Discrete (100%)	Gradient (78%)
	HI	Discrete (100%)	Gradient (77%)
	HF	Gradient (77%)	Gradient (81%)

Table 4: Summary of the judgment process simulations of Experiments 1 and 2. Percentages reported show which model provided a better fit for the empirical distribution. Inconclusive is framed in terms of the discrete model.

Head-initial RCs and *wh*-questions. We found that the ratings distributions of voice-mismatch agent-extractions are better thought of as having discrete distributions. Speakers were either accepting or rejecting them in any given trial. For comparison, the ratings distributions

[18] A reviewer asked about the extent to which the results of our simulations were affected by our choice of reference distributions. Our choice did not have a profound impact on our simulations. Earlier simulations used voice-match extractions as the acceptable reference distribution, and patient-extractions in voice-mismatch contexts as the unacceptable reference distribution, and only agent-extractions in voice-mismatch contexts as the test distribution. The results are qualitatively similar: discrete models provided a better fit than gradient models for the empirical distribution of agent-extractions in voice-mismatch context. Ultimately, we chose a different set of reference distributions (extraction out of complement clauses and out of islands for acceptable and unacceptable, respectively) because our case is made more compelling if we juxtapose and show which of the quantitative models better capture the empirical distribution of agent-extractions in voice-mismatch contexts and of non-agent-extractions in voice-mismatch contexts.

of voice-mismatch patient-extractions are better thought of as having gradient distributions. Speakers were treating them uniformly, either as an overall downward shift in acceptability from the grammatical reference distribution or overall upward shift in acceptability from the ungrammatical reference distribution.

Head-final RCs. We found that voice-mismatch extractions—both agent- and patient-extractions—when the verb exhibits applicatives are better thought of as having gradient distributions. Voice-mismatch patient-extractions when the verb exhibits PV also had a gradient distribution. Speakers were treating them uniformly, either as an overall downward shift in acceptability from the grammatical distribution or overall upward shift in acceptability from the ungrammatical distribution. However, the simulations could not decide on whether the ratings distribution of voice-mismatch agent-extractions when the verb has PV is better thought of as having a gradient or discrete distribution.

One interpretation consistent with these findings is that there are two separable types of Tagalog speakers with respect to extraction: Tagalog A (those who conform to the strict version of the restriction) and Tagalog B (those who conform to the less strict version). We schematize these two types of speakers in **Figure 6**, along with a yet-to-be attested type of speaker, Tagalog C. Again, note that Tagalog A is a proper subset of Tagalog B, which in turn, is a proper subset of Tagalog C.

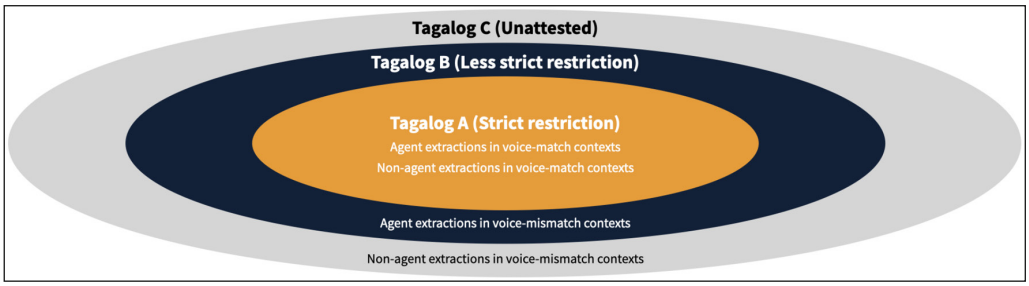


Figure 6: Visualization of the two Tagalogs with respect to extraction.

In this section, we have essentially identified a piece of grammatical knowledge (i.e., restrictions on extraction) that is systematic within an individual speaker but varies unpredictably across a population of Tagalog speakers. When these simulations are taken together with our own work with speakers and the previous descriptions about the variation in extraction restriction that Ceña & Nolasco (2011; 2012), Tanaka (2016), Hsieh (2019), Bondoc (2021) and others have observed, a hypothesis where there are two types of Tagalog extractors seems very tenable. We return to this in the general discussion.

6 General discussion and conclusion

In this section, we first summarize the results of our experiments and our simulations. We then propose a way in which we can think of how the two types of Tagalog speakers emerge despite being exposed to presumably the same or similar linguistic input. We view the two types of Tagalog speakers as a form of grammar competition that arises via the idea that the agent-first bias could affect how child learners parse input strings under noisy conditions during acquisition. We then demonstrate how agent-first is an independently-motivated principle that is at the epicenter of Tagalog grammar. We then discuss the bifurcation in our results and the implications for theories of extraction.

In our experiments, we have replicated the findings of Pizarro-Guevara & Wagers (2020) in broad strokes, and found that voice-mismatch extractions incurred greater mismatch penalties in AV than in PV. We also extended their findings and found that extracting the patient incurred greater mismatch penalties than extracting the agent when the verb had applicative morphology. Taken together, these results strongly suggest that the extraction restriction is not as clean-cut as previously described.

In our simulations, we have found that the ratings distributions of agent-extractions in voice-mismatch contexts involving head-initial RCs and *wh*-questions can better be thought of as having discrete distributions. In other words, in these dependencies, speakers were either accepting or rejecting them in any given trial. On the other hand, those involving head-final RCs can better be thought of as having gradient distributions. In other words, speakers were treating them uniformly—either as an overall downward shift in acceptability from the grammatical reference distribution or an overall upward shift in acceptability from the ungrammatical reference distribution.

6.1 Agent-first garden-paths Tagalog learners

In section 5, we interpreted the results from our simulations as being consistent with having two separable types of Tagalog speakers. A speaker of Tagalog A only allows voice-match extractions, conforming to the strict version of the extraction restriction. On the other hand, a speaker of Tagalog B allows what a speaker of Tagalog A allows, plus agent-extraction in voice-mismatch contexts. See **Figure 6** for a visualization of these two types of Tagalogs.

A natural question to ask is why two types of Tagalog extractors develop based on the same or similar linguistic input. We suggest that the type of extractor a speaker of Tagalog ultimately becomes is a form of grammar competition in learning, in the sense of Han et al. (2007), that arises via the idea that an agent-first bias could affect how child learners parse input strings

under noisy conditions. Because the core of our proposal draws heavily from language acquisition and sentence processing, we first give an overview of the intuition behind each ingredient. We will then discuss how the two types of Tagalog extractors emerge from the interaction of these ingredients.

The first ingredient is grammar competition in learning (Han et al. 2007). The main intuition behind this is that sometimes the primary linguistic data that a child in a language *L* is exposed to are compatible with at least two hypotheses, H_1 and H_2. When linguistic experience cannot adjudicate between the two, some learners acquire a grammar based on H_1 and others, based on H_2.

The second ingredient is the idea of a noisy-channel in acquisition. Production and perception errors are quite common in everyday language use. Noisy-channel models of sentence processing (e.g., Levy 2008) start with the assumption that some of the linguistic input we as comprehenders receive is noise. Thus, our task is to determine which parts of the linguistic input are considered noise and which parts are considered signal. Before children have fully acquired an adult-like grammar, they may parse some of the noise in the input as signal (Perkins et al. 2022). In other words, early in their language development, children may have a combination of signal that is correctly parsed as signal and noise that is parsed incorrectly as signal.

The third ingredient is the agent-first preference, a general bias to order the agent before any other arguments. In real-time comprehension, this manifests as a pressure to identify the most agent-like argument as quickly as possible. In other words, all else equal, assume that the first noun is the thematic agent (Ferreira 2003; Bornkessel & Schlesewsky 2006). This is an independently-motivated pressure that rears its head in many areas of the Tagalog grammar (Pizarro-Guevara & Garcia 2024). More on this in the subsection below.

Lastly, the fourth ingredient is a cognitively costly process called reanalysis (Frazier & Rayner 1982; 1987). In real-time comprehension, we make commitments to what the intended parse is even in the absence of fully disambiguating information. Sometimes, we are garden-pathed, and pursue an incorrect parse. Later on, when we are confronted with evidence against an incorrectly pursued parse, we must reanalyze and find a different analysis that is compatible with the current evidence we have. We know that children find abandoning their first analysis quite challenging, a phenomenon called the kindergarten-path effect (Trueswell et al. 1999). This has been argued to be a general cognitive limitation (Phillips & Ehrenhofer 2015).

Now let us consider how these four ingredients interact to give us the two types of Tagalog extractors. There are at least two hypotheses Tagalog children can consider with respect to what arguments can be extracted. These are provided in (22):

(22) Two hypotheses that child learners of Tagalog entertain vis-à-vis extraction[19]
   a. H_1 ultimately leads to being a speaker of Tagalog A, where only voice-match extractions are allowed
   b. H_2 ultimately leads to being a speaker of Tagalog B, where agent-extractions in voice-mismatch contexts are allowed *on top of* what Tagalog A allows.

When the extracted argument is cross-referenced by voice, the linguistic output of these two hypotheses converge. Their output diverges only in contexts where the extracted argument is not cross-referenced by voice. However, these types of voice-mismatch extractions are rare in the input. For example, in an ongoing corpus analysis of Tagalog conversations, based on 105 minutes of recording and 22,663 words, Nagaya (2019) found 0 instances of voice-mismatch extractions. We do know that they exist, as we have seen examples of voice-mismatches "in the wild" in Section 2. We also know that child and adult speakers produce them in elicited production experiments (Tanaka 2016). With little or no evidence favoring one hypothesis over the other, different child learners may adopt different hypotheses. Some child learners might adopt H_1, ultimately leading to their becoming Tagalog A speakers. Some child learners might adopt H_2, ultimately leading to their becoming Tagalog B speakers. This is grammar competition in action.

Why would H_2 even be a part of the hypothesis space? Framed differently, why would some child learners give agents privileged status? We maintain that the agent-first pressure biases how comprehenders parse input strings under noisy conditions during acquisition. First consider how child learners arrive at the generalization that extractions need to, at the very least, be voice-matched. Recall that this corner of the extraction restriction is shared by both types of extractors. By hypothesis, when they were child learners, they must have correctly parsed the signal as signal. Perhaps this is unsurprising since this is presumably the type of input that they had strong positive evidence for. We argue that correctly parsing signal as signal is shared by learners who eventually become speakers of both Tagalog A *and* B.

How might child learners add the extra layer needed to allow agent-extractions in voice-mismatch contexts? Under a noisy channel assumption in acquisition, child learners also incorrectly parse noise in the input as signal. Our main claim is that the misparsing of noise to signal is caused by the agent-first bias. Consider child learners that encounter what would eventually be a head-initial RC, where the head noun functions as the patient of the verb inside

[19] For reasons of space, we only consider these two hypotheses. It is a logical possibility, for example, to only entertain a hypothesis where only voice-mismatch extractions are allowed. We believe that this is a very implausible hypothesis since that would mean that child learners are ignoring strong positive evidence (i.e., the presence of voice-match extractions) in their input to entertain only these.

the RC, as in (23a). This is the linguistic input that they are confronted with. Upon encountering the head noun *katulong* and the linker *na*, they may have realized that they are likely in a head-initial RC configuration. At this point in time, the head noun *katulong* is still role-ambiguous. They might be biased to interpret this as the agent of the upcoming predicate in the RC, per agent-first.

(23) a. Maldita ang katulong na sinipa ng bata
naughty NOM maid LNK kicked.PV GEN child
'The maid that the child kicked is naughty' *Input*

b. … katulong na sinipa **ng** bata
maid LNK kicked.PV GEN child
'… maid that the child kicked…' *Veridical representation of input*

Upon encountering the RC verb *sinipa*, they might realize that their initial parse is inconsistent with the input, assuming voice-matching, and thus, it needs revision. In (23b), they are able to recover from their incorrect parse and reanalyze the role of the head noun. Thus, they are able to develop a veridical representation of the input. From these types of tokens, over time, child learners learn that when the verb has PV, patients are extractable (i.e., a voice-match extraction). We argue that successful reanalysis is shared by child learners who eventually become speakers of Tagalog A *and* B.

There is evidence in the RC-processing literature that is consistent with the idea that comprehenders temporarily misparse the head-noun as the agent, per agent-first, but successfully recover from that initial parse upon encountering more information (i.e., the verb and the disambiguating co-argument). Consider two minimally different examples from Pizarro-Guevara (2020) in (24); these are string-identical, save the verb. Many researchers have found that RCs involving a patient head-noun, as in (24b), are more difficult to process than RCs involving an agent head-noun, as in (24a) (Pizarro-Guevara 2014; Tanaka 2016; Bondoc et al. 2018; Pizarro-Guevara 2020). We can interpret this difficulty—at least, in part—as reanalysis cost. In (24b), comprehenders posited that the head noun is the agent, per agent-first. Upon encountering the verb (and coargument), their initial parse was confirmed. On the other hand, in (24b), comprehenders posited that the head noun is the agent, per agent-first. Upon encountering the verb (and coargument), they realized that their initial parse was incorrect and had to revise.

(24) RCs with patient head nouns are harder than RCs with agent head nouns

a. baboy na sumisipa ng kambing
pig LNK kicking.AV GEN goat
'pig that is kicking the goat' *RC with agent head-noun*

b. baboy na sinisipa ng kambing
pig LNK kicking.PV GEN goat
'pig that the goat is kicking' *RC with patient head-noun*

Opposite successful reanalysis, an alternative scenario involves children's failure to reanalyze their initial parse. Sometimes, even after encountering the verb, child learners might be unable to let go of their initial parse—a common feature of child sentence processing that many researchers have documented before (Trueswell et al. 1999; Phillips & Ehrenhofer 2015). Consider again the linguistic input that child learners are confronted with in (23a), repeated here as (25a). Upon encountering *ng bata*, they may try to revise the parse again, given that they now have more evidence in the input. Or they may try to save their initial parse by assuming that their interlocutor had intended to say *ang bata* [ʔaŋ.ba.taʔ]—instead of *ng bata* [naŋ.ba.taʔ]. The distortion of the input seems plausible enough since *ang* and *ng* are phonetically very similar; the only difference is the onset. One advantage that this distortion affords the child learners is that they get to form a locally coherent parse with the string *sinipa ang bata* and crucially, the locally coherent parse—even though globally ungrammatical from the perspective of a strict version of the restriction—allows them to maintain their initial parse, where *katulong* is the agent. After accumulating enough of these misparsed noise, child learners might transform these tokens into bona fide signal. Over time, these child learners might learn that when the verb has PV, agents are also extractable (i.e., a voice-mismatch extraction). We argue that failing to reanalyze *and then* reclassifying noise as signal is a characteristic of child learners who eventually become speakers of Tagalog B.

(25) a. Maldita ang katulong na sinipa ng bata
naughty NOM maid LNK kicked.PV GEN child
'The maid that the child kicked is naughty' *Input*

b. … katulong na sinipa ang bata
maid LNK kicked.PV NOM child
'…maid that kicked the child…' *Misparsed representation of input*

There is evidence in the RC-processing literature that is consistent with the idea that children initially parse head-nouns as agents and that they sometimes fail to abandon their initial parse (i.e., failing to reanalyze despite more evidence in the input). Thematic reversals are the most common errors children make. In particular, they have a bias to interpret the head noun as the agent of the predicate inside the RC. For example, Tanaka et al. (2019) found that 5 to 6 year olds interpreted an RC where the head noun was the patient as an RC where the head noun was the agent three times more often than the other way around. In other words, when the head noun was a patient, children often interpreted this as an agent. Misinterpreting an agent head noun was significantly less common. These results are corroborated by an earlier study by Pizarro-Guevara (2014): the rate of thematic reversals 4 to 9 year olds made was higher on average when it was patient-to-agent (i.e., a patient head noun being misinterpreted as an agent head noun) than when it was agent-to-patient (i.e., an agent head noun being misinterpreted as a patient head noun). Furthermore, apart from the naturalistic data presented in section 2.1, there

is evidence that speakers—both children and adults—do produce structures like (25b), as shown by Tanaka (2016) in RC-production studies. The production of such structures is consistent with the idea that learners may have failed to reanalyze, and that they may have reclassified noise as signal.

To summarize, we argued that the language development of a speaker of Tagalog A is characterized by the processes in (26). On the other hand, the language development of a speaker of Tagalog B is characterized by the processes in (27).

(26) Speakers of Tagalog A, which conforms to the strict version of the restriction
  a. Correct parsing of signal as signal
  b. Temporary misparsing of noise as signal, followed by successful reanalysis

(27) Speakers of Tagalog B, which conforms to the less strict version of the restriction
  a. Correct parsing of signal as signal
  b. Temporary misparsing of noise as signal, followed by successful reanalysis
  c. Misparsing of noise as signal, followed by reclassification of noise as signal

A natural question to ask now is how the proposal developed above can account for the participants' behavior in head-final RCs. Recall that voice-mismatches are *not* allowed—even for those who do allow them in head-initial RCs. In head-initial RCs, schematized below in (28a), upon encountering the first noun, this noun is still role-ambiguous. Only upon encountering the verb do comprehenders know its role (assuming voice-matching), as discussed above. On the other hand, in head-final RCs, schematized in (28b), upon encountering the first noun, this noun is no longer role-ambiguous. Comprehenders will have already encountered the verb with voice morphology and the case marker of the noun, which they can then use to infer its role. We claim that this lack of temporal role-ambiguity is the reason why the effect of agent-first is attenuated.

(28) Schematization of head-initial and head-final RCs in Tagalog
  a. Noun [$_{RC}$ Verb Case-Noun]
  b. [$_{RC}$ Verb Case-Noun] Noun

There is evidence in the literature consistent with the idea that the garden-path effect caused by the agent-first bias is attenuated in head-final RCs, compared to its effect in head-initial RCs. For example, Pizarro-Guevara (2020) compared how head-initial and head-final RCs were processed, using items like (29). He found that in head-initial RCs, as in (29a), RCs with a patient head-noun were more difficult to process than those with an agent head-noun. However, in head-final RCs, as in (29b), there was no evidence that one was more difficult than the other.

(29) RCs with patient head nouns are harder than RCs with agent head nouns

a. baboy na sumisipa/sinisipa ng kambing
pig LNK kicking.AV/PV GEN goat
'pig that is kicking the goat (with AV)' *head-initial RC with agent head-noun*
'pig that the goat is kicking (with PV)' *head-initial RC with patient head-noun*

b. sumisipa/sinisipa ng kambing na baboy
kicking.AV/PV GEN goat LNK pig
'pig that is kicking the goat (with AV)' *head-final RC with agent head-noun*
'pig that the goat is kicking (with PV)' *head-final RC with patient head-noun*

In this section, we presented a way for us to think about how the two types of Tagalog extractors came to be by drawing insights from language acquisition and sentence processing. We placed the locus of variation on the agent-first bias. This may seem like an entirely ad hoc proposal. However, as we will see in the next subsection, the agent-first bias has a far-reaching influence in the grammar of Tagalog.

6.2 Agent-first is at the epicenter of Tagalog grammar

The agent-first bias has been argued to be rooted in an evolutionary bias to attend to agentive features and is central in general event perception and cognition (Sauppe et al. 2023). In the proposal above, it was at the epicenter. Here, we exemplify how it is an independently-motivated pressure that seems to have a far-reaching influence in Tagalog grammar (Pizarro-Guevara & Garcia 2024).

First, it influences the post-verbal word order in declarative sentences, along with two other pressures: the *ang*-last and heavy NP-shift, the tendency to have "heavier" elements follow "lighter" elements (Kroeger 1993). For example, when the verb has PV, the verb-agent-patient order, as in (30a), is reported to be more natural than the verb-patient-agent order, as in (30b). The former satisfies both agent-first and *ang*-last, holding heaviness constant, while the latter violates both. On the other hand, when the verb has AV, both verb-agent-patient, as in (30c), and verb-patient-agent, as in (30d), have been reported to be equally natural. This is corroborated by experimental evidence from judgment studies (Hsieh 2016), sentence completion (Bondoc & Schafer 2022), and the visual world paradigm (Sauppe 2016).

(30) Examples adapted from Hsieh (2016)

a. Pinatay ng balyena ang pating
killed.PV GEN whale NOM shark
'The whale killed the shark' *Verb-Agent-Patient*

b. Pinatay ang pating ng balyena
killed.PV NOM shark GEN whale
'The whale killed a shark' *Verb-Patient-Agent*

c. Pumatay ang balyena ng pating
killed.AV NOM whale GEN shark
'The whale killed a shark' *Verb-Agent-Patient*

d. Pumatay ng pating ang balyena
killed.AV GEN shark NOM whale
'The whale killed a shark' *Verb-Patient-Agent*

Second, agent-first also plays a role in Tagalog speakers' interpretation of globally ambiguous RCs. Using a picture selection task, Pizarro-Guevara (2020) showed that when participants were presented with a globally ambiguous head-initial RC, as in (31a), they interpreted the head noun as the patient 30% of the time, and as the agent 70% of the time. In other words, there was a strong preference to interpret the head noun of a globally ambiguous head-initial RC as the agent. By contrast, when they were presented with a globally ambiguous head-final RC, as in (31b), they interpreted the head noun as the patient 50% of the time, and as the agent 50% of the time. Unlike in head-initial RCs, there was no agent-first bias in how they interpreted the head noun of a globally ambiguous head-final RC.

(31) RCs with verbs in the recent perfective can be globally ambiguous

a. baboy na kakasipa lang ng kambing
pig that kicked.RP just GEN goat
'pig that just kicked the goat'
'pig that the goat just kicked' *Globally ambiguous head-initial RC*

b. kakasipa lang ng kambing na baboy
kicked.RP just GEN goat that pig
'pig that just kicked the goat'
'pig that the goat just kicked' *Globally ambiguous head-final RC*

Finally, from a developmental perspective, the centrality of agent-first in Tagalog is also evident in the type of errors that children make when comprehending different types of sentences. In the previous subsection, we saw that children made thematic reversal errors. They were more likely to misinterpret a patient head-noun as the agent than misinterpret an agent head-noun as the patient (Pizarro-Guevara 2014; Tanaka 2016). Similar reversal errors are also found when comprehending verb-initial sentences with two arguments in post-verbal position, as indicated by their accuracies in picture-verification (Garcia et al. 2019) and in picture-selection tasks (Garcia et al. 2020).

Lastly, the agent-first bias is not unique to Tagalog. In fact, it can be observed cross-linguistically in both child and adult languages: Äiwoo (Sauppe et al. 2023), Basque (Erdocia et al. 2009), English (Ferreira 2003), German (Haupt et al. 2008), Hindi (Bickel et al. 2015), Mandarin (Huang et al. 2013), Ojibwe (Hammerly et al. 2022), Spanish (Gattei et al. 2015), and Turkish (Demiral et al. 2008), to name a few. These studies suggest that agent-first is a cross-linguistically robust pressure, and thus, it should be unsurprising for it to also influence/be leveraged by Tagalog learners and speakers.

6.3 A bifurcation in the results

Irrespective of dependency type, we found evidence for a mismatch penalty when (i) extracting an agent not cross-referenced by voice, and (ii) when extracting a patient not cross-referenced by voice. Where the dependencies differed is when we consider the difference in mismatch penalties. That is, the difference between (i) and (ii) above. In other words, there is a bifurcation in their difference of differences. We see that in head-initial RCs and for the most part, *wh*-questions, the mismatch penalty in agent-extractions is smaller than in patient-extractions. Meanwhile, in head-final RCs, the mismatch penalties in agent-extractions and patient-extractions are not different. We also see this bifurcation in our simulations: head-initial RCs and *wh*-questions pattern in one way, while head-final RCs pattern in another way.

A natural question to ask is why head-initial RCs and *wh*-questions would form a natural class and exclude head-final RCs. We offer two potential reasons—and they need not be mutually exclusive. The first reason is structural similarity. It is well-established in the literature that argument wh-questions in Tagalog are derived via a (pseudo-)clefting strategy, which means that the interrogative phrase serves as the predicate (i.e., as a predicate nominal) and the remaining material is a headless relative clause (Kroeger 1993; Aldridge 2002). Their similarity is schematized in (32):

(32) Similarity between head-initial RCs and argument *wh*-questions in Tagalog
   a. Predicate [$_{NP}$ Head$_i$ [$_{RC}$ Predicate $\_\_{}_i$ …]]
   b. *Wh*-phrase$_i$ [$_{NP}$ $\varnothing_i$ [$_{RC}$ Predicate $\_\_{}_i$ …]]

As briefly mentioned in section 1.2, it is an open question whether head-initial and head-final RCs have different derivations (Aldridge 2017) or they share a single one (Law 2016). If Aldridge is on the right track, head-final RCs are derived differently from head-initial RCs (and by extension, argument *wh*-questions). Thus, head-initial RCs and *wh*-questions can form a natural class and exclude head-final RCs.

The second reason is processing similarity. As discussed in section 6.1, in head-initial RCs and *wh*-questions, Tagalog comprehenders are first confronted with the filler (i.e.,

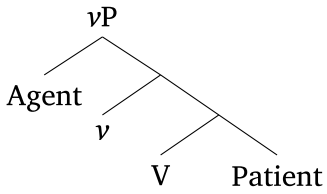
the argument displaced from the position where it is interpreted), followed by the gap (i.e., the position where it is interpreted). In a sense, comprehenders engage in a forward search for where to interpret the displaced argument. In contrast, in head-final RCs, Tagalog comprehenders are confronted with the gap first, followed by the filler. In a sense, comprehenders engage in a backward search for where to interpret the filler. Viewed this way, we have another reason as to why head-initial RCs and *wh*-questions can form a natural class and exclude head-final RCs.

In the proposal above, we have alluded indirectly to the processing similarity in head-initial RCs and argument *wh*-questions. In these dependencies, they are confronted with the filler first, which accentuates the garden-path effect caused by the agent-first bias.

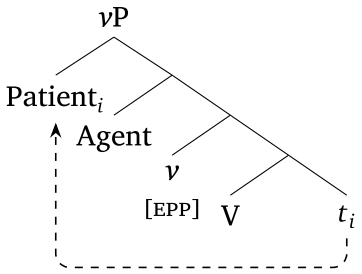
6.4 Implications for theories of extraction

Before we conclude, we briefly discuss the implications of our current findings for theories of A-bar extraction in Austronesian languages. In many of the formal accounts in the literature (Aldridge 2002; Rackowski & Richards 2005; Aldridge 2012), the argument cross-referenced by voice is syntactically privileged because it is the highest DP in the structure. We use "highest DP approach" as an umbrella term for this class of proposals. Even though the finer details of each proposal are different, they have the following features in common. First, AV and non-AV voices are structurally different, and we provided partial derivations in (11) and in (12), repeated here as (33) and (34), for AV and for PV (which we take as representative of non-AV voices), respectively. In PV, the patient vacates the VP and moves to the outermost specifier of *v*P, above the agent.

(33) Partial derivation of AV

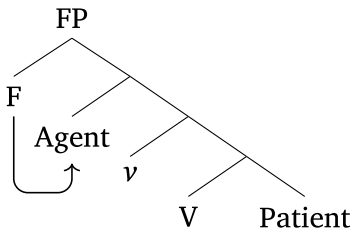


(34) Partial derivation of PV

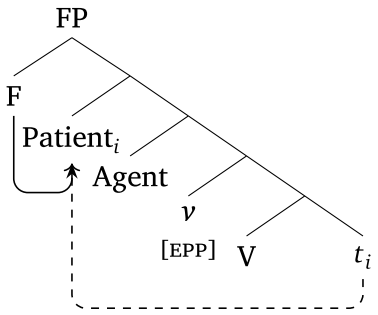


Second, the highest DP receives *ang*-marking. This captures the generalization that when the verb has AV, the agent is *ang*-marked, and when the verb has PV, the patient is. Third (and perhaps most importantly), voice-match extraction is accounted for using Attract Closest. The basic idea is that the highest DP is the closest goal for an A-bar probe. With this formal machinery in place, the strict version of the restriction is a natural consequence. If an A-bar probe like F were to search for a goal, as in (35), only the agent is visible to the probe since it is the highest DP in the structure and is thus, the closest goal. This effectively allows agent-extractions under AV (+MATCH) and rules out patient-extractions under AV (–MATCH). If an A-bar probe like F were to search for a goal, as in (36), the patient, which has moved to a position higher than the agent, has become the highest DP and has become the closest goal. This effectively allows patient-extractions under PV (+MATCH) and rules out agent-extraction under PV (–MATCH).

(35) Partial derivation of AV



(36) Partial derivation of PV



The highest DP approach, in its current form, undergenerates. It cannot generate agent-extractions under non-AV voices for those who do allow them. The present study has demonstrated that these speakers do exist. There are (at least) two ways in which this approach can be modified to allow these structures to be generated: (i) modify the calculus of how the distance between the goal and the probe is determined; or (ii) posit another operation that places the agent in a structurally more prominent position after the patient has vacated the VP. In a recent proposal that takes into account these voice-mismatch structures, Hsieh (2023) makes a case for (ii) and argues that agents that are not cross-referenced by voice undergo what he calls genitive inversion, an independently attested operation in the language, which places the agent in a structurally more prominent position. This operation, in turn, feeds A-bar extraction.

A more thorough consideration of Hsieh's proposal is outside the immediate scope of this paper. A main feature of his system is how intricately tied agent-extraction under NAV is to what he calls genitive inversion. Thus, his system would predict that the extent to which speakers allow agent-extraction under non-AV voices is tightly correlated with the extent to which they allow genitive inversion in their Tagalog. We leave it to future work to investigate the extent to which this prediction obtains.

There are also proposals in the literature, which we subsume under the umbrella term "in situ approach," where the argument cross-referenced by voice need not be the highest DP in the structure (Chen 2017; Aldridge 2018). The finer details of each proposal differ, but they share some core features. First, of particular relevance is that in NAV, the patient does not occupy a structural position higher than the agent. The patient remains in-situ, irrespective of voice. Second, whatever proposal-specific mechanism is employed to determine voice is the same mechanism used to account for voice-mismatch extractions. Framed differently, whatever argument is agreed with by the probe that determines voice in Aldridge's system—or probes in Chen's system—is the same argument that satisfies the feature responsible for extraction. In AV, this argument is the agent and thus, the agent is extractable. In PV, this argument is the patient and thus, the patient is extractable. One way to modify the in situ approach is to decouple the mechanisms responsible for voice determination and for implementing the extraction restriction. They need not be a single process. That is, they could be separate but concurrent processes instead. It is outside the immediate scope of this paper to provide a detailed treatment of these proposals and to fully explore the theoretical ramifications of decoupling the two mechanisms.

6.5 Conclusion

To summarize, we re-examined the extraction restriction in Tagalog. We situated our investigation in the context of the received view of the restriction and provided an alternative generalization. We provided naturally occurring data of voice-mismatch extractions. Second, we replicated the findings of Pizarro-Guevara & Wagers (2020) in broad strokes, and found that voice-mismatch extractions incurred greater mismatch penalties in AV than in PV. Third, we extended their findings and found that extracting the patient incurred greater mismatch penalties than extracting the agent when the verb had applicative morphology. Together, these results strongly suggest that the restriction is not as clear-cut as previously described. At the very least, the generalizations we present should figure in any theory of how voice interacts with A-bar extraction in the language.

We also developed quantitative models that estimated the participants' judgment process and the results suggest that participants were treating agent-extractions under PV and applicatives categorically. This is a novel application of the type of computational modeling that Dillon et

al. (2017) performed for an unrelated phenomenon in an unrelated language (i.e., agreement attraction in English). Our results can thus be viewed as further proof of concept for that method and how a distributional analysis of ratings can provide evidence of multiple types of speakers in a language community.

We identified a piece of grammatical knowledge (i.e., extraction) that is systematic within an individual speaker but varies unpredictably across a population of Tagalog speakers. We proposed that this is a form of grammar competition that arises via the idea that the agent-first bias, an independently motivated pressure in Tagalog that also has cross-linguistic basis, is at the epicenter of this variation. We advanced the claim that it affects how child learners parse input strings under noisy conditions during acquisition.

If we take seriously the claim that there are two types of Tagalog extractors, we should expect to find some evidence of their emergence in how children judge voice-mismatch extractions. There are studies that examine how children judge sentences involving extraction, mainly involving head-initial RCs (Pizarro-Guevara 2014; Bondoc et al. 2018; Tanaka et al. 2019). Future work can investigate how children judge sentences when the extracted argument is not cross-referenced by voice. A comparison of head-initial and head-final RCs would be particularly fruitful. Future work can also systematically investigate if there are any sociolinguistic correlates of this point of variation or whether this truly is stochastic in nature.

Abbreviations

APPL = applicative, AV = agent voice, AY = *ay*-inversion marker/*ay*-inverted sentence, BF = Bayes factor, BIC = Bayesian Information Criterion, CrI = credible interval, DAT = dative, DEM = demonstrative, GEN = genitive, HF = head-final relative clause, HI = head-initial relative clause, LNK = linker, NAV = non-AV voice, NOM = nominative, PL = plural, PN = proper name, PV = patient voice, RC = relative clause, RP = recent perfective aspect, SG = singular, WHQ = argument *wh*-question

Data availability

The experimental items, anonymized data, the visualization and analysis scripts, and other supplementary files associated with this article are openly available in Open Science Framework: https://osf.io/me62k/.

Ethics and consent

This study was granted exemption by the University of California Santa Cruz's Office of Research Compliance Administration (UCSC IRB Protocol # 3255: Extraction asymmetry in Tagalog).

Funding information

This work was supported in part by NSF BCS #1251429 to MW and Sandra Chung, by NSF BCS #1941485 to Brian Dillon, and by NSF SPRF #2204112 to JPG and Brian Dillon.

Acknowledgements

We thank Henrison Hsieh, Kristina Gallego, Soleil Davíd, Rowena Garcia, and other speakers of Tagalog for their help with the facts about the language; and the Linguistics Department at the University of the Philippines Diliman (especially to Jurekah Chene Abrigo and Farah Cunanan) for their help with participant recruitment. We also thank Sandy Chung, Brian Dillon, our three anonymous reviewers, and the editor Lyn Tieu for their questions and comments.

Competing interests

The authors have no competing interests to declare.

Author contributions

Conceptualization: JPG (lead) & MW (supporting); *Data curation:* JPG; *Formal analyses:* JPG; *Funding acquisition:* JPG & MW; *Investigation:* JPG; *Methodology:* JPG (lead) & MW (supporting); *Project administration:* JPG (lead) & MW (supporting); *Software:* JPG; *Visualization:* JPG; *Writing (original draft):* JPG; *Writing (review & editing):* JPG (lead) & MW (supporting).

References

Aldridge, Edith. 2002. Nominalization and *wh*-movement in Seediq and Tagalog. *Language and Linguistics* 3(2). 393–426.

Aldridge, Edith. 2004. *Ergativity and word order in Austronesian languages*. Ithaca, NY: Cornell University dissertation.

Aldridge, Edith. 2008. Phase-based account of extraction in Indonesian. *Lingua* 118(10). 1440–69. DOI: https://doi.org/10.1016/j.lingua.2007.08.006

Aldridge, Edith. 2012. Antipassive and ergativity in Tagalog. *Lingua* 122(3). 192–203. DOI: https://doi.org/10.1016/j.lingua.2011.10.012

Aldridge, Edith. 2017. Internally and externally headed relative clauses in Tagalog. *Glossa: A Journal of General Linguistics* 2(1). 1–33. DOI: https://doi.org/10.5334/gjgl.175

Aldridge, Edith. 2018. ϕ-feature competition: A unified approach to the Austronesian extraction restriction. In *The Proceedings of the 52nd Meeting of the Chicago Linguistics Society*. Chicago, IL: Chicago Linguistics Society.

Bickel, Balthasar & Witzlack-Makarevich, Alena & Choudhary, Kamal K. & Schlesewsky, Matthias & Bornkessel-Schlesewsky, Ina. 2015. The neurophysiology of language processing shapes the evolution of grammar: Evidence from case marking. *PLOS ONE* 10(8). e0132819. DOI: https://doi.org/10.1371/journal.pone.0132819

Bondoc, Ivan Paul. 2021. Relativization asymmetries in Philippine-type languages: A preliminary investigation. *The Archive* 1(1–2). 1–34.

Bondoc, Ivan Paul & O'Grady, William & Deen, Kamil & Tanaka, Nozomi. 2018. Agrammatism in tagalog: voice and relativisation. *Aphasiology* 32(5). 598–617. DOI: https://doi.org/10.1080/02687038.2017.1366417

Bondoc, Ivan Paul & Schafer, Amy. 2022. Differential effects of agency, animacy, and syntactic prominence on production and comprehension: Evidence from a verb-initial language. *Canadian Journal of Experimental Psychology* 76(4). 302–26. DOI: https://doi.org/10.1037/cep0000280

Bornkessel, Ina & Schlesewsky, Matthias. 2006. The extended argument dependency model: A neurocognitive approach to sentence comprehension across languages. *Psychological Review* 113(4). 787–821. DOI: https://doi.org/10.1037/0033-295X.113.4.787

Bürkner, Paul-Christian. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80(1). 1–80. DOI: https://doi.org/10.18637/jss.v080.i01

Ceña, Resty M. 1979. Tagalog counterexamples to the Accessibility Hierarchy. *Studies in Philippine Linguistics* 3(1). 119–24.

Ceña, Resty M. & Nolasco, Ricardo Maria Duran. 2011. *Gramatikang Filipino: Balangkasan*. Quezon City, Manila: The University of the Philippines Press.

Ceña, Resty M. & Nolasco, Ricardo Maria Duran. 2012. *Sintaks ng Filipino*. Manila, Philippines: The National Commission for Culture and the Arts.

Chen, Victoria. 2017. *A re-examination of the Philippine-type voice system and its implications for Austronesian primary-level subgrouping*. Honolulu, HI: University of Hawai'i at Manoa dissertation.

Chen, Victoria & McDonnell, Bradley. 2019. Western Austronesian voice. *Annual Review of Linguistics* 5. 173–95. DOI: https://doi.org/10.1146/annurev-linguistics-011718-011731

Demiral, Sükrü Bari Ş. & Schlesewsky, Matthias & Bornkessel-Schlesewsky, Ina. 2008. On the universality of language comprehension strategies: evidence from Turkish. *Cognition* 106(1). 484–500. DOI: https://doi.org/10.1016/j.cognition.2007.01.008

Dillon, Brian & Staub, Adrian & Levy, Joshua & Clifton, Charles. 2017. Which noun phrases is the verb supposed to agree with?: Object agreement in American English. *Language* 93(1). 65–96. DOI: https://doi.org/10.1353/lan.2017.0003

Drummond, Alex. 2016. Ibex farm: Internet Based EXperiments (Version 0.3.9) [Computer program]. http://spellout.net/ibexfarm/.

Erdocia, Kepa & Laka, Itziar & Mestres-Missé, Anna & Rodriguez-Fornells, Antoni. 2009. Syntactic complexity and ambiguity resolution in a free word order language: Behavioral and electrophysiological evidences from Basque. *Brain and Language* 109(1). 1–17. DOI: https://doi.org/10.1016/j.bandl.2008.12.003

Erlewine, Michael Yoshitaka & Levin, Theodore & van Urk, Coppe. 2017. Ergativity and Austronesian-type voice systems. In Coon, Jessica & Massam, Diane & deMena Travis, Lisa (eds.), *Oxford handbook of ergativity*, 373–96. Oxford, UK: Oxford University Press. DOI: https://doi.org/10.1093/oxfordhb/9780198739371.013.16

Ferreira, Fernanda. 2003. The misinterpretation of noncanonical sentences. *Cognitive Psychology* 47. 164–203. DOI: https://doi.org/10.1016/S0010-0285(03)00005-7

Foley, William A. 2008. The place of Philippine languages in a typology of voice systems. In *Voice and grammatical relations in Austronesian languages*, 22–44. Stanford, CA: CSLI Publications.

Foley, William & Van Valin, Jr., Robert D. 1984. *Functional syntax and universal grammar*. Cambridge, UK: Cambridge University Press.

Frazier, Lyn & Rayner, Keith. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology* 14(2). DOI: https://doi.org/10.1016/0010-0285(82)90008-1

Frazier, Lyn & Rayner, Keith. 1987. Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of Memory and Language* 26(5). DOI: https://doi.org/10.1016/0749-596X(87)90137-9

Galang, Rosita. 1988. The language situation of Filipino Americans. In McKay, Sandra Lee & Wong, Sau-ling Cynthia (eds.), *Language diversity, problem or resource?*, 229–51. Cambridge, MA: Newbury House Publishers.

Galang, Rosita. 2001. Language of instruction in the Philippines in the 20th century: policies, orientations, and future directions. In Brainard, Cecilia Manguerra & Litton, Edmundo F. (eds.), *Journey of one hundred years: Reflections on the centennial of Philippine independence*, 97–117. Santa Monica, CA: Philippine American Women Writers and Artists.

Garcia, Rowena & Roeser, Jens & Höhle, Barbara. 2019. Thematic role assignment in the L1 acquisition of Tagalog: Use of word order and morphosyntactic markers. *Language Acquisition* 26(3). 235–61. DOI: https://doi.org/10.1080/10489223.2018.1525613

Garcia, Rowena & Roeser, Jens & Höhle, Barbara. 2020. Children's online use of word order and morphosyntactic markers in Tagalog thematic role assignment: an eye-tracking study. *Journal of Child Language* 47(3). 533–55. DOI: https://doi.org/10.1017/S0305000919000618

Gattei, Carolina A. & Dickey, Michael W. & Wainselboim, Alejandro J. & París, Luis. 2015. The thematic hierarchy in sentence comprehension: A study on the interaction between verb class and word order in Spanish. *Quarterly Journal of Experimental Psychology* 68(10). 1981–2007. DOI: https://doi.org/10.1080/17470218.2014.1000345

Gelman, Andrew & Lee, Daniel & Guo, Jiqiang. 2015. Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics* 40(5). 530–43. DOI: https://doi.org/10.3102/1076998615606113

Hammerly, Christopher & Staub, Adrian & Dillon, Brian. 2022. Person-based prominence guides incremental interpretation: Evidence from obviation in Ojibwe. *Cognition* 225. 105122. DOI: https://doi.org/10.1016/j.cognition.2022.105122

Han, Chung-Hye & Lidz, Jeffrey & Musolino, Julien. 2007. V-raising and grammar competition in Korean: Evidence from negation and quantifier scope. *Linguistic Inquiry* 38(1). 1–47. DOI: https://doi.org/10.1162/ling.2007.38.1.1

Haupt, Friederike S. & Schlesewsky, Matthias & Roehm, Dietmar & Friederici, Angela D. & Bornkessel-Schlesewsky, Ina. 2008. The status of subject–object reanalyses in the language comprehension architecture. *Journal of Memory and Language* 59(1). 54–96. DOI: https://doi.org/10.1016/j.jml.2008.02.003

Hsieh, Henrison. 2016. Prosodic indicators of phrase structure in Tagalog transitive sentences. In Nomoto, Hiroki & Miyauchi, Takuya & Shiohara, Asako (eds.), *Proceedings of the 23rd Annual Meeting of Austronesian Formal Linguistics Association*, 111–22. Canberra: Asia-Pacific Linguistics.

Hsieh, Henrison. 2018. *Wh*-relative clauses in tagalog. Paper presented at *the 28th Annual Meeting of the Southeast Asian Linguistics Society*.

Hsieh, Henrison. 2019. Distinguishing nouns and verbs: A Tagalog case study. *Natural Language and Linguistic Theory* 37(2). 523–69. DOI: https://doi.org/10.1007/s11049-018-9422-3

Hsieh, Henrison. 2023. Locality in exceptional Tagalog Ā-extraction. *Linguistic Inquiry*. DOI: https://doi.org/10.1162/ling_a_00505

Huang, Yi Ting & Zheng, Xiaobei & Meng, Xiangzhi & Snedeker, Jesse. 2013. Children's assignment of grammatical roles in the online processing of Mandarin passive sentences. *Journal of Memory and Language* 69(4). 589–606. DOI: https://doi.org/10.1016/j.jml.2013.08.002

Jeffreys, Harold. 1998. *The theory of probability*. Oxford: Oxford University Publishing. DOI: https://doi.org/10.1093/oso/9780198503682.001.0001

Kaufman, Daniel. 2009. Austronesian nominalism and its consequences: A Tagalog case study. *Theoretical Linguistics* 35(1). 1–49. DOI: https://doi.org/10.1515/THLI.2009.001

Keenan, Edward & Comrie, Bernard. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry* 8(1). 63–99.

Kroeger, Paul. 1993. *Phrase structure and grammatical relations in Tagalog*. Stanford, CA: Center for the Study of Language and Information (CSLI).

Latrouite, Anja. 2011. *Voice and case in Tagalog: The coding of prominence and orientation*: Heinrich-Heine-Universität Düsseldorf dissertation.

Law, Paul. 2016. The syntax of Tagalog relative clauses. *Linguistics* 54(4). 717–68. DOI: https://doi.org/10.1515/ling-2016-0016

Levy, Roger. 2008. A noisy-channel model of human sentence comprehension under uncertain input. In Lapata, Mirella & Ng, Hwee Tou (eds.), *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 234–243. Honolulu, Hawaii: Association for Computational Linguistics. https://aclanthology.org/D08-1025. DOI: https://doi.org/10.3115/1613715.1613749

Nagaya, Naonori. 2019. Relativization in Tagalog conversation: A typological perspective. Paper presented at *13th Conference of the Association for Linguistic Typology*.

Nicenboim, Bruno & Vasishth, Shravan & Rösler, Frank. 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* 142. 107427. DOI: https://doi.org/10.1016/j.neuropsychologia.2020.107427

Perkins, Laurel & Feldman, Naomi H. & Lidz, Jeffrey. 2022. The power of ignoring: filtering input for argument structure acquisition. *Cognitive Science* 46. e13080. DOI: https://doi.org/10.1111/cogs.13080

Philippine Statistics Authority. 2021. *Philippines in figures*. Quezon City, Philippines: Republic of the Philippines, Philippine Statistics Authority. https://psa.gov.ph/sites/default/files/2021_pif_final.pdf.

Phillips, Colin & Ehrenhofer, Lara. 2015. The role of language processing in language acquisition. *Linguistic Approaches to Bilingualism* 5(4). 409–453. DOI: https://doi.org/10.1075/lab.5.4.01phi

Pizarro-Guevara, Jed Sam. 2014. *The acquisition of Tagalog relative clauses: A comprehension study*. Long Beach, CA: California State University, Long Beach MA thesis.

Pizarro-Guevara, Jed Sam. 2020. *When human universal meets language specific*. Santa Cruz, CA: University of California, Santa Cruz dissertation.

Pizarro-Guevara, Jed Sam & Garcia, Rowena. 2024. Philippine psycholinguistics. *Annual Review of Linguistics* 10. DOI: https://doi.org/10.1146/annurev-linguistics-031522-102844

Pizarro-Guevara, Jed Sam & Wagers, Matthew. 2020. The predictive value of tagalog voice morphology in filler-gap dependency formation. *Frontiers in Psychology* 11. 517. DOI: https://doi.org/10.3389/fpsyg.2020.00517

R Core Team. 2020. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/.

Rackowski, Andrea. 2002. *The structure of Tagalog: Specificity, voice, and the distribution of arguments*. Cambridge, MA: Massachusetts Institute of Technology dissertation.

Rackowski, Andrea & Richards, Norvin. 2005. Phase edge and extraction: A Tagalog case study. *Linguistic Inquiry* 36(4). 565–99. DOI: https://doi.org/10.1162/002438905774464368

Sauppe, Sebastian. 2016. Verbal semantics drives early anticipatory eye movements during the comprehension of verb-initial sentences. *Frontiers in Psychology* 7. 95. DOI: https://doi.org/10.3389/fpsyg.2016.00095

Sauppe, Sebastian & Næss, Åshild & Roversi, Giovanni & Meyer, Martin & Bornkessel-Schleswesky, Ina & Bickel, Balthasar. 2023. An agent-first preference in a patient-first language during sentence comprehension. *Cognitive Science* 47(9). e13340. DOI: https://doi.org/10.1111/cogs.13340

Schachter, Paul. 1977. Reference-related and role-related properties of subjects. In *Syntax and semantics: Grammatical relations*, vol. 8, 279–306. New York, NY: Academic Press. DOI: https://doi.org/10.1163/9789004368866_012

Schachter, Paul & Otanes, Fe T. 1983. *Tagalog reference grammar*. Berkeley, CA: University of California Press.

Tanaka, Nozomi. 2016. *An asymmetry in the acquisition of Tagalog relative clauses*. Honolulu, HI: University of Hawai'i at Manoa dissertation.

Tanaka, Nozomi & O'Grady, William & Deen, Kamil & Bondoc, Ivan Paul. 2019. An asymmetry in the acquisition of relative clauses: Evidence from Tagalog. *First Language* 39(6). 618–32. DOI: https://doi.org/10.1177/0142723719859090

Trueswell, John C. & Sekerina, Irina & Hill, Nicole M. & Logrip, Marian L. 1999. The kindergarten-path effect: studying on-line sentence processing in young children. *Cognition* 73. 89–134. DOI: https://doi.org/10.1016/S0010-0277(99)00032-3

Zeijlstra, Hedde. 2024. Two varieties of Korean: Rightward head movement or polarity sensitivity? *Linguistic Inquiry* 55(3). 622–41. DOI: https://doi.org/10.1162/ling_a_00471