2023. *Proc Ling Soc Amer* 8(1). 5533. https://doi.org/10.3765/plsa.v8i1.5533.

Examining voice choice in Tagalog: A corpus of web-based Tagalog

Norielle Adricula*

Abstract. This study is a corpus-based analysis of web-based Tagalog (Austronesian) investigating how different prominence features influence voice in basic, declarative, transitive clauses. A large sample of these structures were extracted from a web-based corpus of Tagalog. The arguments were annotated for animacy, definiteness, and other factors proposed to influence voice choice. Preliminary results suggest that despite the morphosyntactic symmetry in voice alternations in the language, the Undergoer voice appears to be the preferred structure regardless of these factors in Tagalog. Moreover, there may be highly constrained contexts in which the Actor voice is used when describing two-participant, transitive events. This work has implications for how we understand the notion of prominence more generally and how languages might have specific requirements for the mappings between different prominence hierarchies.

Keywords. Tagalog; corpus study; prominence hierarchies; syntax-semantics

1. Introduction. Across many languages, the coding of arguments relies not just on semantic and syntactic roles, but on referential prominence such as animacy, definiteness, topicality, and others. Prominence refers to a hierarchical ranking of features such that entities that are higher on a feature are more prominent compared to those having lower ranked features. For example, across many languages, agents on the semantic role hierarchy are frequently more prominent compared to other semantic roles: Agent > benefactive/goal/experiencer > patient/theme (Riesberg & Primus 2015; Jackendoff 1975; Grimshaw 1990; many others).

Role prominence is often realized syntactically. Across languages, agents are typically the unmarked choice for syntactically privileged positions (e.g., sentence-initial position) and/or placed before other verbal arguments (e.g., Riesberg et al. 2019; Primus 1999); the unmarked choice for privileged syntactic arguments (PSA) such as grammatical subject; the unmarked or less marked case (e.g., nominative); and often the argument that the verb agrees with (Riesberg et al. 2019). In addition to its semantic role, the encoding of an argument may also depend on its referential prominence features such as animacy, definiteness, person, etc., (also proposed to be hierarchical). For example, on the animacy scale, humans tend to be more prominent than other animates (subject to language and culture specific constraints) and animates tend to be more prominent than inanimates (e.g., Aissen 2003; see Aissen 2003 for definiteness scale). Crosslinguistically, these properties frequently correlate with the Actor role such that they are usually +animate/+human, +definite, +1st person, +nominative in nominative-accusative languages, etc. (Primus 1999; Bornkessel-Schlesewsky and Schlesewsky 2009). That is, higher ranked semantic roles tend to be more referentially prominent (Haspelmath 2021) and when they are not, languages may have ways of coding this less expected mapping. For example, when a patient (Undergoer) is high on the animacy or definiteness hierarchy, it is often marked or specially coded (Haspelmath 2021) to indicate an atypical mapping between role and referential prominence

* Thank you to my advisors Bhuvana Narasimhan and Laura Michaelis, and my colleagues in the LDC Lab and the Semiotic Syntax Working Group who provided valuable feedback on this work. Thanks as well to the reviewers of the Colorado Research in Linguistics Journal who had commented on an earlier version of this work. All remaining errors are my own. Authors: Norielle Adricula, University of Colorado Boulder (norielle.adricula@colorado.edu).

(e.g., Differential Object Marking, Aissen 2003). The mappings between prominence hierarchies appear to be complexified, if not challenged, by symmetrical voice languages such as Tagalog (Austronesian) where there is not a clear preference for mapping referentially prominent Actor arguments to the highest grammatical role.

2. Background and theoretical grounding. Tagalog is part of the Central Philippine subgroup of Philippine languages and is part of the Western-Malayo-Polynesian set of Austronesian languages. It is native to Manila, the largest city of the Philippines and is, along with English, the lingua franca in many cities. Tagalog is spoken by ~ 21.5 million speakers in the Philippines (Sauppe et al. 2013). As of 2008, it was estimated that over 90% of the population in the Philippines is either a first- or second-language speaker of Tagalog (Schachter & Reid 2008). Speakers tend to be multilingual in Tagalog, English, and/or another Philippine language.

Grammatical relations such as subject or object may not be applicable to Tagalog (e.g., Schachter 1976, 1977; Schachter & Otanes 1972; Naylor 1995; Kroeger 1993; Himmelmann, 2008). Therefore, I draw on a few key concepts from the framework of ROLE AND REFERENCE GRAMMAR (RRG, e.g., Foley and Van Valin 1984, Van Valin and La Polla 1997, etc.). RRG defines two types of semantic roles: thematic relations in the traditional sense of agent, theme, patient, experiencer (Fillmore 1968; Gruber 1965) and generalized semantic roles called Semantic Macroroles. The macroroles play a central role by acting as the interface between the arguments in a verb's structure (in RRG, Logical Structure) and syntactic representations. The macroroles Actor and Undergoer each subsumes specific semantic relations. In this paper, different voice options will be discussed using Actor and Undergoer macrorole terms. Within RRG, grammatical relations such as subjects, objects, etc. are replaced with the notion of a privileged syntactic argument (PSA), which is a "construction-specific relation and is defined as a restricted neutralization of semantic roles and pragmatic functions for syntactic purposes" (Van Valin 2002, p. 18).

Tagalog has more than one transitive construction, whereby transitive refers to two+ participant constructions that is used to describe one entity acting on another entity. Here, I use a semantic based understanding of voice (cf. Klaiman 1991; Shibatani 2005) in which voice is a system that regulates the ways nominal positions in basic sentences are assigned or correlated with roles that pertain to participants in the event (Latrouite 2011). I will focus on two constructions, the Actor and Undergoer voice forms which are functionally very different from active/passive alternations that you might find in languages like English. On morphological grounds, no verbal voice form in Tagalog can be considered basic, as all verbs consist of a verb stem plus a distinct voice affix. Furthermore, the Actor is neither demoted nor dropped as is the case with actors in passive sentences. This suggests that Tagalog is morphosyntactically symmetrical, contrasting with asymmetrical voice systems such as the English active/passive pattern (e.g., Latrouite 2011; Schachter 1976, 1977, 1995; Himmelmann 2008). For example:

(1) Tagalog (Undergoer voice)
B<in>ili ng guru ang libro
buy<UV>.PFV. NG teacher. ANG book
'A/The teacher bought the book.'

(2) Tagalog (Actor voice)
B<um>ili ang. guru ng libro
buy<AV>.PFV. ANG. teacher NG book
'The teacher bought a book'

Arguments preceded by the marker *ang* are analyzed as the PSA since it is the argument that "agrees" with the verb's morphology and is the target of a range of syntactic operations (Himmelmann 2008; Shibatani 1991; Kroeger 1993; Schachter, 1976; 1977; 1995). Arguments that are preceded by *ng* (and potentially *sa,* see Latrouite 2011) are understood as core, but non-privileged syntactic arguments (NPSA). In Tagalog, pronouns and proper names are not marked by *ang, ng,* or *sa,* however they have corresponding forms to the three markers[1]. Undergoer voice structures have *ang*-marked (PSA) Undergoers and Actor voice structures have *ang*-marked (PSA) Actors. If a predicate has voice affixation, the semantic role of the argument that is *ang*-marked is overtly marked by the voice affix on the predicate (Himmelmann 2008).

Actors and non-Actors (patients, themes, goals, etc.) have potential for *ang*-marking. However, when the Actor is not *ang*-marked (as in patient/locative/instrumental "voice" structures where non-Actor roles are *ang*-marked), first, the Actor is not demoted or dropped as is the case with actors in passive sentences. Furthermore, the *ng*-Actor retains many subject-like properties, such as reflexive binding, control of an actor gap in the second coordinated clause, deletion in imperatives, and control of a gap in subordinated clauses (Schachter 1977; Shibatani 1991; Kroger 1993; Shibatani 2005; Latrouite 2011). On the other hand, *ang*-marked non-Actor arguments also show several subject properties such as verb agreement, extractability, control of floating quantifiers and gaps in *samptan* ('while') clauses (Shibatani 1991). This suggests that there is a neutralization of the Actor/Undergoer macroroles distinction in specific constructions where the Undergoer roles are *ang*-marked.

2.1. THE UNDERGOER VOICE PREFERENCE. Basic sentences typically have one *ang*-phrase (Schachter & Otanes 1972). The *ang*-marked argument "agrees" with the verb and is understood to be who or what the verb is about. The argument that is *ang*-marked is understood to be a main determiner of the voice structure: when the Actor is *ang*-marked then the clause is in the Actor voice. When the Undergoer is *ang*-marked, the Undergoer voice. Some analyses posit the *ang*-marked argument as the most prominent (Latrouite 2011) or "salient" (Wouk 1986) argument. Patterns of prominence features such as definiteness, animacy, topicality, etc., that are often associated with these (macro)roles also play a significant role in how semantic roles are mapped to syntactic roles which will be briefly reviewed below. Tagalog exhibits a preference for mapping Undergoers to the privileged syntactic argument and Actors to the non-privileged syntactic argument (cf. "patient primacy" Cena 1977). A range of linguistic and psycholinguistic studies provides ample evidence that Tagalog speakers prefer to use the Undergoer voice (Katagiri 2005; Wouk 1986; Cooreman et al. 1984; Sauppe et al. 2013) even if the Undergoer may be low on animacy or definiteness (Adricula 2022). Similarly, children produce and comprehend Undergoer voice structures more easily than Actor voice structures (Marzan 2013; Garcia et al. 2019; Kidd & Garcia 2021). This preference seems to run contrary to crosslinguistic tendencies to map Actors to the highest grammatical role, particularly when they are high on different prominence hierarchies. The research described below suggests a complex mapping process between different prominence hierarchies in Tagalog syntactic choice.

2.2. ANIMACY AND VOICE. Animacy, often correlated with definiteness and other referential features, is proposed to play a role in prominence and *ang*-marking. Generally, the Actor voice (*ang*-marked Actors) tends to be less acceptable with a human undergoer. For example,

[1] For example, *ang*-form pronouns include *ako* PSA.1sg; *ka* PSA.2sg; *siya* PSA.3sg; *ng*-form pronouns include *ko* NPSA.1sg; *mo* NPSA.2sg; *niya* NPSA.3sg (e.g., Schachter & Otanes, 1972; Schachter & Reid, 2008)

(3) Tagalog (Saclot 2006)

a. K<um>agat ang aso ng/sa. buto
<AV>PFV.bite PSA dog. NPSA/SA. bone
'The dog bit a bone.'

b. K<in>agat ng aso ang buto/si Lena
<UV>PFV.bite NPSA dog PSA bone/ANG Lena
'A dog bit the bone/Lena.'

c. ??K<um>agat. ang aso sa akin/kay Lena
<AV>PFV.bite PSA dog SA 1SG.SA/SA Lena

Examples (3a-c) show different kinds of acceptability between the Actor voice and Undergoer voice given different animacy values for the Undergoer. 3a shows that Actor voice is acceptable when the Undergoer is inanimate but 3c shows that it is much less acceptable when the Undergoer is human. Sentence production experiments with Tagalog speakers show that participants prefer to use the Undergoer voice when the Undergoer is human, even when the Actor is also human and the Undergoer is human or non-human (Sauppe 2017).

2.3. DEFINITENESS AND VOICE. Early analyses of Tagalog *ang*-marked noun phrases propose that *ang*-marked phrases are definite while *ng*-marked phrases, particularly objects, are indefinite (Bowen 1965; Schachter & Otanes 1972; Naylor 1975, etc.). Other Philippine languages such as Ilokano (Schwartz, 1976), Hiligaynon (Wolfenden 1971) and Cebuano (Wolff 1966) appear to exhibit this correlation between definiteness and marking structures analogous to *ang*. Schachter (1976) and Schwartz (1976) proposed that definite patients[2] are *ang*-marked and in the case where none of the noun phrases is definite, Tagalog may resort to *ang*-less existential constructions (Schachter & Otanes 1972). Obligatory definiteness has been cited as one of the reasons the Tagalog *ang*-marked noun phrase is different from canonical subjects (e.g., Schachter, 1976; Richards, 2000; many others). Broadly speaking, definite descriptions necessitate that the hearer is not free to assign just any value to the discourse referent introduced by the noun phrase and are subject to a familiarity requirement (e.g., Latrouite 2011; Paul, Cortes & Milambiling 2015).

Although definiteness appears to play a significant role in voice choice, it is not the case that *ang*-phrases and *ng*-phrases are straightforwardly marked for (in)definiteness. Bell (1978) and Adams & Manaster-Ramer (1988) show that *ang*-phrases can be indefinite and are the generally accepted interpretation when in the presence of indefinite quantifiers such as *isa-ng* 'one' or *iba-ng* 'other,' *marami-ng* 'many,' or *anuman* 'anything,' and others:

(4) Tawag-an ang isa-ng pediatric dermatologist. kung…
call-UV.IMP PSA ONE-LNK. pediatric dermatologist. COND…
'Call a pediatric dermatologist if…' (SketchEngine tlTenTen2019, website: krikids.com)

In (4), only an indefinite interpretation for *ang isang pediatric dermatologist* is possible despite its *ang*-marked status, suggesting that *ang* is not necessarily marked for definiteness (see Paul et al.,2015 who suggests that *ang* is also unmarked for familiarity). Additionally, under a discourse-based definition of definiteness, definiteness appears to be weighed differently between the Un-

[2] A similar argument has been made for specificity. However, Sabbah (2026) and Ross (2002) claim that the marker *ng* does not seem to encode non-specificity. Instead, the *ng*-Undergoer is interpreted as non-specific through pragmatic inference because a specific common noun phrase patient would normally be the pivot (Ross, 2002; p. 27).

dergoer voice and Actor voice such that highly definite Undergoers tend to be *ang*-marked, but indefinite Undergoers do not necessarily mean Actors will be *ang*-marked (Wouk 1986, see similar proposal for specific Undergoers vs. Actors in Latrouite 2011). In summary, while (in)definiteness appears to be a significant factor in voice choice, it does not seem to be sole defining factor for voice and *ang/ng*-marked phrases.

2.4. ACCESSIBILITY (ARGUMENT REALIZATION) AND VOICE. Accessibility, or how accessible an entity is in a discourse (and perhaps cognitively, see Ariel 1991) is a related measure to definiteness, animacy and many others (e.g., givenness, topicality, etc.). Arguments that are referred to using pronouns, zeros, demonstratives, etc. are understood to be more accessible (e.g., given) than an argument that is realized as a lexical realization (e.g., full NP, clausal arguments, free relative clauses, etc.). Cooreman and colleagues' (1984) work on Tagalog determined that Undergoers tended to be realized by lexical NPs and Actors tended to be realized by non-lexical forms (i.e., pronouns, zeros, etc.). In general, they found that most definite full NPs in the text in Undergoer voice clauses represented discontinuous, low-topicality patient NPs. An argument that is more accessible tends to be more likely to be assigned a prominent grammatical function, such as subject or pivot. Riesberg and colleagues (2021) use the lexical/non-lexical distinction in addition to the new/given distinction to code argument realization in Totoli (Austronesian). This may more accurately measure the accessibility of an entity than the new/given distinction since any re-mention of an entity may be labeled as "given" but it is unlikely that a referent that has not been mentioned in some time would be referred to using a pronoun. Riesberg and colleagues found that accessibility (for them, activation state) was a significant predictor of voice choice such that when both referents are non-lexical (which was the most common Actor/Undergoer combination), there remains a persistent preference for Undergoer voice. However, when the Undergoer is lexical (thus, likely new), Actor voice is increased.

2.5. OPEN QUESTIONS. Given the prior literature, it remains an open question as to how different prominence features map onto each other and how they map onto voice choice in Tagalog. Is it the case that Undergoers are more prominent than Actors in these constructions? How do these different prominence hierarchies map onto each other? Do we see these similar effects of prominence mappings in large samples of naturalistic text? To address these questions, I conduct corpus-based analyses of web-based Tagalog to examine how different prominence features influence voice in basic, transitive structures (e.g., Hopper & Thompson 1980). Despite the morphological symmetry of the language, there appears to be a significant asymmetry in Undergoer and Actor voice usage, and in how prominence influences voice choice.

3. Methods. All data were extracted from the tlTenTen 2019 Tagalog (Filipino) Web-based corpus which is part of the TenTen corpora (Jakubíček et al. 2013) and made up of web-crawled texts collected from the Internet. The corpus has 232,854,660 tokens collected from websites such as Wikipedia, news articles, narratives, blog posts, and many others (Jakubíček et al., 2013). The corpus was previously POS-tagged using a Filipino-tagger model (Go & Nocon 2017) which was previously based on the Stanford parser (e.g., Toutanova & Manning 2000). The overall tagging accuracy of this model was reported at 96.15% (Go & Nocon 2017). All collocational analyses and extractions were performed using the SketchEngine concordance and Corpus Query Language (CQL) search tools which allows you to search for grammatical or lexical patterns in the corpus. A randomized sample of tagged Actor voice and Undergoer voice clauses (n = ~2300) were added to a preliminary sample that I extracted for an earlier corpus study on verbs (Adricula 2022). The window for each extraction included two to three sentences

preceding and following each token to provide context for the token. Of that sample, 631 clauses were processed as basic, verb-initial, transitive clauses containing Actor and Undergoer arguments. I excluded clauses where the predicate form was nominalized, where the PSA choice is grammatically determined (such as in relatives), clauses whose word order are not considered basic, such as *ay*-clauses, "focused" clauses, and others.

3.1. ANNOTATION SCHEME. Each token was annotated for a variety of features with respect to the verb and the PSA and NPSA. An abbreviated coding scheme of the prominence features annotated for are given below:

- **Macroroles:** *ang, ng,* and *sa*-marked arguments were coded for their macrorole (Actor/Undergoer) in the clause, subsuming more specific roles like agent, cause/patient, theme etc.
- **Animacy:** each argument was coded using a typological animacy hierarchy: Human >Animate> Inanimate > Abstract (e.g., Primus, 1999; Aissen, 2003)
- **Definiteness**: broadly defined as a feature of a referent in which the hearer is not free to assign any value to the referent. They are often subject to a familiarity requirement where the value of a referential term is determined by previous discourse and/or context (Aissen, 2003) and their absence, presence, and individuation in the context (Wouk 1986). Aissen's (2003) definiteness scale was used: Personal pronoun > Proper name > Definite NP > Indefinite specific NP > Non-specific NP (e.g., Aissen, 2003).
- **Argument Realization (proxy for accessibility)**: each argument was coded using a form-based analysis of their referential form (cf. Ariel, 1991). Forms were coded as being lexical (full NPs or free relative clauses) or non-lexical (pronouns, zeros, demonstratives, etc.). This analysis may provide a more accurate representation of an entity's accessibility compared to the given/new distinction (Riesberg et al., 2021).

4. Results. Figure 1 shows the proportions of occurrence for Undergoer (blue) and Actor voice (orange) in the sample (n = 631). Similar to prior research, there were significantly more Undergoer voice structures (n = 520, prop. = .82) compared to Actor voice (n = 111, prop. = .18). The following sections will further examine the extent to which the referential prominence features of definiteness, animacy, and accessibility (argument realization) of the Actor and Undergoer influence which will be the PSA (*ang*-marked).

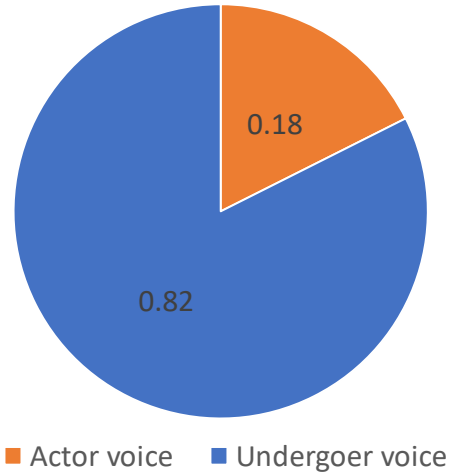

Figure 1. Proportion of Undergoer voice and Actor voice

4.1. ANIMACY AND VOICE. Figure 2 shows the proportions of Undergoer and Actor voice forms when they are distributed by different animacy configurations, beginning with human Actors acting on human Undergoers (H-H), human Actors acting on non-human Undergoers (H-NH), and so on.

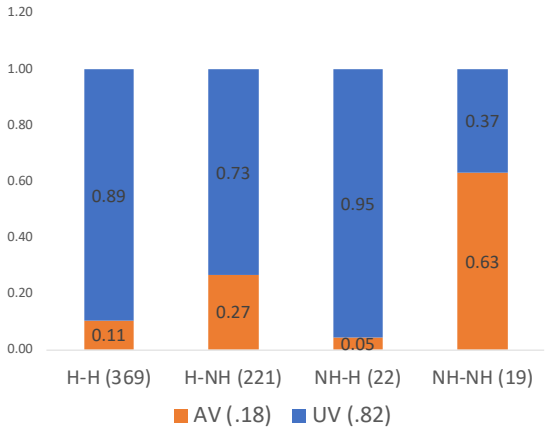

Figure 2. Proportions of Undergoer voice and Actor voice across animacy configurations

Typically, the most common animacy configuration is one in which human Actors act on human Undergoers. This is followed by human Actors acting on non-human Undergoers, a tendency which aligns with both linguistic and cognitive associations between Actor arguments and the animacy hierarchy. Although Actors tended to have equal or higher animacy compared to Undergoers, in most cases, there is a preference for the Undergoer voice. When both Actors and Undergoers are human, there is a decrease in Actor voice. If we expect that the Actor is both the prominent thematic argument and is prominent on the animacy hierarchy, then we would expect that the Actor would more frequently map to the PSA, and thus the Actor voice would be the more frequent structure. However, there is an increase in the Actor voice only when the Actor is human and the Undergoer is non-human, though the preference does not flip. When both entities are non-human (column 4), there appears to be more Actor voice, though the sample is comparatively small. These results suggest that the use of Actor voice may be more constrained to situations where the Actor is higher on the animacy hierarchy than the Undergoer. Given the suggestive proportions in the Non-human A/ Non-human U column, Actor voice may be broadly constrained to scenarios where the Undergoer is lower on the animacy hierarchy. Below are examples of Actor voice and Undergoer voice in different animacy configurations.

(5) Ni-yakap na lang niya ako
UV.PFV-hug only NPSA.3SG PSA.1SG
'He just hugged me.'

Example (5) shows the most frequent scenario: Undergoer voice with human Actors and human Undergoers. Here, we see the Undergoer voice form of *yakap* 'hug,' which has an Undergoer voice affix which agrees with the 1sg PSA-pronoun *ako*. The Actor is referred to using a *ng*-pronoun and is the non-privileged syntactic argument. In this instance, both Actor and Undergoer are human and the Undergoer is the privileged syntactic argument.

In the next example, we see the Actor voice with human Actors and human Undergoers, a relatively infrequent occurrence in the sample:

(6) ...kapag s<in>abihan mo-ng naka-patay ka ng tao…
…when <UV>.PRV.say NPSA.2SG-LNK AV-kill PSA.2SG NPSA person
'[A true friend,] when you say you killed a person…'

In example (6), we see that both the Actor and Undergoer to the verb *patay* 'kill' are human. And though the verb *patay* may exhibit a verb-specific bias towards the Undergoer voice (Latrouite 2011; Adricula 2022), the Actor *ka* 'you' is the addressee and in its *ang*-form and the PSA. The Undergoer *tao* 'human' is clearly human, so just based on animacy it is not clear why the Actor may be the PSA. However, *tao* is also a generic reference to any human. Here, another feature like definiteness (identifiability) or specificity, might be more influential than animacy in mapping the Actor to the PSA. In particular, a non-specific or indefinite Undergoer may influence the mapping of the Actor to the PSA.

Example (7) shows an example of the Actor voice with Non-animate Actors and Non-animate Undergoers, a configuration that seems to allow (proportionally) for more Actor voice uses.

(7) ang neoliberalismo, halimbawa nag-tu~tulak pa rin ito
PSA neoliberalism for example AV-IPFV~push also PSA.DEM

ng mga patakaran sa bansa
NPSA PL policy to the nation
'Neoliberalism, for example, this pushes policy to the nation…"

Here, the verb *nagtutulak* 'push,' has an Actor voice affix *nag-* and takes as arguments an abstract causer *ito* (neoliberalism) and an abstract Undergoer *patakaran* 'policy.' Both entities may be considered low on the animacy hierarchy, so neither entity would be considered eligible for *ang*-marking given this feature. However, information structure may also be at work here. Specifically, 'neoliberalism' is introduced at the very beginning of the clause, perhaps in a kind of topicalization and/or presentational function, which is again referred to using a demonstrative *ito* 'this' as the PSA argument to the verb 'push.' This may be an instance of an Actor argument functioning as a topic such that its prominence on an information structural level is increased, thus licensing its mapping to the PSA in the clause (see Latrouite 2011, p. 111 on principles for Actor voice Selection).

In sum, these results suggest that animacy may play a very specific role in voice marking. That is, maximally different Actors and Undergoers in terms of animacy, where the Actor is high on animacy and the Undergoer is low on animacy might increase the likelihood of Actor voice. However, in general, the Undergoer voice preference persists. Furthermore, the role of animacy may interact or be influenced by other referential prominence features such as definiteness and accessibility, which I examine below.

4.2. DEFINITENESS AND VOICE. Figure 3 shows the proportions of Undergoer and Actor voice forms distributed by different definiteness configurations, beginning with Definite Actors acting on Definite Undergoers, Definite Actors acting on Indefinite Undergoers, and so on.

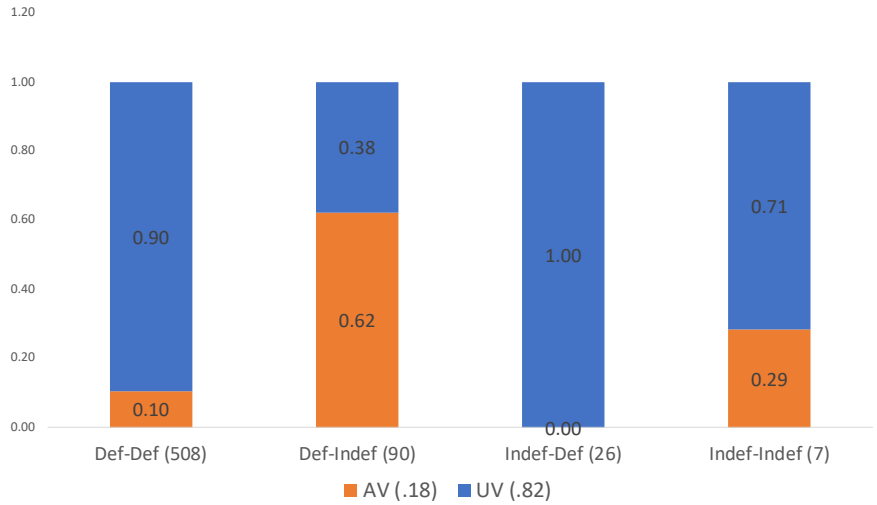

Figure 3. Proportions of Undergoer and Actor voice across definiteness configurations

In this sample, both Actors and Undergoers tend to be definite entities. Specifically, both entities tend to be referred to using pronominal forms and definite NPs (given and identifiable). In this configuration, the Undergoer voice appears to be the most prevalent voice. In the second column, the Actor voice increases when the Actor is definite and the Undergoer is indefinite, a similar pattern to what was seen with animacy. Specifically, when definiteness is low for the Undergoer, we see increases in the Actor voice. Below are relevant examples:

(8) H<in>awak-an ni Clyde ang aking kamay
<UV.REAL>hold-UV NPSA Clyde PSA 1POSS hand
'Clyde held my hand.'

Example (8) represents the most frequent configuration in terms of definiteness, specifically the Undergoer voice with a Definite Actor and Definite Undergoer. Here, we see the Undergoer voice affix for the verb *hawak* 'hold,' which is co-indexed with the definite Undergoer argument *ang aking kamay* 'my hand.' The Actor *Clyde* is *ng*-marked and is the NPSA. Under Aissen's (2003) definiteness hierarchy, it may be the case that *Clyde* as a personal name may be higher on the definiteness hierarchy compared to the definite noun phrase *ang aking kamay.* However, it is also possible that the first-person possessive pronoun which is referential with the speaker makes this phrase more definite and thus more prominent and eligible for mapping to the PSA role.

There appears to be an increase in Actor voice when the Actor is definite and the Undergoer appears indefinite:

(9) B<um>unot siya ng isa-ng itim na wallet…
<AV>PFV.pickup PSA.3SG NPSA. INDF-LNK black LNK wallet
'He just picked up a black wallet…'

In example (9), we see the *ang*-form of the third person singular pronoun *siya*, which is the Actor argument to the verb *b<um>unot*. The Undergoer is preceded by the NPSA *ng*-marker and the indefinite quantifier *isang.* This suggests that the *ng*-marker can be marked for indefiniteness, while the following example (10) shows that it can also refer to a definite entity:

(10) Nag-punas na lang ako ng pawis
AV.PRV-wipe just 1SG.PSA NPSA sweat
'I just wiped (my) sweat.'

In example (10), the verb *punas* 'wipe,' has as arguments the Actor who did the wiping and the Undergoer which was either a body part or bodily fluid, such as tears or sweat, that was often implicitly co-referential with the Actor. We infer that the possession of the Undergoer is implied even if it is not formally marked. It is not the case that the Actor is somehow wiping someone else's mouth or sweat. Possession, and thus definiteness, is only implied here, providing supporting evidence for Adams & Manaster-Ramer (1988) and Paul et al.'s (2015) observations that *ng*-arguments are not necessarily marked for indefiniteness. Furthermore, it is not necessarily the case that definiteness marking is required for *ang*-marking. The following example suggests that even when the Undergoer is explicitly marked as being co-referentially possessed by the Actor and thus definite, the Actor voice is still available:

(11) Um-order din siya ng sarili niya-ng pagkain.
AV.PRV-order also PSA.3SG NPSA REFL 3POSS-LNK food
'He also ordered his own food'

These results suggest that like animacy, definiteness may play a specific role in voice marking. In general, there may be much less variation in the definiteness configurations between Actors and Undergoers since they both tend to be definite. However, there is an increased likelihood of Actor voice when Undergoers are indefinite, especially when Actors are definite and Undergoers are indefinite. This accords with prior work that shows that the Tagalog Actor voice is less frequent and more constrained (Latrouite 2011; 2016) and perhaps more marked. The contexts for mapping the Actor to the PSA may rely more on the Actor having higher definiteness than the Undergoer, though the reverse may not be the case (Wouk 1986).

4.3. ACCESSIBILITY (ARGUMENT REALIZATION) AND VOICE. Figure 4 shows the proportion of Actor and Undergoer voice forms distributed among argument realization (non-lexical/nL and Lexical/L NP) configurations.

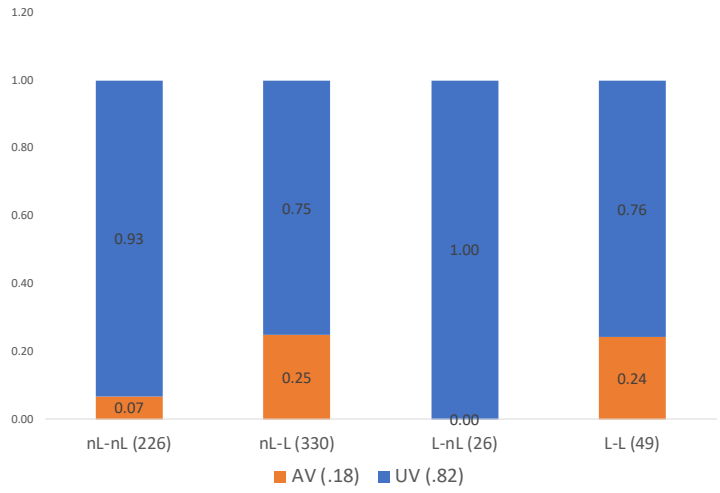

Figure 4. Proportions of Undergoer and Actor voice in argument realization configurations

In general, there are relatively few instances where both arguments are lexical. This is not that surprising given that full lexical NPs tend to be relatively rare compared to pronominal and 0-forms in naturalistic language, particularly for Actors (e.g., DuBois 2003). There is a significant proportion of non-lexical Actors and Undergoers, of which a high percentage of these are non-lexical (highly accessible) actors. Comparatively, columns two and four demonstrate that there are quite a few lexical (less accessible) Undergoers. Furthermore, these columns show that when the Undergoer is lexical there are proportionately more Actor voice instances, which is similar to the findings in Totoli (Riesberg et al. 2021). When the Undergoer is non-lexical, there is more Undergoer voice. Additionally, the distributions of lexical/non-lexical Actor and Undergoer configurations differ substantially from the numbers of Definite/Indefinite Actors and Undergoers. This suggests that accessibility (argument realization) and definiteness may still capture slightly different phenomena even if they are related constructs that contribute to an argument's prominence. Utilizing statistical tests will be able to tell us if these associations are statistically significant and the extent to which a feature individually contributes to prominence and its influence on the mapping from argument to grammatical role (see Riesberg et al. 2021). Like the prominence features of animacy and definiteness, the increase in the Actor voice appears to be affected by the Undergoer and whether it is high or low accessibility. More broadly, there does not appear to be a default mapping between the PSA and highly accessible Actors. Below are representative examples:

(12)	H<in>anap	ko	ang	cellphone	sa	bag.
	<UV.PFV>find	NPSA.1SG	PSA	cell phone	PREP	bag
	'I looked for the cellphone in the bag.'					

Example (12) represents the frequent occurrence of the Undergoer voice when the Actor is non-lexical and the Undergoer is lexical. Here, we see the Undergoer voice affix for the verb *hanap,* used with the bare, lexical noun phrase *ang*-marked 'cellphone' and the Actor *ko* which is the *ng*-form of the first-person pronoun. Because the cellphone is a lexical noun phrase form, we would expect that it would be less accessible compared to the speaker. However, it is not readily apparent based on its proxy accessibility why the cellphone would be the PSA in this instance.

Like with animacy and definiteness, there appears to be an increase in the Actor voice when the Actor is non-lexical and the Undergoer is lexical. For example,

(13)	k<um>ain	na	ako	ng	almusal
	<AV>PFV.eat	already	PSA.1sg	NPSA	breakfast
	'I already ate breakfast.'				

We see the Actor voice affix with the verb *kain,* used with the *ang*-form of the first-person pronoun. The *ng*-marked Undergoer *almusal* is a bare, lexical noun phrase and presumably is less accessible (and generic), thus mapping to the NPSA. By contrast, although the *cellphone* argument in (12) is lexical (less accessible), it may be specific in comparison to *almusal* in (13), allowing it to map to the PSA.

In the lexical Actor and lexical Undergoer configuration, there is also a relatively higher proportion of Actor voice. One interesting example of this is shown below:

(14)	Ayon sa	isang	mangingisda,	b<um>angga	ang	kanyang
	PTCL.about	one	fisherman	b<AV>PFV.bump	PSA	POSS

bangka	ng	buwaya,	na	naka-harang	sa	daanan	nito.
boat	NPSA	crocodile	REL	blocking	PREP	way	POSS

'Regarding the fisherman, his boat bumped a crocodile, which was blocking his way.'

Here, although both *ang kanyang bangka* 'his boat' and *buwaya* 'crocodile' are both lexical noun phrases, the voice of the structure is in the Actor voice. This may be in part due to the fact that *isang mangingisda* is left-dislocated for the purposes of introducing the fisherman as a topic (Gregory & Michaelis 2001): there is a fisherman and it is his boat that bumped into the crocodile. Although the following relative clause description of the crocodile might further individuate the crocodile, the information appears to function more like background information to explain the presence of the crocodile (though this is a subjective reading).

As with animacy and definiteness, we see that the Actor is often more accessible compared to the Undergoer. Despite that, there remains a persistent Undergoer voice preference. There is an increase in Actor voice under circumstances where the Undergoer is realized with a lexical form, suggesting that voice marking appears to be constrained by the features of the Undergoer. Again, it appears as though it is the features of the Undergoer and whether it is low or high on a feature that appears to drive whether the structure will be Undergoer voice or Actor voice.

5. Discussion and conclusions. Overall, there is a robust Undergoer voice preference in this large web-based corpus. Looking across different prominence features, we see a similar pattern: when Actors and Undergoers are both high on a prominence hierarchy, the Undergoer is likely to be mapped to the PSA. When the Actor is high on a prominence feature and the Undergoer is low on a prominence feature, there is more Actor voice, but the preference does not seem to flip. Overall, this suggests that the Actor voice is more likely to occur when the Undergoer is non-human, inanimate, indefinite, and has low accessibility. That is, highly restricted contexts. This supports prior research that has suggested that the Actor voice occurs when Undergoers have significantly lower topicality (as measured by topic continuity, Cooreman et al. 1984), or is minimally individuated or even absent altogether (Wouk 1986). These patterns appear to exhibit some divergence from broader cross-linguistic patterns of how thematic role prominence hierarchies interact with other prominence hierarchies such as animacy, definiteness, topicality, etc. and the mapping of these arguments to grammatical functions. Below is a schematic for the mappings of these prominence hierarchies for Tagalog Actor voice and Undergoer voice:

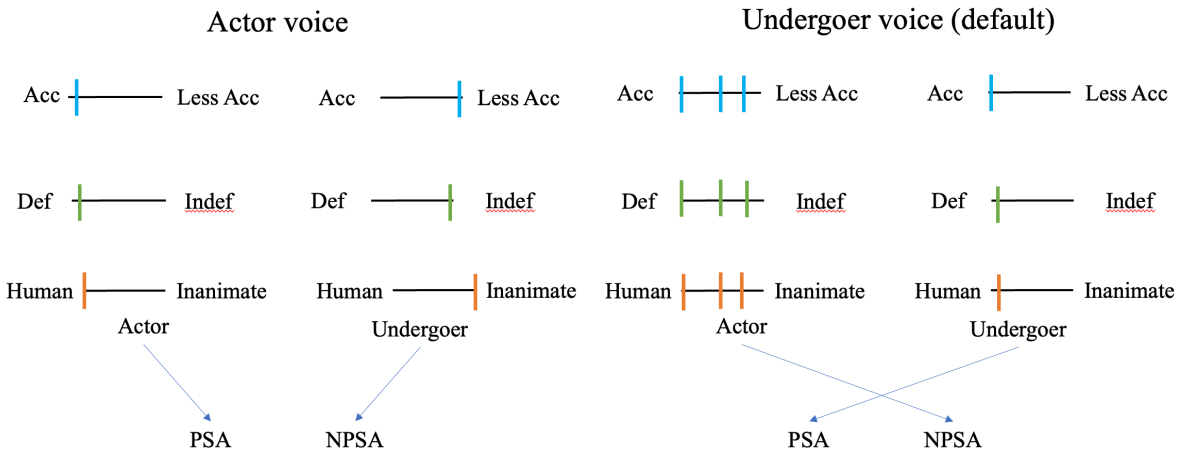

Figure 6. Proposed mappings for Actor and Undergoer voice

In the Actor voice, we see highly constrained contexts where the Actor is high on a prominence feature and the Undergoer is low on a prominence feature. The Undergoer voice on the other hand, is seemingly a default structure for transitive events. That is, if the Undergoer is referred to explicitly, the properties of the Actor are less influential on its capacity for mapping to the PSA, especially when the Undergoer is high on a prominence hierarchy.

On its face, these patterns suggest that perhaps the Undergoer role is more prominent than the Actor role in these structures, at least for the purposes of voice and *ang*-marking in basic, transitive events. Prior work by Riesberg & Primus (2015) suggests that Actor properties, such as volitionality, ability, and control manifest in other parts of the grammar and structure, such as in the verbal affix morphology, even in Undergoer voice constructions in which the Actor is not the PSA. Speculatively, the notion of prominence may be divided between different indices of prominence, such as *ang*-marking and word order. For example, other work has shown that in Tagalog, the Actor most frequently precedes the Undergoer in Undergoer voice structures (i.e., exhibits VAU word order), but word order is more variable in the Actor voice (e.g., Sauppe et al. 2013; 2017; Kroeger 1993). Other areas for further research include examining constructional influences on prominence, such as the role of the presentational construction in increasing the topic-ness of a particular argument. Early prior work (Adricula 2022) examined the potential for verb-specific patterns of occurrence in both voices and showed preliminary data that verb-specificity may play a role in voice marking. In sum, there is a complex interplay between these prominence hierarchies that results in these syntactic choices. Examining languages where these hierarchies may be "mismatched" provides evidence that the notion of "prominence" is a complex, multidimensional construct and future research should examine its multifaceted role in syntactic choice.

References

Adams, Karen L., and Alexis Manaster-Ramer. 1988. Some questions of topic/focus choice in Tagalog. *Oceanic Linguistics* 27 (1/2). 79–101. https://doi.org/10.2307/3623150.

Adricula, Norielle. 2022. Examining the Tagalog Undergoer voice preference in the context of verbs and referential features. *Colorado Research in Linguistics* 26. https://journals.colorado.edu/index.php/cril/article/view/1579.

Aissen, Judith. 2003. Differential object marking: Iconicity vs. economy. *Natural Language & Linguistic Theory* 21. 435–483. https://doi.org/10.1023/A:1024109008573.

Ariel, Mira. 1991. The function of accessibility in a theory of grammar. *Journal of Pragmatics* 16(5). 443–463. https://doi.org/10.1016/0378-2166(91)90136-L.

Bell, Sarah J. 1978. Two differences in definiteness in Cebuano and Tagalog. *Oceanic Linguistics* 17(1). 1–9. https://doi.org/10.2307/3622824

Bowen, Donald J. (ed.) 1965. *Beginning Tagalog*. Berkeley/Los Angeles: University of California Press.

Bornkessel-Schlesewsky, Ina & Matthias Schlesewsky. 2009. The role of prominence information in the real-time comprehension of transitive constructions: A crosslinguistic approach. *Language and Linguistics Compass* 3(1), 19–58. https://doi.org/10.1111/j.1749-818X.2008.00099.x.

Cena, Resty. M. 1977. Patient primacy in Tagalog. Presented at the Annual Meeting of the Linguistic Society of America, Chicago.

Cooreman, Ann, Barbara A. Fox, and Talmy Givón. 1984. The discourse definition of ergativity. *Studies in Language.* 8(1). 1–34. https://doi.org/10.1075/sl.8.1.02coo.

Du Bois, John W. 2003. Argument structure: Grammar in use. In John W. Du Bois, Lorraine E. Kumpf & William J. Ashby (eds.), *Preferred argument structure: Grammar as architecture for function,* 11–60. Philadelphia: John Benjamins. https://doi.org/j7rp.
Fillmore, Charles J. 1968. Lexical entries for verbs. *Foundations of Language* 4. 373–393.
Foley, William A. & Robert D. Van Valin Jr. 1977. On the viability of the notion of 'subject' in universal grammar. *Berkeley Linguistics Society (BLS)* 3. 293–320. https://doi.org/10.3765/bls.v3i0.3297.
Garcia, Rowena, Irina Sekerina, J. Dery, Jens Roeser & B. Höehle. 2015. Thematic role assignment in the L1 acquisition of Tagalog. Poster presentation. Architectures and Mechanisms for Language Processing.
Garcia, Rowena & Evan Kidd. 2020. The acquisition of the Tagalog symmetrical voice system: Evidence from structural priming. *Language Learning and Development* 16(4) 399–425. https://doi.org/10.1080/15475441.2020.1814780.
Go, Matthew Phillip & Nicco Nocon. 2017. Using Stanford part–of–speech tagger for the morphologically-rich Filipino language. *Proceedings of the Pacific Asia Conference on Language, Information and Computation* 31. 81–88.
Gregory, Michelle L. & Laura A. Michaelis. 2001. Topicalization and left-dislocation: A functional opposition revisited. *Journal of Pragmatics* 33(11). 1665–1706. https://doi.org/d8hpjr.
Grimshaw, Jane. 1990. *Argument structure*. Cambridge, MA: The MIT Press.
Gruber, Jeffrey Steven. 1965. *Studies in lexical relations*. Cambridge, MA: MIT dissertation.
Haspelmath, Martin. 2021. Role–reference associations and the explanation of argument coding splits. *Linguistics* 59(1). 123–174. https://doi.org/10.1515/ling-2020-0252.
Himmelmann, Nikolaus P. 2008. Lexical categories and voice in Tagalog. In Peter K. Austin & Simon Musgrave (eds.), *Voice and grammatical relations in Austronesian languages*, 247–293. Stanford, CA: CSLI.
Hopper, Paul J. & Sandra A. Thompson. 1980. Transitivity in grammar and discourse. *Language* 56. 251–299. https://doi.org/10.2307/413757.
Jackendoff, Ray S. 1975. Semantic interpretation in generative grammar. Cambridge, MA: The MIT Press.
Jakubíček, Miloš, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý & Vít Suchomel. 2013. The TenTen corpus family. *International Corpus Linguistics Conference* 7. 125–127.
Katagiri, Masumi. 2005. Voice, ergativity and transitivity in Tagalog and other Philippine languages: A typological perspective. In I Wayan Arka & Malcolm Ross (eds.), *The many faces of Austronesian voice systems: Some new empirical studies*, 153–174. Canberra: Pacific Linguistics.
Klaiman, Miriam H. 1991. *Grammatical voice.* Boston, MA: Cambridge University Press.
Kroeger, Paul R. 1993. Another look at subjecthood in Tagalog. *Philippine Journal of Linguistics* 24(2). 1–16.
Latrouite, Anja. 2011. *Voice and case in Tagalog: The coding of prominence and orientation.* Düsseldorf: Heinrich Heine University dissertation.
Latrouite, Anja, and Arndt Riester. 2018. The role of information structure for morphosyntactic choices in Tagalog. In Sonja Riesberg, Asako Shiohara & Atsuko Utsumi (eds.), *Perspectives on information structure in Austronesian languages,* 247–284. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.1402548.
Marzan, Jocelyn C. 2013. *Spoken language patterns of selected Filipino toddlers and preschool children*. Diliman, Quezon City: University of the Philippines Diliman dissertation.

Naylor, Paz Buenaventura. 1975. Topic, focus, and emphasis in the Tagalog verbal clause. *Oceanic Linguistics* 14(1). 12–79. https://doi.org/10.2307/3622792.
Paul, Ileana, Key Cortes & Lareina Milambiling. 2015. Definiteness without D: The case of *ang* and *ng* in Tagalog. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 60 (3). 361–390. https://doi.org/10.1017/S0008413100026256.
Primus, Beatrice. 1999. *Cases and thematic roles: Ergative, accusative and active*. Tübingen: Niemeyer.
Reid, Lawrence & Paul Schachter. 2008. Tagalog. In Bernard Comrie (ed), *The world's major languages* (2nd edn.), 833–855. London: Routledge.
Riesberg, Sonja, & Beatrice Primus. 2015. Agent prominence in symmetrical voice languages. *STUF–Language Typology and Universals* 68(4). 551–564. https://doi.org/gdz639.
Riesberg, Sonja, Kurt Malcher & Nikolaus P. Himmelmann. 2019. How universal is agent–first? Evidence from symmetrical voice languages. *Language* 95(3). 523–561. https://doi.org/10.1353/lan.2019.0055.
Riesberg, Sonja, Maria Bardají i Farré, Kurt Malcher & Nikolaus P. Himmelmann. 2021. Predicting voice choice in symmetrical voice languages: All the things that do not work in Totoli. *Studies in Language* 46(2). 453–516. https://doi.org/10.1075/sl.20061.rie.
Saclot, Maureen Joy. 2006. On the transitivity of the actor focus and patient focus constructions in Tagalog. In Linguistic Society of the Philippines (eds.), *Tenth International Conference on Austronesian Linguistics,* 17–20. Online: Linguistic Society of the Philippines and SIL International.
Sauppe, Sebastian. 2017. *The role of voice and word order in incremental sentence processing: Studies on sentence production and comprehension in Tagalog and German*. Nijmegen: Radboud University Nijmegen Dissertation.
Sauppe, Sebastian, Elisabeth Norcliffe, Agnieszka E. Konopka, Robert D. Van Valin Jr. & Stephen C. Levinson. 2013. Dependencies first: Eye tracking evidence from sentence production in Tagalog. *Proceedings of the Annual Meeting of the Cognitive Science Society* 35. https://escholarship.org/uc/item/9z68g7q5.
Schachter, Paul. 1976. The subject in Philippine languages: Topic, actor, actor–topic, or none of the above. In Charles Li (ed.), *Subject and topic*. 493–518. New York: Academic Press.
Schachter, Paul. 1977. Reference-related and role-related properties of subjects. *Grammatical relations*, 279–306. Leiden: Brill. https://doi.org/10.1163/9789004368866_012.
Schachter, Paul. 1996. The subject in Tagalog: Still none of the above. *UCLA Occasional Papers in Linguistics* 15.
Schachter, Paul & Fe Otanes. 1972. *Tagalog reference grammar.* Berkeley: University of California Press.
Shibatani, Masayoshi. 1991. Grammaticization of topic into subject. In Elizabeth C. Traugott & Bernd Heine (eds.) *Approaches to grammaticalization,* 93–133. Amsterdam: John Benjamins.
Tanaka, Nozomi. 2016. *An asymmetry in the acquisition of Tagalog relative clauses*. Manoa: University of Hawaii dissertation.
Toutanova, Kristina, and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 63–70. https://doi.org/10.3115/1117794.1117802.
Van Valin, Robert D. & Randy J. LaPolla. 1997. *Syntax: Structure, meaning, and function*. Cambridge: Cambridge University Press.

Wouk, Fay. 1986. Transitivity in Batak and Tagalog. *Studies in Language* 10(2). 391–424. https://doi.org/10.1075/sl.10.2.06wou.
Wolfenden, Elmer P. 1971. *Hiligaynon reference grammar*. Honolulu: University of Hawaii Press.
Wolff, John U. 1966. *Beginning Cebuano, Part 1*. New Haven: Yale University Press.