1. Overview

Research on Tagalog has shown that when describing transitive events, an event wherein an entity acts on another entity, speakers exhibit a strong preference for mapping Undergoers (encompassing patients, themes, goals, etc.) instead of Actors (agents, experiencers, causers, etc.) to the privileged syntactic argument function (as indicated by *ang*-marking). However, the role of individual verbs and their co-occurrence patterns with their Actor and Undergoer arguments within these voice structures remains an open question. To what extent do individual verbs prefer mapping Undergoers to the privileged syntactic argument? To what extent does this preference vary across verbs and are potentially verb-specific behaviors modulated by referential properties known to affect Undergoer and Actor voice selection? This study uses corpora methods (Gries & Stefanowitsch 2004; Bresnan et al. 2004; Bybee 2006; Colleman 2009) to examine the Tagalog Undergoer voice preference for frequently occurring semantically transitive verbs (e.g., *bangga* 'bump,' *tawag* 'call,' *tulak* 'push,' etc.). The data were extracted from the tlTenTen 2019 Tagalog web corpus and coded for several morphosyntactic and semantic features. Preliminary results (n = 10 verbs, tokens = 685) suggest that preference for *ang*-marked Undergoers is not monolithic. Each verb exhibits specific patterns of *ang*-marking Undergoers and Actors that vary somewhat per verb and the relative weighting of their arguments' referential features. Furthermore, the contexts for mapping Actors to the privileged syntactic argument appear to be much more highly constrained. These results suggest that complex interactive relationships between these factors (and others) must be examined in order to explain the Undergoer and Actor voice distributions in Tagalog.

2. Background & Significance

2.1. Tagalog Background Information

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.

2.2. THEORETICAL GROUNDING AND TERMINOLOGY

This paper focuses on Tagalog *ang*- and *ng-/sa*- marking[1] on arguments, the co-indexation of those arguments on semantically transitive verbs via voice affixation, and the referential properties of those arguments. 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 two macroroles Actor and Undergoer each subsumes specific semantic relations. 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). Although RRG focuses on mapping of semantic roles to grammatical relations, patterns of semantic features, such as definiteness, animacy, topicality, etc., that more are often associated with these (macro)roles also play a significant role in how semantic roles are mapped to syntactic roles. With respect to Tagalog, *ang*-marked arguments are analyzed as the PSA since it is the only argument co-indexed with the verb and the target of a range of syntactic operations (though non-PSA Actors retain several subject-like properties[2], e.g., Himmelmann 2008; Shibatani 1991; Kroeger 1993; Schachter 1976; 1977; 1995). Likewise, arguments that are marked by *ng* or *sa* will be referred to as the non-privileged syntactic arguments (NPSA). 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):

(1) a.

B<in>ili	ng	guru	**ang**	**libro**
buy<UV>.PRV[3]	**NG**	teacher	**ANG**	**book**

'The teacher bought the book'

b.

B<um>ili	**ang**	**guru**	ng	libro
buy<AV>.PRV	**ANG**	**teacher**	**NG**	book

'The teacher bought a book'

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. I will focus on two constructions, the Actor and Undergoer voice forms which have some analogy to the active/passive alternation in languages like English, but which are functionally very different. 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, which is taken as evidence that Tagalog has a symmetrical voice system, as opposed to an asymmetrical voice system as is seen with the English active/passive patterns (e.g., Latrouite 2011; Schachter 1976, 1977, 1995; Himmelmann 2008).

2.3. TAGALOG UNDERGOER VOICE PREFERENCE

In Tagalog, Undergoers are the preferred privileged syntactic argument (e.g., Cena, 1977; Cooreman et al., 1987; Wouk, 1986; Garcia & Kidd, 2021), which seems contrary to robust cross-linguistic patterns for Actors as PSA (see Riesberg & Primus 2015). The Undergoer voice preference has been seen in narrative text (Katagiri 2005; Wouk 1986; Cooreman et al., 1984: 17), in child speech (Marzan 2013; Garcia et al. 2018), and psycholinguistic experiments (Tanaka et al., 2016). Various semantic-pragmatic factors have been proposed to affect *ang*-marking in Tagalog, including, but not limited to topicality, specificity, definiteness, animacy, and (to some degree) verb semantics. However, apart from Latrouite (2011), little work has explored the Undergoer voice preference with respect to their verbs and their co-occurrence patterns with their arguments' semantic-pragmatic factors.

2.4. CHARACTERISTICS OF *ANG*-MARKED ARGUMENTS

Basic sentences typically have one *ang*-phrase (Schachter & Otanes 1972)[4]. The *ang*-marked argument is understood as the most "prominent" (Latrouite 2011) or "salient" (Wouk 1986) argument. Here, prominence can be understood broadly in terms of marking the argument that has the most relevance to the message or utterance (Latrouite 2011; relevance theory, Sperber & Wilson, 2004). To *ang*-mark an entity is to indicate who or what the verb is about. Generally, argument prominence is measured in multiple ways, including definiteness, animacy, topicality, and others[5], which will be briefly described below. In Tagalog, pronouns and proper names are not marked by *ang, ng,* or *sa*, however they have corresponding forms to the three markers (here, glossed as *ANG, NG, SA*). Though prominence-marking is primarily a pragmatic notion, prominence can be analyzed along some semantic scales that allows us to examine a complex weighting system between features to explain patterns in the data.

DEFINITENESS AND *ANG*-MARKING

Highly definite entities[6], particularly Undergoers, are more likely to be *ang*-marked (Bowen 1965; Schachter & Otanes 1972; Naylor 1975, etc.). The role of definiteness in *ang*-marking seems to be exhibited in other Philippine languages, e.g., Ilokano (Schwartz 1976), Hiligaynon (Wolfenden 1971) and Cebuano (Wolff 1966). Schachter (1976) and Schwartz (1976) proposed that definite Undergoers will be *ang*-marked and in the case where none of the nominals is definite, Tagalog may resort to *ang*-less existential constructions (Schachter & Otanes, 1972). However, Adams & Manaster-Ramos (1988) show that indefinite readings of *ang*-marked nouns are not only possible, but the generally accepted interpretation, when there is an indefinite quantifier such as *isa-ng* ‘one,’ *marami-ng* ‘many,’ or *anuman* ‘anything,’ and others:

(2)

Tawag-an	**ang**	**isa-ng**	pediatric dermatologist	kung	na-pansin
call-**UV.IMP**	**ANG**	**one-LNK**	pediatric dermatologist	**COND**	**UV.PFV**-notice

mo ang anuman-ng naaangkop ng abcde

2SG ANG anything-LNK appropriate NG abcde

‘Call a pediatric dermatologist if you notice something that looks like whatever is labeled in ABCDE (context: pictures associated with different health issues)’ (SketchEngine tlTenTen2019, website: krikids.com)

In 2, only an indefinite interpretation for *ang isang pediatric dermatologist* is possible despite its *ang*-marked status, suggesting that Undergoers, regardless of definiteness can be *ang*-marked. Furthermore, under a discourse-based definition of definiteness, definiteness appears to be weighed differently between the Undergoer 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).

ANIMACY AND *ANG*-MARKING

Animacy, often correlated with definiteness and other referential features, also plays a role in *ang*-marking. Generally, the Actor voice (*ang*-marked Actors) tends to be less acceptable with a human undergoer. For example, examples from Saclot (2006) shows:

(3) a.

K<um>agat	**ang**	**aso**	ng/sa	buto
<AV>bite	**ANG**	**DOG**	NG/SA	bone

'The dog bit a bone'

b.

K<in>agat	ng	aso	**ang**	**buto/si Lena**
<UV>bite	NG	dog	**ANG**	**bone/**ANG **Lena**

'A dog bit the bone/Lena'

c.

??**K<um>agat**	**ang**	**aso**	sa	akin/kay Lena
<AV>bite	**ANG**	**dog**	SA	1SG.SA/SA lena

??'The dog bit me/Lena'

3a,b show that Actor voice and Undergoer voice forms of *kagat* 'bite' are acceptable when the Undergoer is inanimate. 3a shows that Actor voice is acceptable when the Undergoer is inanimate but 3b 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).

PREDICATE-INHERENT ORIENTATION AND *ANG*-MARKING

Slightly less explored are these Undergoer/Actor voice structures with respect to verbs and their arguments. Latrouite (2011, 2016) proposes an analysis that incorporates aspects of the verb's meaning along with the semantic-pragmatic features discussed above to explain asymmetrical patterns of Undergoer and Actor voice patterns ("voice marking gaps" Latrouite 2011). For example:

(4) a.

*P<um>atay	**ang**	**mga**	**bata**	ng	aso
<AV>PFV.kill	**ANG**	**PL**	**child**	**NG**	dog

Intended: 'The children killed a dog.'

b.

P<in>atay	ng	mga	bata	**ang**	**aso**
<UV>PFV.kill	NG	**PL**	child	**ANG**	**dog**

'The children killed the dog.' (cf. Saclot 2005:3)

4a,b contrast with the examples in 3a-c. Even though *ang mga bata* 'the children' has higher animacy and/or definiteness than *aso* 'dog', the Actor voice form of *patay* 'kill' is generally unacceptable (except in certain constructions, e.g., focus construction). Verbs *takot* 'frighten' and *sira* 'break' pattern similarly to *patay.* But either form is acceptable with the verb *suntok* 'hit' even when both arguments are referentially prominent as in 5a,b (Saclot 2006: 10, cited from Latrouite 2011):

(5) a.

S<um>untok	si	Pedro	kay	Jose
<AV>PFV.hit	ANG	Pedro	SA	Jose

'Pedro hit Jose'

b.

S<in>untok	ni	Pedro	si	Jose
<UV>PFV.hit	NG	Pedro	ANG	Jose

'Pedro hit Jose'

These examples suggest that in addition to definiteness and animacy, an argument can be measured based on its prominence in the predicate structure, or its centrality to the predication (Latrouite 2011 p. 194). Undergoer-oriented verbs such as *patay* 'kill,' or *takot* 'frighten,' *sira* 'destroy,' etc., highlight the state of the Undergoer compared to the Actor (Latrouite 2011), increasing its likelihood of occurring in the Undergoer voice. By contrast, activity verbs such as *kain* 'eat' or *sulat* 'write,' which allow for incremental interpretations with individuated Undergoers, describe activities that profile information about the Actor instead of the Undergoer. This increases the likelihood that the Actor will be the PSA. Punctual contact verbs like *suntok* 'hit,' *kagat* 'bite,' and others, which denote punctual contact between the Actor and Undergoer may have no clear predicate-inherent focus of attention and thus might occur readily in either voice form. The choice between Actor and Undergoer voice might then come down to prominence features.

Given the prior research, Undergoer and Actor voice structures are a result of a complex interplay between verbs and the referential properties of their arguments. The current corpus study is a first attempt at examining the extent to which the Undergoer voice preference is informed by these factors across a range of verbs and large samples.

3. METHODS

3.1. DATA

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 198,303,250 words (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). 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. The resulting concordance searches were pre-processed by the author to ensure the token was a valid sample of the target verb prior to annotation. The window for each extraction included two

to three sentences preceding and following each token to provide some minimum context for the token.

3.2. Verb Selection

Latrouite (2011) provides only a few verbs in her predicate-inherent orientation categories, so most verbs examined here are not a priori categorized. Verbs were selected based on their corpus frequency and their *potential* for denoting causatively transitive actions (*kain* 'eat' was included here). The top 1000 most frequent verbs were extracted, and then causatively transitive verbs were chosen for analyses. That is, selected verbs denoted actions that had causation and affectedness (e.g., Hopper & Thompson, 1980) and which had potential to take (at least) two participants as arguments in a clause. If they met the previous criteria and were also analyzed in Latrouite (2011), they were also extracted. Valid verbs under these criteria included *bangga* 'bump,' *karga* 'carry,' *buhat* 'lift,' *patay* 'kill,' *kain* 'eat,' and others, and excluded highly frequent experiencer-theme verbs such as *kita* 'see,' *sabi* 'say,' and others.

3.3. Annotation Scheme

Each token was annotated for a variety of features with respect to the verb and the PSA and NPSA. Given how impoverished discourse-pragmatic information can be in this kind of corpora, important discourse-pragmatic features like topicality and specificity could not be coded for. Instead, the author coded for related factors such as definiteness and animacy which can exhibit explicit morphosyntactic coding in minimal context. An abridged annotation scheme is shown below:

> **Verb Voice Affixes:** voice affixes (if they existed since verbs can be bare) were coded for and co-indexed with the *ang*-marked argument. Actor voice affixes included but were not limited to *-um-*, *mag-*, *maka-*, and others. Undergoer voice affixes included *-in-*, *-an*, *i-*, and others
>
> **Macroroles** Actor and Undergoer roles were assigned to the *ang*, *ng*, and *sa* arguments (if present) based on their role in the sentence.
>
> **Definiteness** was 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). Definiteness codes were broadly based on a typological definiteness scale: Personal pronoun > Proper name > Definite NP > Indefinite specific NP > Non-specific NP (e.g., Aissen 2003). Arguments were coded as "Definite" if they were a personal/demonstrative pronoun, proper name, discourse-old (Bhatia et al., 2014), syntactically individuated as heads of relative clauses, NPs marked by certain quantifiers, etc. (Wouk, 1986). Arguments were marked as "Indefinite" if they were common nouns that were new in the context, preceded by quantifiers such as *isang* 'one,' or *kahit* 'any,' and others. Arguments were marked as "Other" if the definiteness values could not be determined or if that argument did not exist.

Animacy was coded following a typological animacy hierarchy: Human >Animate> Inanimate > Abstract (e.g., Primus, 1999; Aissen, 2003).

Hierarchical values for the referential properties allow us to calculate relative weights of those features between arguments and derive different measures of "prominence." Furthermore, taking the verb and clausal properties into account with these weights provides us a way of understanding how these features might inform a verb's occurrence in Actor and Undergoer voiced structures based on the semantic properties of their arguments.

4. PRELIMINARY RESULTS

A total of ten verbs and 685 tokens were annotated and analyzed here. Figure 1 shows the proportions of occurrence for Undergoer (green bars) and Actor voice (blue bars) and "Other" uses (yellow bars) for each verb in the sample. The "Other" category encompassed infrequent voice forms (e.g., *ang*-marked locations or instruments), forms where the *ang*-entity was a reciprocal pronoun (e.g. *kita* 1SG.2SG pronoun), when the clause was *ang*-less, or if the clause had double-*ang* arguments. In general, in line with the prior research, *ang*-marked Undergoers were more frequent than Actors in the entire sample (green bars, 50.7% of all usages compared to 27.8% of all usages).

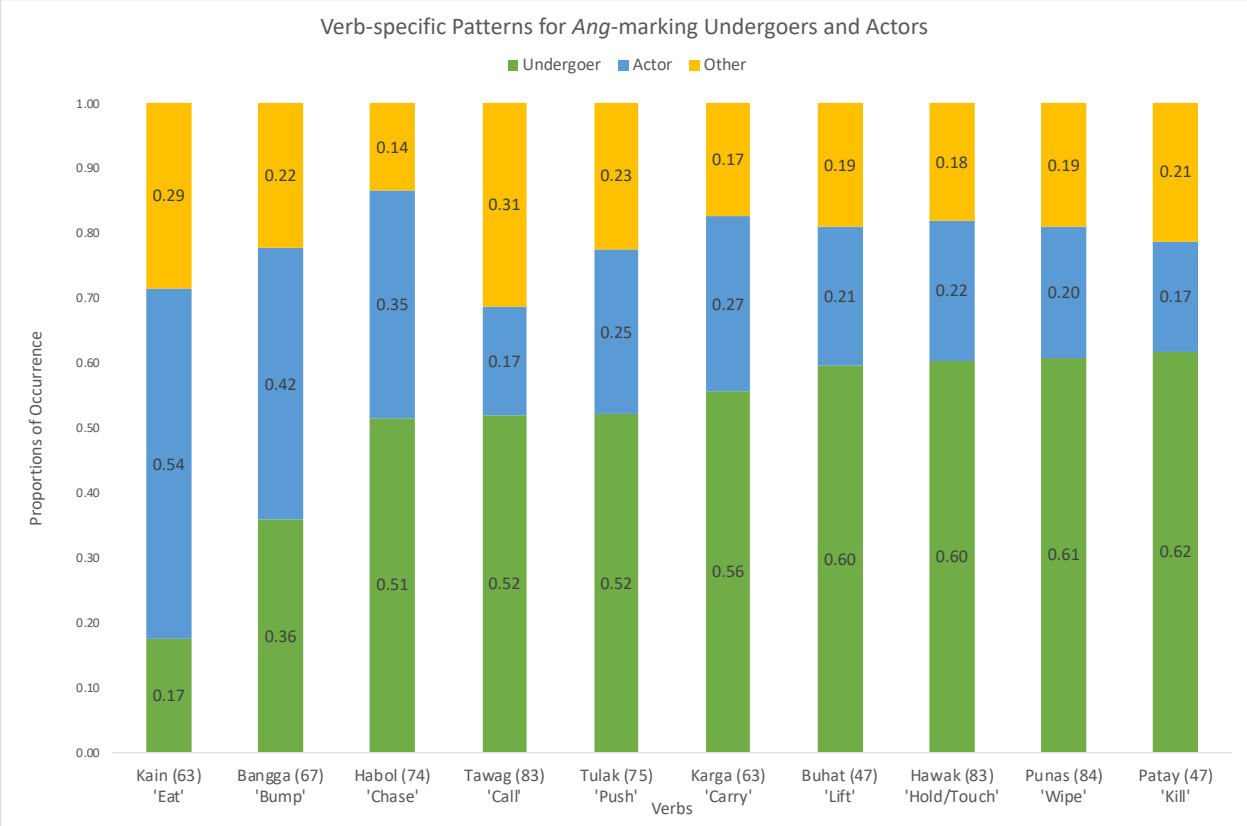

Figure 1. Verb-specific Patterns for Ang-marking Undergoers and Actors

Eight of ten verbs, *habol* 'chase,' *tawag* 'call,' *tulak* 'push,' *karga* 'carry,' *buhat* lift,' *hawak* 'hold/touch,' *punas* 'wipe,' and *patay* 'kill,' occurred between 50-60% of the time in the Undergoer voice. In comparison, the verbs *kain* 'eat' and *bangga* 'bump' tend to have more *ang*-marked Actors (54% and 42% respectively). The verb *kain* occurred in both intransitive and transitive uses, which contributes to the high proportion of *ang*-marked Actors. The result for *kain* provides evidence for Latrouite's (2011) analysis that *kain* is more Actor-oriented as an incremental activity verb. The verb *bangga* 'bump' appears to have less of a preference for *ang*-marking Undergoers compared to the other verbs. This verb might pattern similarly to *suntok* 'hit,' since *bangga* denotes a meaning of surface contact. *Bangga*'s distribution between Actor and Undergoer voice structures might provide support for verbs of surface contact exhibiting "neutral" predicate-orientation. Broadly speaking, there appears to be relatively little variability between verbs in how they are used in the Undergoer and Actor voice sentences. The following sections will further examine the extent to which the relative features of definiteness and

animacy have an influence on an individual verb's distributions in Undergoer and Actor voice structures.

4.1. Relative Definiteness

Relative Definiteness was calculated by comparing the definiteness feature between the *ang*-argument and the *ng*- or *sa*- argument if present. If there was no explicit *ng-/sa*-entity, they were coded as being absent and having a value akin to "0" in the weight calculations. To better understand how each verb might be affected by relative definiteness and its role in *ang*-marking Actors versus Undergoers, the data was further separated by the proportions of relative definiteness on Undergoer voice (Figure 2) and Actor voice (Figure 3) per verb.

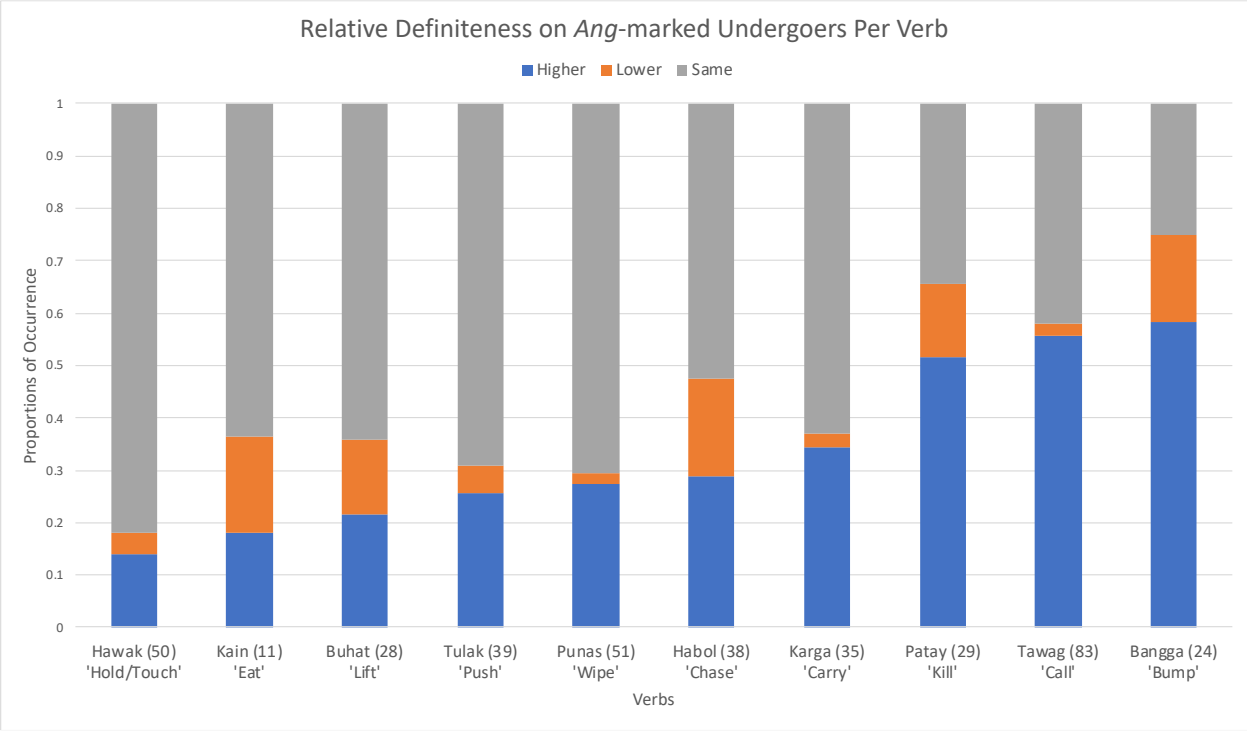

Figure 2. Relative Definiteness on *Ang*-marked Undergoers per Verb

Figure 2 shows the proportions of occurrence of Undergoer voice structures where the PSA-Undergoer had higher (blue), lower (orange), or equal (gray) definiteness to the Actor. Generally, Undergoers and Actors tended to be relatively equal in definiteness across verbs, but there is variation among the verbs in how often the Undergoer was higher or lower in

definiteness than the actor. Verbs like *patay* 'kill', *tawag* 'call', and *bangga* 'bump' often appear with Undergoers that are higher in definiteness compared to the Actor. The event-semantics of *patay* 'kill' suggests that it may exhibit a natural preference towards the Undergoer voice regardless of the referential status of its arguments. This is demonstrated in example 6 below:

(6)

Dapat	**patay-in**	na	rin	yan
Should	**kill-UV.IRR**	**LNK**	also	**DEM.ANG**

'That one (previously mentioned criminal) should also be killed'

Example 6 demonstrates an instance of Actor absence in the Undergoer voice. Actor dropping can sometimes result from Actor being relevant to a context such that they do not need to be referred to further. In this example, the Actor is not so much "dropped" as it is just irrelevant to refer to in this context. The main focus or the most prominent entity is the entity to be killed, indicated by the *ang*-form demonstrative pronoun *yan* along with the Undergoer voice marking on the verb *patay-in*.

While *bangga* also occurs with Undergoers higher in definiteness, the pattern for *bangga* 'bump' differs from what we see with *patay*. Whereas *patay* exhibits an Undergoer preference and argument definiteness seems to reinforce that pattern, *bangga* has a slight preference for the Actor voice (Figure 1). When it does appear in the Undergoer voice, there are many examples of the Undergoer with higher definiteness than the Actor:

(7)

Na-bangga	ng	6 wheeler truck	**ang**	**tricycle**	na
UV.PFV-bump/hit	**NG**	6 wheeler truck	**ANG**	**tricycle**	**REL**

m<in>a~maneho	ng	biktima-ng	si	Clemente Enerio	ng	Antipolo
IPFV<UV>drive	**NG**	victim-**LNK**	ANG	Clemente Enerio	GEN	Antipolo

'The tricycle that was being driven by the victim, Clemente Enerio of Antipolo, was bumped/hit by a 6-wheeler truck'

In example 7, the Undergoer *ang tricycle* has higher definiteness given its further elaboration through the additional relative clause *na minamaneho ng biktimang...* The nature of the Undergoer having higher definiteness in these *bangga* cases often occur due to some relativization process that provides further elaboration and emphasizes the importance of the

Undergoer in these instances (Wouk 1986). Because *bangga* may be a neutral verb, we might speculate that the role of relative definiteness plays a slightly more important role for Undergoer uses of *bangga* compared to *patay*. However, the presence of Undergoers with lower and equal definiteness to the Actors suggest that this is not the entire story for *bangga*.

That a more definite entity would be *ang*-marked is not surprising in and of itself. More interesting are the patterns for the verbs where Undergoer and Actor definiteness are equal. Across verbs like *hawak* 'hold/touch,' *buhat* 'lift,' *tulak* 'push,' *punas* 'wipe,' and a couple others, Undergoers and Actors often were equal in definiteness (gray bar). Both arguments were generally referred to using pronouns, possessed body parts, personal names, or descriptive NPs in situations where there was physical contact between the arguments:

(8)

H<in>awak-an	ni	Clyde	**ang**	**aki-ng**	**kamay**
<REAL>hold-UV	**NG**	Clyde	**ANG**	**1GEN-LNK**	**hand**

'Clyde held my hand'

Given that these verbs denote physical contact but not necessarily result-oriented action, we might analyze these verbs as having less of a predicate-inherent orientation and expect Undergoers to have higher definiteness when the verb occurs in Undergoer voice. Instead, the general Undergoer voice preference despite the equal weights potentially suggests that other factors may affect Undergoer uses of these verbs or that the Undergoer voice takes precedence over argument referential properties and predicate semantics. The results in Figure 2 suggest that these verbs generally exhibit a preference for the Undergoer voice and that relative definiteness may be a factor, but not always a defining factor of Undergoers for these verbs.

The co-occurrence patterns of relative definiteness in the Undergoer voice contrasts heavily with the Actor Voice (Figure 3). Except for *punas* 'wipe,' *ang*-marked Actors almost always have higher definiteness compared to the Undergoer. 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 Actor *ang*-marking may rely more on the Actor having higher definiteness than the Undergoer in comparison to the Undergoer voice (Wouk 1986).

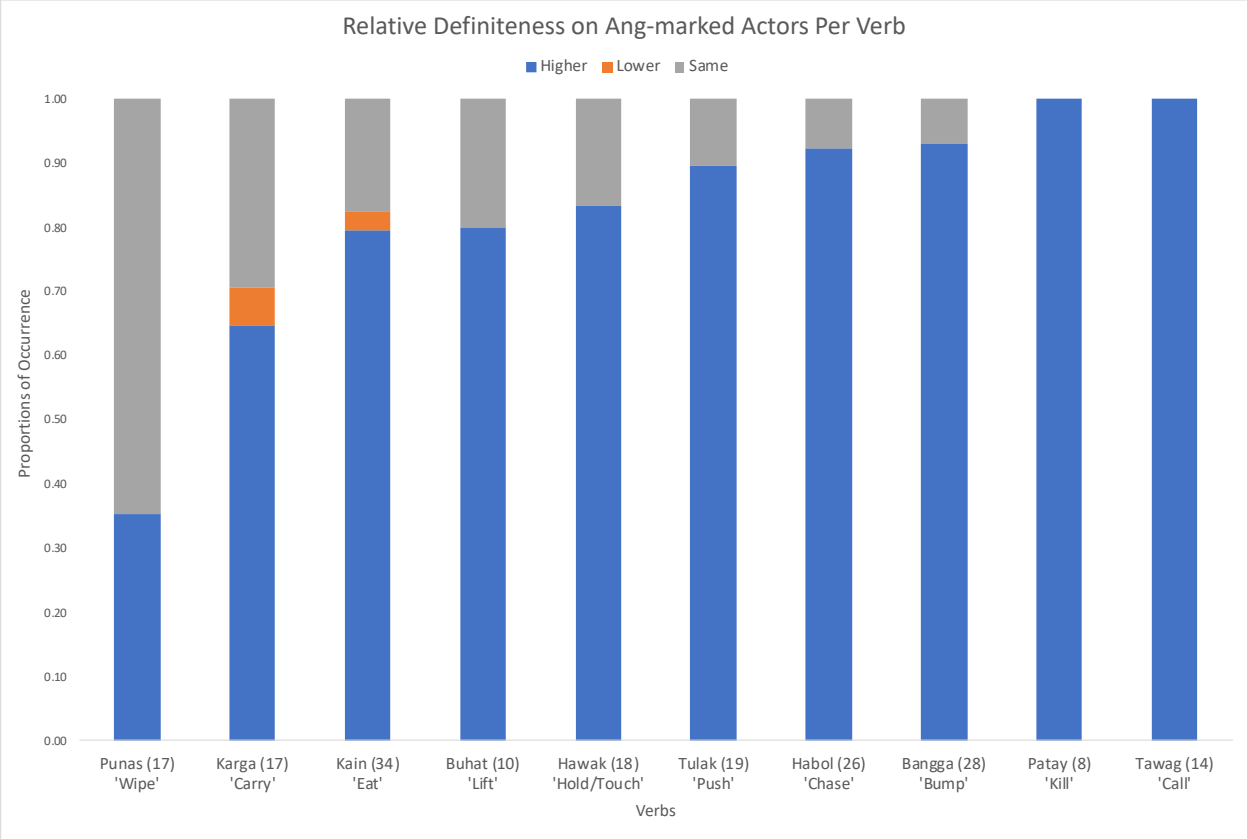

Figure 3. Relative Definiteness on *Ang*-marked Actors Across Verbs

The verb *punas* 'wipe' appears to be the primary exception to this. Many of the Undergoers in these instances were often body parts that were implicitly co-referential with the Actor:

(9)

Nag-punas	siya	ng	bibig
AV-wipe	3.ANG	NG	mouth

'She wiped (her) mouth'

Both Actors and Undergoers in such examples exhibit equal definiteness values. But as we saw in Figure 1, *punas* has a strong preference for Undergoer voice relative to Actor voice, and in Figure 2 looking at the relative definiteness values, *punas* tends to have *ang*-marked Undergoers even when relative definiteness was equal between the arguments. There are too few samples to draw strong conclusions, however, this may suggest that relative definiteness is not as strong a differentiating factor between *ang*-marking Undergoers and Actors for *punas*.

In sum, the role of relative definiteness for *ang*-marking Undergoers (Figure 2) seems to be variable across verbs. By contrast, higher relative definiteness on Actors is generally the default when the verb is used in the Actor Voice. Certain verbs, such as *punas* 'wipe' may exhibit verb-specific behaviors that do not follow this.

4.2. Relative Animacy

Relative animacy was calculated by comparing the animacy features (Human, Animate, Inanimate, Abstract) between the PSA (*ang*-marked) and the NPSA (*ng* or *sa*-marked) if present. Figure 4 shows the relative animacy values for verbs in Undergoer voice (*ang*-marked Undergoers).

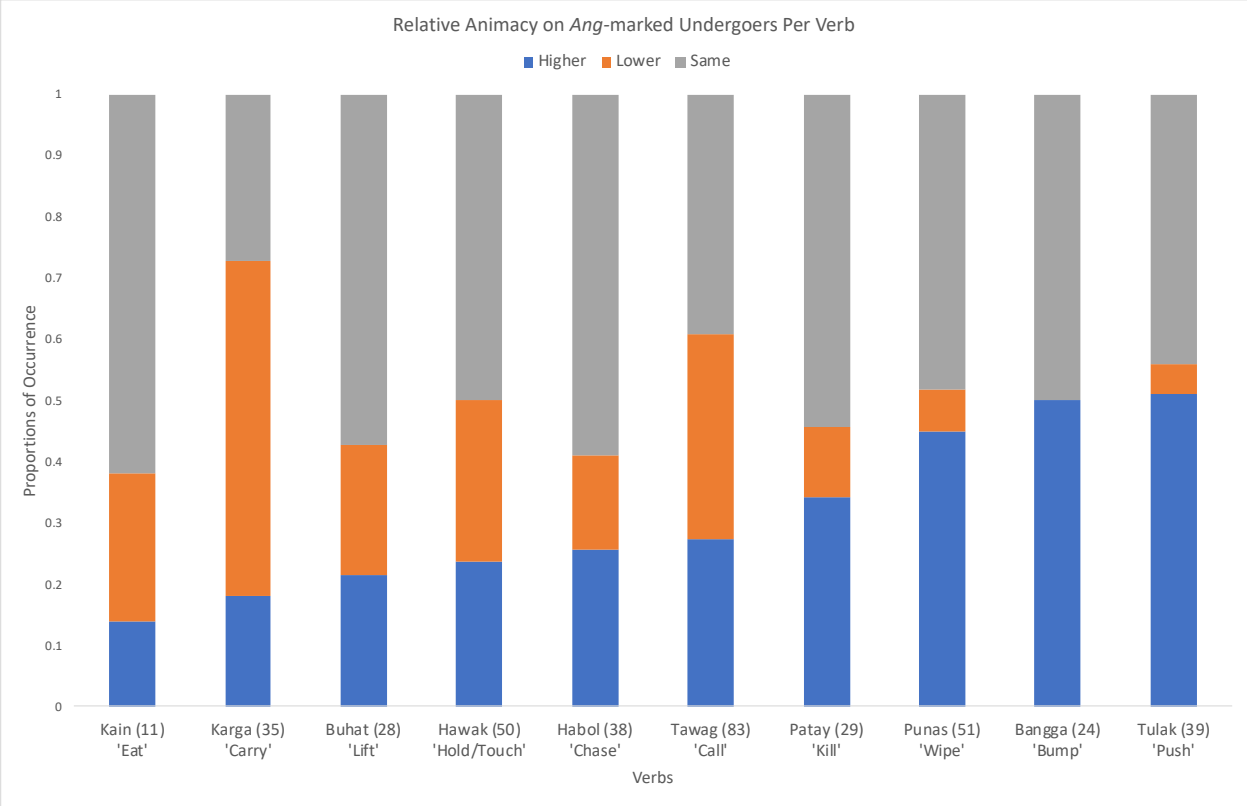

Figure 4. Relative Animacy on *Ang*-marked Undergoers Across Verbs

Figure 4 shows that though Undergoers and Actors were often equal in animacy, there is variation across verbs much like relative definiteness on Undergoer voice marking. The few instances of Undergoer voice *kain* 'eat' initially seem odd. How could the Undergoer (eatee) have

equal or higher animacy compared to the Actor (eater)? Upon further inspection, these were not examples of cannibalism. Instead, situations of intimacy used *kain* less literally where the Undergoer was often a body part like *labi* 'lips' (coded as human). Other examples of when the Undergoer was of higher animacy than the Actor was when the Actor was non-referential:

(10)

G<um>awa	ng	sabaw	ng	karne	kapag	kailangan	lamang	sa
<ACT>make	NG	soup	GEN	meat	ADV.when	need	ADV.only	SA

dahil(a)-ng	masarap	ito-ng	**kain-in**	n(an)g	bagong-luto
because-LNK	tasty	DEM.ANG-LNK	**eat-UV.NFIN**	ADV	newly-cooked

'Make some meat soup when needed because it is delicious to eat it when it is freshly cooked'

10 shows *kain-in* used with *ito-ng*, a demonstrative pronoun which refers to the previously mentioned *sabaw ng karne* 'meat soup'. Although *kain* predicate-inherently profiles the Actor, when it is absent as in 10, the Undergoer must be the most prominent argument. The verb *karga* 'carry' seems to pattern differently from the other verbs. That is, there is a high proportion of *ang*-marked Undergoers that are lower in animacy compared to the Actor:

(11)

Matapos	**i-karga**	ang	gamit	ko	sa	kotse
Finished	**UV-carry**	**ANG**	**things**	1.**GEN**	**SA**	car

nag-paalam	na	ako	kay	Dave
AV-goodbye	**ADV**	**1SG.ANG**	**3SG.SA**	Dave

'After I carried my things to the car, I said goodbye to Dave'

Karga denotes an action wherein the object changes location and thus might be understood as affected. Since carrying objects is a common occurrence, Undergoers lower in animacy are not surprising and may potentially reinforce the Undergoer voice preference.

Figure 5 shows the relative animacy of *ang*-marked Actors compared to the Undergoer if present.

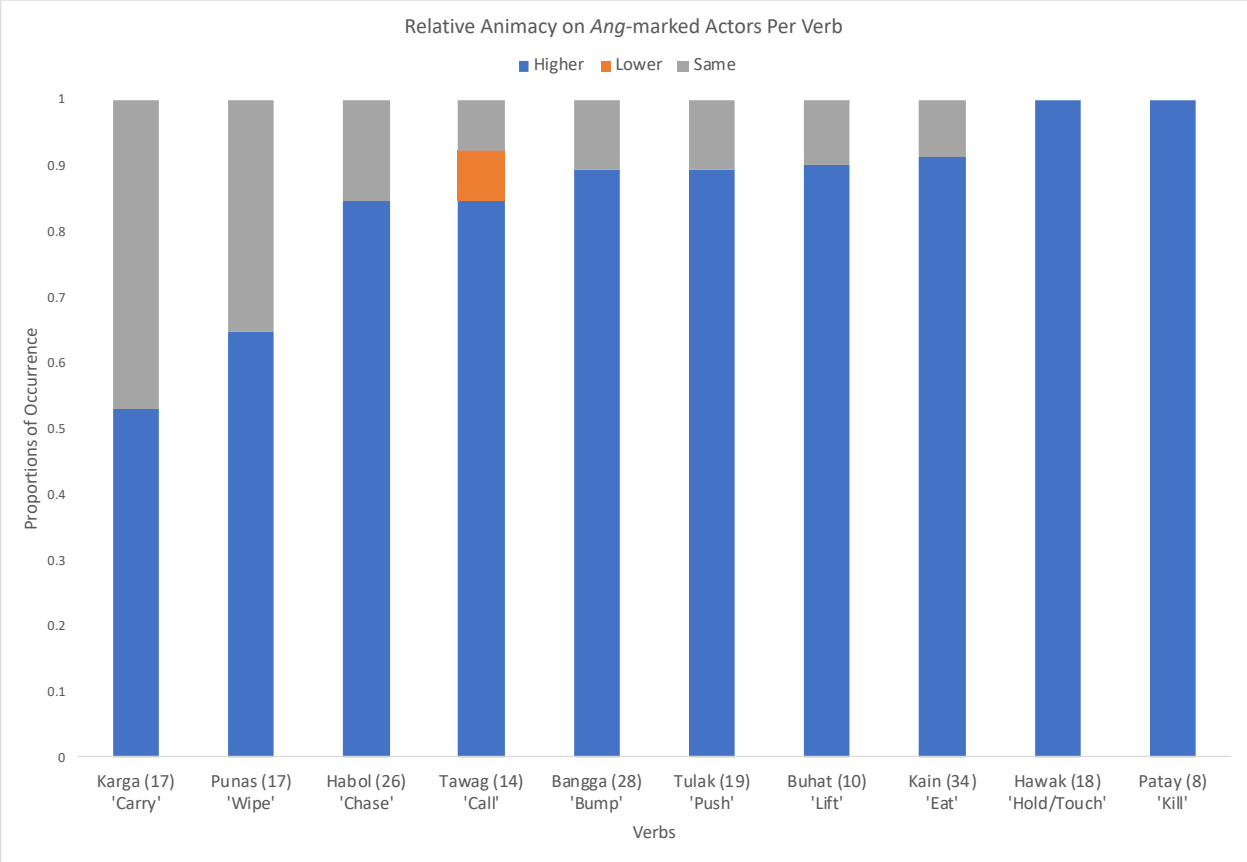

Figure 5. Relative Animacy on *Ang*-marked Actors Across Verbs

Like relative definiteness on *ang*-marked Actors, there is far less variation in the role of relative animacy of Actors compared to Undergoers in Actor voice sentences. Actors tended to have higher (or at least equal) animacy values compared to Undergoers across verbs, suggesting again that the Actor voice appears to be more constrained such that Actor Voice usage tends to occur when Actors are higher in prominence. This may vary with verb. The pattern for the verb *karga* 'carry' exhibits a complementary pattern to what we saw in the Undergoer voice. That is, in the Actor Voice, Actors frequently have equal animacy to Undergoers and higher animacy does not appear to be the defining factor for *ang*-marked Actors. This pattern along with what we see in the Undergoer voice for *karga* (Figure 4) provides further evidence that relative animacy may not have as strong a relationship with *karga* uses.

The Actor voice uses of *patay* 'kill' bear examining. As previously mentioned, under an event-structural analysis, *patay* highlights the resultant state of the Undergoer which contributes to its preference for the Undergoer voice. The Actor voice examples found here show that when *patay* occurs in the Actor voice, those Actors have high definiteness (Figure 4) and animacy (Figure 5). However, it may be more accurate to characterize these occurrences as when the Undergoer is non-referential (and therefore not definite/animate):

(12)

…kapag	**naka-patay**	**ka**	para	proteksyunan
ADV	ABIL.AV.PFV-**kill**	2SG.ANG	PREP	protection

ang	sarili	mo	o	ang	pamilya	mo,	that is just
ANG	self	2SG.GEN	CONJ	ANG	family	2.GEN	that is just

'…when you've been able to kill in order to protect yourself or your family, that is just'

12 shows the Actor *ka* 'you' as the only argument to the verb. There is no Undergoer explicitly or implicitly mentioned in this example. This is further co-indexed by the *maka-* Actor voice affix on *patay*. We see another example of Undergoer absence in 13:

(13)

Oo	hindi	cla (sila)	**p<um>a~patay**	pero	may
Yes	NEG	3PL.ANG	IPFV<UV>~**kill**	but	EXIST

go signal	cla (sila)
go signal	3PL.ANG

'Yes, they do not kill, but they have a go signal…'

Again, the only argument to *patay* is the pronoun *cla*[7] (*sila*) 'they,' which is co-indexed by the Actor voice affix *-um-*. Although *patay*'s semantics licenses and perhaps profiles the Undergoer, the absence of an Undergoer in these examples suggests a pattern similar to English existential null complementation constructions where the patient argument of an accomplishment verb is omissible (Goldberg 2005; Michaelis 2006). What this shows is that *patay* exhibits highly specific behaviors in the Actor voice that involve referential features as well as larger constructional patterns. In sum, these results show that relative animacy may play a greater role in the Actor voice compared to the Undergoer voice across verbs. Furthermore, each verb's event

semantics interact with the relative animacy between its arguments, though there is greater variation in its co-occurrence with Undergoer voice compared to the Actor voice.

5. Discussion & Conclusions

This corpus study provided a preliminary understanding of how verbs and the referential features of their co-occurring arguments interact to produce verb-specific patterns with respect to the Undergoer voice and the Actor voice. The data showed that across most verbs, there was a preference for the Undergoer voice. An examination of the Undergoer voice and Actor voice show that relative definiteness and animacy of the verbs' co-occurring arguments is more indicative of the patterns for Actor voice compared to the Undergoer voice. That is, *ang*-marked Actors occur more frequently when Actors are more prominent in terms of definiteness and animacy compared to the Undergoer. Otherwise, the Undergoer voice appears to be the default for most verbs, which is in line with prior work on Tagalog (Latrouite 2011; 2016; Wouk 1986; Katagiri 2005; Cooreman, Fox, & Givón, 1984: 17). However, an examination of each verb's patterns of occurrence in both voices and the semantic properties of their arguments showed that each verb exhibited its own behaviors within these broader general patterns. In sum, there is a complex interplay between verbs, their semantics, the referential properties of their arguments, and other factors which shed light on these distributions of Undergoer and Actor voice structures.

This study is just a start, but many more verbs and samples must be annotated and analyzed to better understand these patterns. Furthermore, *ang*-marking is likely influenced by other factors, such as different construction types (focus constructions, e.g., Latrouite 2011; null complementation; other constructions, Garcia & Kidd 2021) and of course, discourse factors. An important note here is that these results reflect co-occurrence patterns and not patterns of causation. That is, the presence of one factor with a certain voice marking shows that they happen to co-occur together. We cannot establish causal or directional links between these features and voice marking. If we are to better understand a causal link between these factors and what "triggers" how these forms are used, other methods such as experiments, must complement the corpus methods.

References

Adams, K. L., & Manaster-Ramer, A. (1988). Some Questions of Topic/Focus Choice in Tagalog. *Oceanic Linguistics*, *27*(1/2), 79. https://doi.org/10.2307/3623150

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

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (n.d.). *Predicting the Dative Alternation*. 33.

Bybee, J. L. (2006). From Usage to Grammar: The Mind's Response to Repetition. *Language*, *82*(4), 711–733. https://doi.org/10.1353/lan.2006.0186

Cena, R. M. (1977). Patient primacy in Tagalog. *LSA Annual Meeting, Chicago*, 28–30.

Colleman, T. (2009). Verb disposition in argument structure alternations: A corpus study of the dative alternation in Dutch. *Language Sciences*, *31*(5), 593–611. https://doi.org/10.1016/j.langsci.2008.01.001

Cooreman, A., Fox, B. A., & Givón, T. (1984). The Discourse Definition of Ergativity. *Studies in Language*, *8*(1), 1–34. https://doi.org/10.1075/sl.8.1.02coo

Foley, W., & Van Valin Jr, R. (1977). On the organization of "subject" properties in universal grammar. *Annual Meeting of the Berkeley Linguistics Society*, *3*, 293. https://doi.org/10.3765/bls.v3i0.3297

Garcia, R., Roeser, J., & Höhle, B. (2018). Thematic role assignment in the L1 acquisition of Tagalog: Use of word order and morphosyntactic markers. *Language Acquisition*, 1–27. https://doi.org/10.1080/10489223.2018.1525613

Goldberg, Adele. 1995. Constructions: A Construction Grammar Approach to Argument Structure. Chicago: University of Chicago Press.

Gries, S. Th., & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on `alternations'. *International Journal of Corpus Linguistics*, *9*(1), 97–129. https://doi.org/10.1075/ijcl.9.1.06gri

Himmelmann, N. P. (2008). Lexical categories and voice in Tagalog. *Voice and Grammatical Relations in Austronesian Languages*, 247–293.

Hopper, P. J., & Thompson, S. A. (1980). Transitivity in Grammar and Discourse. *Language*, *56*(2), 251–299. https://doi.org/10.1353/lan.1980.0017

Katagiri, M. (2005). Voice, ergativity, transitivity in Tagalog and other Philippine languages: A typological perspective. *The Many Faces of Austronesian Voice Systems. Pacific Linguistics*.

Kroeger, P. R. (1993). *Another look at subjecthood in Tagalog*. 15.

Latrouite, A. (2011). *Voice and Case in Tagalog: The coding of prominence and orientation* [Dissertation]. https://d-nb.info/106308511X/34

Latrouite, A. (2016). Shifting Perspectives: Case Marking Restrictions and the Syntax-Semantics-Pragmatics Interface. In J. Fleischhauer, A. Latrouite, & R. Osswald (Eds.), *Explorations of the Syntax-Semantics Interface* (pp. 289–318). De Gruyter. https://doi.org/10.1515/9783110720297-011

Marzan, Jocelyn C. 2013. Spoken language patterns of selected Filipino toddlers and preschool children. Diliman, Quezon City: University of the Philippines Diliman dissertation.

Michaelis, L. (2006). Complementation by Construction. *Annual Meeting of the Berkeley Linguistics Society*, *32*(1), 529. https://doi.org/10.3765/bls.v32i1.3461

Nagaya, N. (2006, January). Topicality and reference-tracking in Tagalog. In *9th philippine linguistics congress. Quezon City: University of the philippines diliman*.

Naylor, P. B. (1975). Topic, Focus, and Emphasis in the Tagalog Verbal Clause. *Oceanic Linguistics*, *14*(1), 12. https://doi.org/10.2307/3622792

Reid, Lawrence and Paul Schachter. "Tagalog." In The World's Major Languages (2nd edition), edited by Bernard Comrie, chapter 47. London: Routledge, 2008.

Riesberg, S., & Primus, B. (2015). Agent prominence in symmetrical voice languages. *STUF - Language Typology and Universals*, *68*(4). https://doi.org/10.1515/stuf-2015-0023

Saclot, M. J. (2006). On the transitivity of the actor focus and patient focus constructions in Tagalog. *Tenth International Conference on Austronesian Linguistics, Palawan, Philippines, January*, 17–20.

Sauppe, S. (2017). Word Order and Voice Influence the Timing of Verb Planning in German Sentence Production. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01648

Sauppe, S., Norcliffe, E., & Konopka, A. E. (n.d.). *Dependencies First: Eye Tracking Evidence from Sentence Production in Tagalog*. 6.

Schachter, Paul. 1976. 'The subject in Philippine languages: Topic, Actor, Actor-Topic, or none of the above.' In Li, Charles (ed.) *Subject and Topic*. New York: Academic Press, 493-518.

Schachter, P. (1977). Reference-Related and Role-Related Properties of Subjects. *Grammatical Relations*, 279–306. https://doi.org/10.1163/9789004368866_012

Schachter, Paul. 1996. 'The subject in Tagalog: still none of the above.' *UCLA Occasional Papers in Linguistics* No.15. Los Angeles: University of California.

Schachter, Paul & Fe Otanes. 1972. *Tagalog Reference Grammar*. Berkeley: University of California Press.

Shibatani, Masayoshi. 1991. 'Grammaticization of topic into subject.' In Traugott, Elizabeth C. & Bernd Heine (eds.) *Approaches to Grammaticalization*, vol.1. Amsterdam/Philadelphia: John Benjamins, 93-133.

Tanaka, N. (2016). *AN ASYMMETRY IN THE ACQUISITION OF TAGALOG RELATIVE CLAUSES*. 177.

Valin, R. D. V. (2009). *A Brief Overview of Role and Reference Grammar*. 30.

Valin, R. D. V. (2005). *A Summary of Role and Reference Grammar*. 30.

Van Valin, R. D. & LaPolla, R. J. (1997). *Syntax: Structure, meaning, and function*. Cambridge University Press.

Wouk, F. (1986). Transitivity in Batak and Tagalog. *Studies in Language*, *10*(2), 391–424. https://doi.org/10.1075/sl.10.2.06wou

Wolfenden, E. P. (2019). *Hiligaynon reference grammar*. University of Hawaii Press.

Wolff, J. U. (1966). BEGINNING CEBUANO, PART 1. YALE LINGUISTIC SERIES, 9.

ENDNOTES

[1] There is some controversy in how to understand and gloss these markers. They've been variously understood as case markers (Latrouite 2011), *ang* as a "topic marker" (Cooreman et al., 1984), "trigger" (Wouk 1986), as well as subject/object, etc. Because the analysis of what the markers are is not a main issue in this paper, and there is not consensus on how to understand these markers, I will attempt to stay close to the language phenomena and just refer to them as *ang*, *ng*, and *sa* marking.

[2] In Tagalog, different subject-like behavioral properties (Keenan, 1976) are distributed between the Actor and the PSA (*ang*-marked argument). For example, the NPSA (non-*ang*-marked) Actor retains many subject-like properties, such as reflexive binding, control of an actor gap in the second coordinated clause, deletion in imperatives, deletion in the second coordinated clause, and control of a gap in subordinated clauses (Schachter, 1977; Shibatani, 1991; Kroger, 1993; Shibatani, 2005; Latrouite, 2011). On the other hand, Undergoer PSA arguments show several subject properties such as verb agreement, extractability, control of floating quantifiers and gaps in *samptan* 'while' clauses (Shibatani, 1991). These behaviors have resulted in Tagalog being classified as varying systems, including, but not limited to, a "focus" system (e.g., Schachter & Otanes, 1972; Schachter, 1976; Naylor, 1995, etc.) or a "trigger" system (e.g., Schachter, 1976; Fox, 1982; Wouk, 1986).

[3] Abbreviations: 1 first person, 2 second person, 3 third person, ABIL abilitative, ADV adverb, ANG ang marker/form, AV actor voice, CONJ conjunction, EXIST existential, GEN genitive, IPFV imperfective, IRR irrealis, LNK linker, NEG negation, NFIN non-finite, NG ng marker/form, PFV perfective, PL plural, PREP preposition, REL relativizer, REAL realis, SA sa marker/form, SG singular, UV undergoer voice

[4] There are some structures that have double-*ang* or only *ng*-marking, which are beyond the scope of this paper.

[5] The nature of the corpus data does not allow for topicality to be reliably measured here, but I would be remiss in not briefly mentioning the literature around *ang*-marking and topicality. The concept of "topic" and "focus" have referred to different functions in Tagalog, resulting in varying analyses of the *ang*-argument, only a couple of which I will cover here. Nagaya (2006) defined a topical referent as a complex feature that denotes an "animate participant and/or an S or A core argument which tends to be referred to by a pronoun" and a non-topical referent as an "inanimate participant and/or an O core argument" which tends to be marked by zero anaphora (Nagaya, 2006, p. 6). Cooreman et al., (1984) measured topicality by looking at referential distance, topical persistence, and deletability. The researchers found that in "transitive" -in- clauses (i.e., Undergoer voice) patient arguments had much lower topicality compared to agents in terms of anaphoric and cataphoric behaviors. That is, the *ng*-marked Actor argument was shown to be more topical than the *ang*-marked Patient argument. In sum, the role of topicality on *ang*-marking is complex and depending on how topicality is defined, may result in different analyses of *ang*-marked and *ng*-marked arguments.

[6] Although the notions of specificity and definiteness are formally separate features, in the Tagalog linguistics literature, the two features have been used nearly interchangeably.

[7] The form *cla* to mean the pronoun *sila* is common in internet usage for this pronoun.