Frequency influences on rule application within and across words[1]

Kie Zuraw
University of California, Los Angeles

1 Introduction

In morphologically complex environments, whether phonology applies can depend on how “unitlike” a sequence is. A well-known word-internal example from English concerns stress-shifting versus stress-neutral suffixes. The stem *párent* shifts its stress to the closed penult in *parént-al*, but not in *párent-hood*. It appears that in *parént-al*, the whole word is the domain of stress assignment, but in *párent-hood* only the stem is. This difference can be analyzed derivationally, with stress being assigned after *–al* suffixation and before *–hood* suffixation (e.g., in the lexical-phonology framework of Pesetsky 1979, Kiparsky 1982, Mohanan 1986, and others); it can be analyzed by assigning a domain of rule/constraint application such as a prosodic word to the whole word in *(parént-al)* but to the stem alone[2] in *(párent)-hood* (e.g., in the prosodic-phonology framework of Selkirk 1980, Nespor & Vogel 1986, Hayes 1989); it can also be analyzed through output-output correspondence, with a strong requirement that the stem in *parent-hood* bear the same stress as *parent* in isolation, but a weak (or no) requirement that the stem in *parent-al* bear the same stress as *parent* in isolation (e.g., in the output-output correspondence approach of Burzio 1996, Kenstowicz 1996, Benua 1997, Crosswhite 1998, Steriade 2000). An example across a word boundary in English concerns optional *v*-deletion in sequences of verb+clitic. In the sequence *gave me*, *v*-deletion is likely (*he ga[v/Ø] me that*), but in *wove me*, it seems unlikely (*he wo[v/??Ø] me that*). (See Hayes 1989 and Peperkamp’s 1997 discussion). In *gave me*, the *v*’s behavior is sensitive to the presence of the following *m* (assuming that this preconsonantal position is the conditioning environment for *v*-deletion), but in *wove me*, it is not.

We could view *parental* and *gave me* as being more unitlike than *parenthood* and *wove me* respectively. What is a “unit” in lexical, grammatical, and processing terms? Some researchers, such as Bybee (2001 and elsewhere), view units as stored lexical items, which may be bigger (or smaller) than words: in this view, *gave me* would be a stored item but *wove me* would have to be composed of

[1] Thanks to Ivan Tam for programming work on the written corpus, Nikki Foster for data entry on *d*-initial stems, Rosie Jones for sharing a seed corpus (Ghani & al. 2004), and Rowena Christina Guevara for sharing an audio corpus (Guevara & al. 2002). Thanks to Kevin Ryan for phonetic investigations and computing on the corpus. Thanks for discussion of various issues in this paper—without blame for its faults—to Anne Christophe, Bruce Hayes, Sharon Peperkamp, Janet Pierrehumbert, Kevin Ryan, Katrin Skoruppa, Donca Steriade, and Colin Wilson, and especially to the organizers and participants of CLS 43. This work was supported by grants from the UCLA Academic Senate’s Committee on Research.

[2] or the stem and suffix separately

the separate items *wove* and *me*; similarly, *parental* would be a stored item but *parenthood* would be composed of the separate lexical items *parent* and *hood*. Phonology like *v*-deletion and stress shifting would be more likely to apply within a stored lexical item. For other researchers, items of various sizes may be stored, but what matters is how an item is accessed—thus, *parenthood* might well be a stored item, and yet the word could still tend to be accessed decompositionally (e.g., Hay 2003). And for yet other researchers, stored items can contain morpheme boundaries (as in *parent+hood*) that influence their phonological behavior (e.g., Baroni 2001, where a word may have competing representations with and without morpheme boundaries: *parenthood* vs. *parent+hood*). The differences between units and non-units can result from frequency (Bybee proposes that frequent sequences become stored items), from relative frequency (Hay proposes that access route is influenced by which is more frequent, the whole sequence or its subparts), from distribution cues that learners use to guess boundary locations (Baroni 2000), or from phonological cues to boundary locations (Trubetzkoy 1939, Raffelsiefen 1999, Hay 2003).

Explanations for frequency and boundary-signal effects on the phonological behavior of morphologically complex sequences all depend on models—explicit to varying degrees—of lexical learning, storage, and access. The model of word recognition proposed by Hay (2003), for example, starts with a fast phonological preprocessor (Pierrehumbert 2002) that chunks the speech stream using purely phonological cues, such as phoneme sequences that would be unusual within a monomorphemic word—this accounts for the effect of boundary signals. Then, lexical items of various sizes—*happy, fortunate, unhappy, unfortunate, un*—become activated according to their similarity to the target and resting activation, which is a function of frequency. A word that is more frequent than its stem (e.g., *unfortunate*) will tend to be accessed directly, whereas a word that is less frequent than its stem (*unhappy*) will tend to be accessed via its component parts (*un*, *happy*): the components become activated and spread activation to the whole word.

This paper presents three optional phonological processes in Tagalog, investigating their application quantitatively. For the first and most extensively discussed rule, tapping (section 2), it will be shown that the rule applies variably and is subject to frequency influence in prefixed words and word+clitic combinations, but applies—or is blocked—more uniformly in suffixed words and two-syllable (σσ-) reduplication. A grammar is proposed that restricts where frequency can have an effect. The second rule, vowel raising (section 3), is also subject to frequency influences and a phonological influence in one environment (σσ-reduplication), but is subject only to phonological influences in another (suffixed words); this lack of frequency influence in suffixed words is predicted by the grammar proposed for tapping. The third rule, nasal substitution (section 4), is of interest because it is applicable only at the prefix-stem boundary: although it applies most often in higher-frequency words—subject to phonological restrictions—the morpheme boundary in these words must not be entirely erased.

In order to achieve a sufficiently large data set, this paper relies on written data, from a corpus composed of web pages (targeted to be 'probably Tagalog') consisting of about 20 million Tagalog words. The corpus is described in Zuraw 2006, although the version used here has additionally been subjected to boilerplate stripping (using Baroni 2005). Although phonetic data is not discussed here, corroboration for the written data on tapping comes from analysis by Kevin Ryan of an audio corpus of spoken Filipino created by Guevara et al. (2002). Glosses are drawn from English 1986.

2 Case study I: tapping

2.1 Overview

The Tagalog phenomenon to be discussed in the most detail here, tapping, concerns the distribution of [d] (spelled *d*) and [ɾ], spelled *r*. The examples below are given in normal spelling (often with additional hyphens to mark morpheme boundaries), except for some phonetic transcriptions, given in square brackets.

Among native monomorphemic words, the two sounds are, with few exceptions, in complementary distribution: *r* occurs intervocalically (*araw* 'sun, day') and *d* elsewhere (*dapat* 'should', *likod* 'back', *ganda* 'beauty', *idlip* 'nap'). Loans have introduced exceptions in both directions, with intervocalic *d* (*barkada* 'group'), and non-intervocalic *r* (*radyo* 'radio', *ambasador* 'ambassador', *sobra* 'too much', *barkada* 'group'). Among the native words, there is an alternation parallel to the complementary distribution, whereby a *d* that becomes intervocalic becomes *r*. For example, the stem *lakad* 'walk' has the suffixed form *lakar-an* 'to be walked on'. Schachter and Otanes (1973) describe some of the environments for tapping. When a *d*-final stem takes a vowel-initial suffix, tapping is described as obligatory, as *lakad*, *lakar-an*. When a *d*-initial stem takes a vowel-final prefix, Schachter and Otanes describe variation, implied by the presentation to be across words rather than within words: *dumi* 'dirt' *ma-rumi* 'dirty' but *dahon* 'leaf' *ma-dahon* 'leafy'. This includes the CV- or CVː- reduplicant: *dambong* 'plunder' *man-da-rambong* 'plunderer' vs. *dula* 'drama' vs. *man-du-dula* 'dramatist. (Unfortunately, there are no *d*-initial suffixes or *d*-final prefixes.) Schachter and Otanes don't discuss the application of tapping in two-syllable (σσ-) reduplication,[3] but all their examples have no reduplication: *dala-dala* 'load carried', *agad-agad* 'at once' (in the *agad-agad* cases, we might not expect tapping anyway, since, at least in careful speech, the stems begin with a glottal stop: [ʔagadʔagad]). A final environment for tapping is clitic-initially: the clitics *daw* 'reportedly' and *din* 'also' have allomorphs *raw* and *rin* that can occur after vowel-final words or clitics (and, less frequently, after consonant-final words): *ako raw* ~ *ako daw* 'me, reportedly' and *ako din* ~ *ako rin* 'me too'.

[3] Most Tagalog roots are two syllables, so this reduplication pattern will often appear to copy the whole root. But, for longer roots, only the first two syllables are copied: *pare-pareho* 'very similar' from *pareho* 'similar'.

2.2 Prefixed words

The written corpus data agree with Schachter and Otanes's description of prefixed words. All word types (unlemmatized) in the corpus were run through a morphological segmenter that undoes nonconcatenative morphology/phonology, and words that could be prefixed forms of known, native roots—with the prefix being vowel-final and the stem being *d*-initial, and additional affixes preceding or following the prefix+root sequence allowed—were identified.

Figure 1 is a histogram showing results for words with a corpus frequency of at least 10 (lower-frequency words are shown below). The words are divided, along the horizontal axis, into bins indicating the rate at which the word is spelled with *r*. Words that are spelled with *r* only 0-5% of the time (i.e., are spelled with *d* 95-100% of the time), such as *ma-dahon*, are counted in the leftmost bin, and words that are spelled with *r* 95-100% of the time, such as *ma-rumi*, are counted in the rightmost bin. There are relatively few words with intermediate rates of *r*.

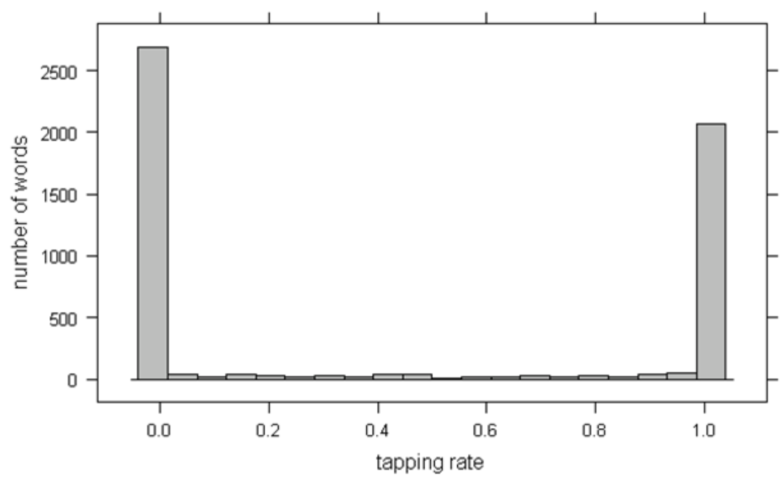


Figure 1: Written-corpus data for prefixed words with frequency ≥ 10.

What determines whether a prefixed word undergoes tapping? As a first approximation to the analysis to be proposed in section 2.5, suppose that tapping is blocked just in case the /VdV/ sequence is split over two separately accessed lexical units. I will assume (following, e.g., Baayen & Schreuder 1999) a dual-route view of lexical access: a prefixed word can be accessed either directly or via its components, and which route is used depends in part of the relative frequencies of the word and its subparts (Hay 2003). A word like *ma-dahon*, which is less frequent in the corpus than its stem *dahon* (9 vs. 1,947 occurrences in the corpus), would probably be accessed via its components *ma* and *dahon*—thus, the /ada/ sequence is split over two accessed units (though it also occurs within the whole word, which is presumably accessed too). By contrast, a word like *ma-runong* 'intelligent', which is more frequent than its stem *dunong* 'knowledge' (9,164 vs. 902) is expected to be accessed directly, and thus no accessed-unit boundary interrupts the /adu/ sequence, and tapping is free to apply. The predictions of such a model for an individual word depend on just how the model is implemented, and this paper will not explore any such predictions in detail, but instead will simply use word frequency as a proxy for lexical-access effects. In general, it will be expected that higher-frequency words should tend to be accessed through the whole-word route, and lower-frequency words should tend to be accessed through the decomposed route.

The effect of frequency in the prefixed words can be seen in Figure 2, which shows a histogram like Figure 1 for five subsets of the prefixed words. The cell labeled "A:2 to 4" shows how many words with frequency from 2 to 4 have each rate of tapping, the cell labeled "B:5 to 9" shows the data for words with frequency 5 to 9, etc. We can see that among the lowest-frequency words (A, B, and C), there are slightly more non-tapped words than tapped. Among the words with intermediate frequency (D) the split is about even, and among the highest-frequency words (E) tapping predominates. Thus, the effect of word frequency is as expected, suggesting a role for lexical access in determining whether a prefixed word undergoes tapping or not.

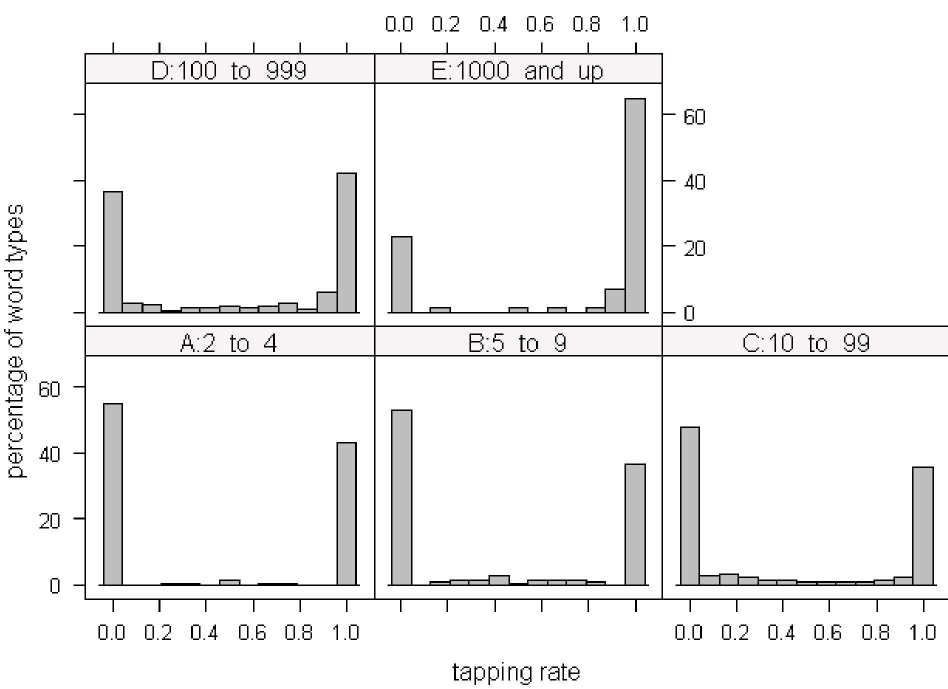


Figure 2: Prefixed words, grouped by frequency.

2.3 Suffixed words

Turning now to the suffixed words (*lakad+an*), recall that tapping is claimed to be obligatory there. This is supported by the corpus results—nearly all the suffixed words are consistently spelled with *r*:

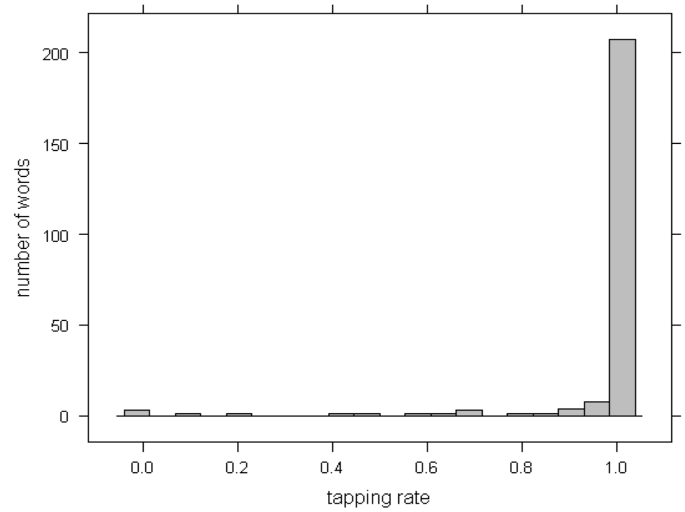


Figure 3: Suffixed words, no frequency cutoff.

2.4 Two-syllable reduplication

Schachter and Otanes don't list σσ-reduplication (*dala-dala*, *agad-agad*) as an environment where tapping can occur. This reduplication is of particular interest because σσ-reduplicated words, impressionistically, appear to have the same prosody as compounds—each copy seems to bear a primary stress—but they are much easier to identify automatically than are compounds.

The spellings in the written corpus are mostly with *d*, as shown in Figure 4, which excludes words of the type *agad-agad* (there are not many), where tapping would not be expected anyway. Furthermore, non-tapping predominates in all the frequency categories; it is not concentrated in lower-frequency words as would be expected if there were a frequency effect. I interpret these results to mean that tapping is essentially not allowed in this environment, though there are numerous exceptions. (There are 11 words in Figure 4 with ≥ 50% tapping; of those, 5 are from the root *dami* 'amount', e.g. *marami-rami* 'very many'. The root *dami* occurs mainly in the extremely high-frequency word *marami* 'many'.)

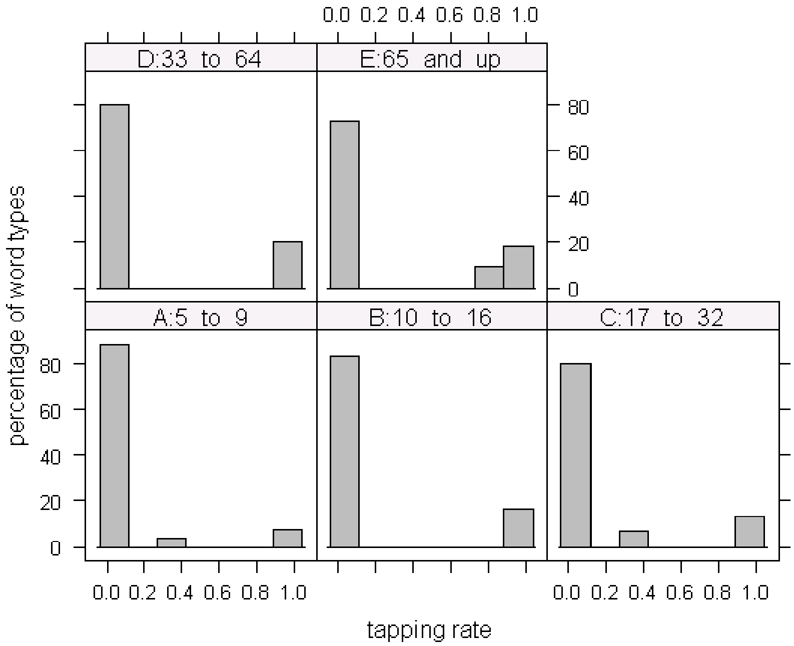


Figure 4: σσ-reduplicated words (84 word types total).

2.5 Grammar

We have seen that in prefixed words, frequency correlates with tapping behavior. In suffixed words, tapping occurs regardless of frequency, and in σσ-reduplicated words, tapping rarely occurs, regardless of frequency. It seems, then, that even if lexical access route can explain some of tapping's distributions, we still need a grammar in the traditional sense. The grammar will be implemented here, for concreteness, using Optimality Theory (Prince & Smolensky 1993/2004), but other approaches might be possible.[4] The essential properties of the grammar proposed here are that it requires the stem+suffix to be in the same domain for tapping, it requires stem+stem (the structure assumed for σσ-reduplication) to form two separate domains, and it allows two options for prefixation.

The proposed analysis will rely on prosodic domains (e.g., Selkirk 1980), following analyses of Northern Italian *s*-voicing by Nespor & Vogel 1986 and Peperkamp 1997. The domain of tapping will be called the prosodic word (or p-word, symbolized ω), since it is roughly the size of a morphological word and primary stress seems to have a similar domain[5] (the prosodic word being the label given to the domain for primary stress in many languages). A key assumption, following these earlier authors, is that domains posited to account for overt prosodic properties such as primary stress, pitch-accent placement, boundary-tone

---

[4] Hayes et al. 2003 was used to check rankings.

[5] The stem+suffix clearly forms a single unit for stress assignment, with stress shifts taking place (e.g., *lákad*, *lakár-an*), suggesting a single p-word. In σσ-reduplicated words, treated here as containing two p-words, each copy, impressionistically, bears a main stress. The stress relationship of prefixes and stems is less clear.

placement, and final lengthening also act as domains for the application of segmental rules, such as tapping.

The constraint against tapping, given in (1), applies only within a prosodic word.[6] The structures that the grammar will derive are shown in (2), in two formats.

(1) *(...VdV...)$_\omega$: the sequence vowel-[d]-vowel, uninterrupted by a p-word boundary, is forbidden.

(2) a. (prefix+stem)$_\omega$

p-word
prefix stem *tapping applies*

b. (prefix+(stem)$_\omega$)$_\omega$

p-word
p-word
prefix stem *no tapping*

c. (stem+suffix)$_\omega$

p-word
stem suffix *tapping applies*

d. (redup$_{\sigma\sigma}$)$_\omega$+(stem)$_\omega$

p-word p-word
redup$_{\sigma\sigma}$ stem *no tapping*

For prefixed words, the choice between the two structures is made by an Alignment constraint (McCarthy & Prince 1993) much like Prince and Smolensky's (1993/2004) LX≈PR, one of whose versions (p. 114) could be rephrased as ALIGN(LexWd,L,PWd,L), requiring the left edge of any lexical word (here, tentatively, noun, verb, adjective) to coincide with the left edge of some prosodic word. The difference here is that rather than depending on the syntactically-defined "LexWd", the constraint refers to accessed lexical units:

(3) ALIGN(AccU,L,PWd,L): the left edge of any accessed lexical unit must coincide with the left edge of some prosodic word.

If the stem has been accessed separately, as in the input to (4), the alignment constraint rules out a simple structure (candidates *c*, *d*). A recursive structure with the prefix doesn't satisfy the alignment constraint either (*f*), and we are left with a recursive structure (*a*, *b*) or separate p-words for the prefix and stem (*e*). Because *e* and the winning *a* are homophonous, this example does not argue for the grammar to choose one over the other. As will be seen later in the discussion of

6 It is crucial that the constraint is violated when the *VdV* sequence occurs uninterrupted by a p-word boundary, as opposed to when the whole *VdV* sequence is dominated by some p-word node, because in recursive structures such as will be proposed below, a sequence can be dominated by the same p-word node and yet interrupted by the boundary of a lower p-word.

suffixing, however, the grammar adopted here rules out monosyllabic p-words through undominated MINIMALITY (7). We are thus down to *a* vs. *b*; because there is no violation in either candidate of the constraint driving tapping, *(...VdV...)$_{\omega}$, candidate *a*, without a tap, is preferred. (The low-ranked constraint *r, which penalized all taps, makes the decision. It would also be possible to have a set of more specific constraints including *($_{\omega}$r, forbidding prosodic-word-initial tap—see Peperkamp on Northern Italian *s*-voicing.)

(4)

accessed: ma, dahon, madahon	*(...VdV...)	MINIMALITY	STEMISHEAD	ALIGN (AccU,L,PWd,L)	NORECURSION	*r
☞ *a* (ma(dahon)$_{\omega}$)$_{\omega}$					*	
b (ma(rahon)$_{\omega}$)$_{\omega}$					*	*!
c (madahon)$_{\omega}$	*!			*		*
d (marahon)$_{\omega}$				*!		
e (ma)$_{\omega}$(dahon)$_{\omega}$		*!				
f ((ma)$_{\omega}$ dahon)$_{\omega}$		*(!)	*(!)	*	*	

When only the whole word has been accessed, the anti-recursion constraint NORECURSION rules out the recursive candidates (*g*, *h*) in favor of the simple candidates (*i*, *j*), and the tapping constraint prefers (*j*).

(5) NORECURSION: a prosodic node of category *n* must not dominate another node of category *n*.

(6)

accessed: madami	*(...VdV...)	MINIMALITY	STEMISHEAD	ALIGN (AccU,L,PWd,L)	NORECURSION	*r
g (ma(dami)$_{\omega}$)$_{\omega}$					*!	
h (ma(rami)$_{\omega}$)$_{\omega}$					*!	*
i (madami)$_{\omega}$	*!					
☞ *j* (marami)$_{\omega}$						*
k (ma)$_{\omega}$(dami)$_{\omega}$		*!				
l ((ma)$_{\omega}$ dami)$_{\omega}$		*(!)	*(!)		*	

In suffixed words, on the other hand, the ALIGN(AccU,L,PWd,L) constraint can't be satisfied. The constraint MINIMALITY is shorthand for two cross-linguistically common requirements: a prosodic word must contain at least one foot, and a foot at least two syllables:

(7) MINIMALITY: a p-word must contain at least two syllables

Support for ranking this constraint high in Tagalog comes from the fact that non-loan content words are all at least disyllabic. Moreover, monosyllabic and disyllabic pronouns behave differently, with monosyllables cliticizing "more closely" than disyllables (Schachter & Otanes 1972, Anderson to appear). Because Tagalog has just two (productive) suffixes, *-in* and *-an*, both monosyllabic, neither can stand on its own as a p-word.

As shown in (8), even if the stem and suffix are both accessed, tapping must occur. MINIMALITY rules out the two candidates that satisfy ALIGN(AccU,L,PWd,L): the suffix-headed recursive candidate (*a*) and the candidate where stem and suffix form two separate p-words (*d*). Note that the ALIGN constraint is not satisfied by candidate (*e*), with a stem-headed recursive structure, because the left edge of the suffix (an accessed unit) does not coincide with the *left* edge of any p-word. A tableau for a suffixed word under whole-word access is also given for comparison (9); the outcome is the same.

(8)

	accessed: lakad, an, lakaran	*(...VdV...)	MINIMALITY	STEMISHEAD	ALIGN (AccU,L,PWd,L)	NORECURSION	*r
	a $(\text{lakad}(\text{an})_\omega)_\omega$		*(!)	*(!)		*	
	b $(\text{lakadan})_\omega$	*!			*		
☞	*c* $(\text{lakaran})_\omega$				*		*
	d $(\text{lakad})_\omega(\text{an})_\omega$		*!				
	e $((\text{lakad})_\omega\ \text{an})_\omega$				*	*!	

(9)

accessed: lakadan	*(...VdV...)	MINIMALITY	STEMISHEAD	ALIGN (AccU,L,PWd,L)	NORECURSION	*r
f (lakad(an)$_\omega$)$_\omega$		*(!)	*(!)		*	
g (lakadan)$_\omega$	*!					
☞ *h* (lakaran)$_\omega$						*
i (lakad)$_\omega$(an)$_\omega$		*!				
j ((lakad)$_\omega$ an)$_\omega$					*!	

Turning to σσ-reduplicated words, here again the outcomes are homophonous under either access route (ignoring the exceptional words that do tap). A constraint STEMISHEAD (10) requires every stem to head its own p-word (other non-stem material may be included in the p-word that the stem head). Assuming that the two-syllable reduplicant has the status of a stem (following Urbanczyk's 2001 treatment of languages with short and long reduplicants; the σσ-reduplicant would be a root in her terms), STEMISHEAD requires the two copies to form separate p-words.[7]

(10) STEMISHEAD: every stem must head its own p-word

When the stem is accessed independently, of course tapping does not apply (11). But even if only the whole word is accessed, as long as its syntactic structure has not become synchronically lost, the STEMISHEAD constraint requires two separate p-words, and thus blocks tapping (12). (There is a difference between the two outcomes in minor-phrase structure, which is irrelevant to tapping but will be discussed below in section 3.)

[7] Base-reduplicant identity (McCarthy & Prince 1995) is another possibility (although it would have to be stronger for σσ-reduplication than for CV-reduplication): perhaps the second copy resists tapping because of pressure to be identical to the first copy, which itself must resist tapping because of the prohibition on (in native words) word-initial tap. This theory predicts that in prefixed compound-reduplicated words—like *kadaki-dakila* from *dakila* 'eminent'—the two copies should still tend to behave alike, but should both show tapping a fair amount of the time. Instead, these words show very little tapping at all (except derivatives of *dami*). Perhaps base-reduplicant identity preservation is at work, but in the opposite direction: the impossibility of tapping in the second copy prevents the first copy from tapping.

(11)

	accessed: dala, daladala	*(...VdV...)	MINIMALITY	STEMISHEAD	ALIGN (AccU,L,PWd,L)	NORECURSION	*r
	a [(dala(dala)$_\omega$)$_\omega$]$_\varphi$			*!		*	
	b [(dalarala)$_\omega$]$_\varphi$			*!	*		*
☞	*c* [(dala)$_\omega$(dala)$_\omega$]$_\varphi$						
	d [(dala)$_\omega$(rala)$_\omega$]$_\varphi$						*!
	e [(dala)$_\omega$]$_\varphi$[(dala)$_\omega$]$_\varphi$						
	f [(dala)$_\omega$]$_\varphi$[(rala)$_\omega$]$_\varphi$						*
	g [((dala)$_\omega$ dala)$_\omega$]$_\varphi$			*!	*	*	

(12)

	accessed: daladala	*(...VdV...)	MINIMALITY	STEMISHEAD	ALIGN (AccU,L,PWd,L)	NORECURSION	*r
	h [(dala(dala)$_\omega$)$_\omega$]$_\varphi$			*!		*	
	l [(dalarala)$_\omega$]$_\varphi$			*!			*
☞	*j* [(dala)$_\omega$(dala)$_\omega$]$_\varphi$						
	k [(dala)$_\omega$(rala)$_\omega$]$_\varphi$						*!
	l [(dala)$_\omega$]$_\varphi$[(dala)$_\omega$]$_\varphi$						
	m [(dala)$_\omega$]$_\varphi$[(rala)$_\omega$]$_\varphi$						*
	n [((dala)$_\omega$ dala)$_\omega$]$_\varphi$			*!		*	

This grammar involves a change from the standard view of how the lexicon feeds into the grammar. Typically, in the phonological literature, it is assumed that lexical access delivers to the grammar a set of morphemes, with no information as to how they were accessed (although there may be information about morpho-syntactic structure): /morpheme+morpheme/. In order for constraint such as ALIGN(AccU,L,PWd,L) to be evaluable, lexical access must instead deliver a set of accessed units, or, equivalently for our purposes, a set of morphemes, annotated for how they were accessed: / morpheme + morpheme / (whole-word) or / morpheme + morpheme / (decomposed), for instance, with ellipses surrounding accessed units.

2.6 Clitics

There is one more environment for tapping to consider. Tagalog has two enclitics that can undergo tapping: *daw*, which marks reported statements, and *din* 'also'. These enclitics have the allomorphs *raw* and *rin*, which can occur after vowel-final words (and, less frequently, after consonant-final words).

(13) ako rin ~ ako din 'me too'
ako raw ~ ako daw 'me, reportedly'

In the written corpus, we can examine all word+clitic combinations to see how often each is spelling with *daw/din* and how often with *raw/rin*. Unlike in the case of prefixed words, here we see the full range of behaviors, with many word+clitic combinations showing intermediate rates of tapping (Figure 5). This suggests that the choice of allomorphs is made online, at least in a substantial fraction of the cases (unlike in the prefixed words, whose behavior may be lexicalized).

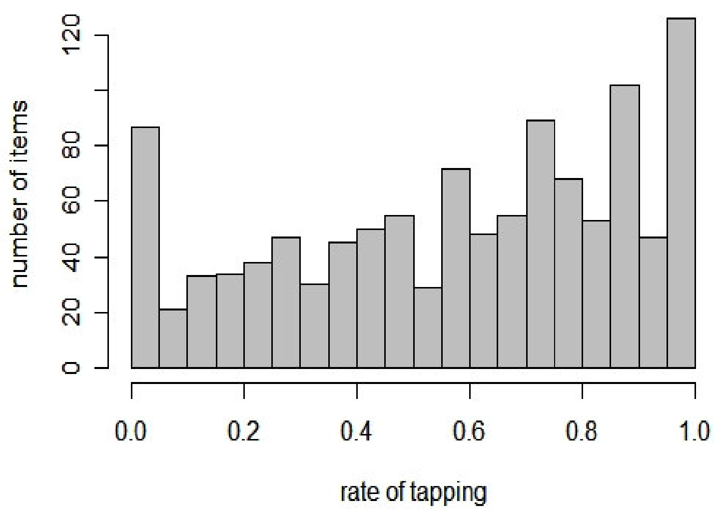


Figure 5: Word+clitic combinations.[8]

Here too, distributional effects apply. Higher frequency of the word+clitic combination is associated with higher rates of tapping, as illustrated in Figure 6, where each point represents one word+clitic combination.

[8] All the clitic data are from a version of the corpus that has not undergone boilerplate stripping. It should make little difference, however, as nearly all boilerplate in the corpus is in English.

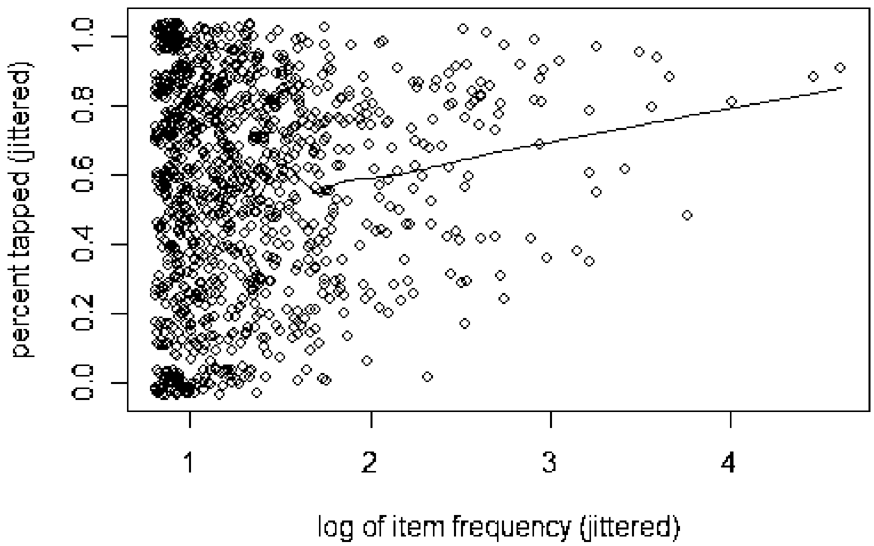


Figure 6: Word+clitic combinations: frequency vs. percent tapped; line: supersmoother.

The frequency effect can be seen more clearly for clitic+*din*/*daw* combinations (when the first clitic is vowel-final), as found in *ako* ***pa rin*** 'still me also', shown in Figure 7.

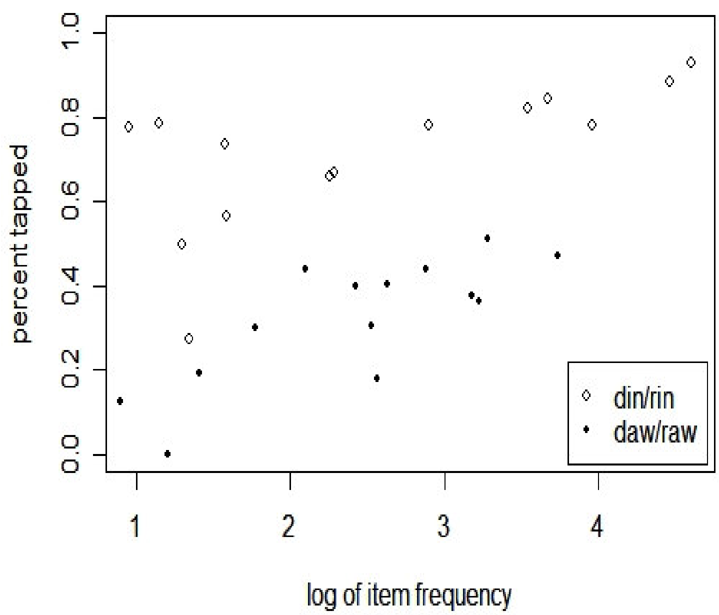


Figure 7: Clitic+clitic combinations: frequency vs. percent tapped.

It remains to be investigated whether there is a detectable prosodic difference between *X+daw/din* and *X+raw/rin* sequences.

3 Case study II: o/u alternations

In native Tagalog words, *o* and *u* are in roughly complementary distribution, with *o* in final syllables and *u* in nonfinal syllables (certain complications and exceptions are discussed below; see Zuraw 2003 for further details). An alternation accompanies this difference in distribution: when a suffix is attached to a stem with *o* in its final syllable, the *o* becomes *u*, because it is no longer in the final syllable, for example *tapos* ‘ending’, *tapus-in* ‘to be finished’.

With σσ-reduplication, we can also see an alternation, as in (14), although there is a great deal of variation, and the variants with *o* seem to be more frequent. It is likely that, unlike tapping, this alternation is phonetically gradient, with many tokens that fall somewhere between *o* and *u*.[9]

(14) halo ‘mixture’ halo-halo ~ halu-halo ‘(a dessert)’

Spelling here is probably not as reliable as it is for tapping, because the alternation gives the impression of being phonetically gradient, with many tokens that fall somewhere between *o* and *u*. The written data should then be interpreted with caution, but have been investigated nonetheless because they at least provide some hypotheses for testing in future research.

Looking first at compounding reduplication, we see in Figure 8 that there is much more within-item variation, compared to tapping in prefixed words (Figure 1). The largest group of words show *o* nearly all the time (the leftmost bin, 0-5% *u* use), and the second-largest group show *u* nearly all the time (the rightmost bin), but a fair number are somewhere in between.

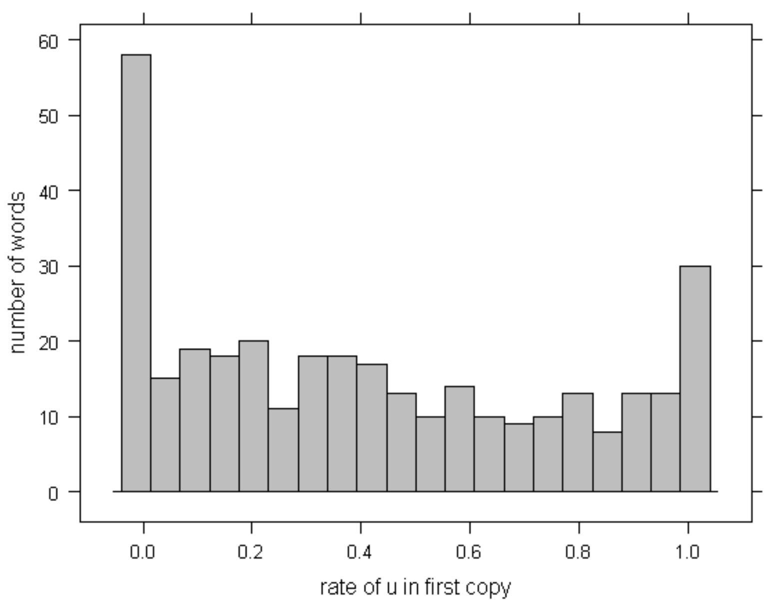


Figure 8: σσ-reduplication—words with frequency ≥ 10 only.

[9] Thanks to Janet Pierrehumbert for discussion of this point, which remains to be researched.

If there were to be a frequency effect, what would we expect it to be? In monomorphemic words, *o* occurs in final syllables; thus, spelling the first copy of a compound-reduplicated words with *o* (*hal**o**-halo* instead of *hal**u**-halo*) should indicate a strong boundary between the two copies, so that both *lo* syllables are treated as word-final. This should correlate with decomposed access, which should be found primarily in low-frequency words. In sum, we expect lower-frequency words to show less *u* and more *o*.

The expected frequency effect is found in the written corpus, with *o* being concentrated in lower-frequency words. Figure 9 shows that the most *o* use in the first copy is concentrated in the words with frequency up to 15 (in group A, 57% of words use *o* more 80% of the time; in group B, 50%; C 41%; D 31%; E 23%, F 25%); words with higher frequencies have more evenly distributed rates of *u* use.

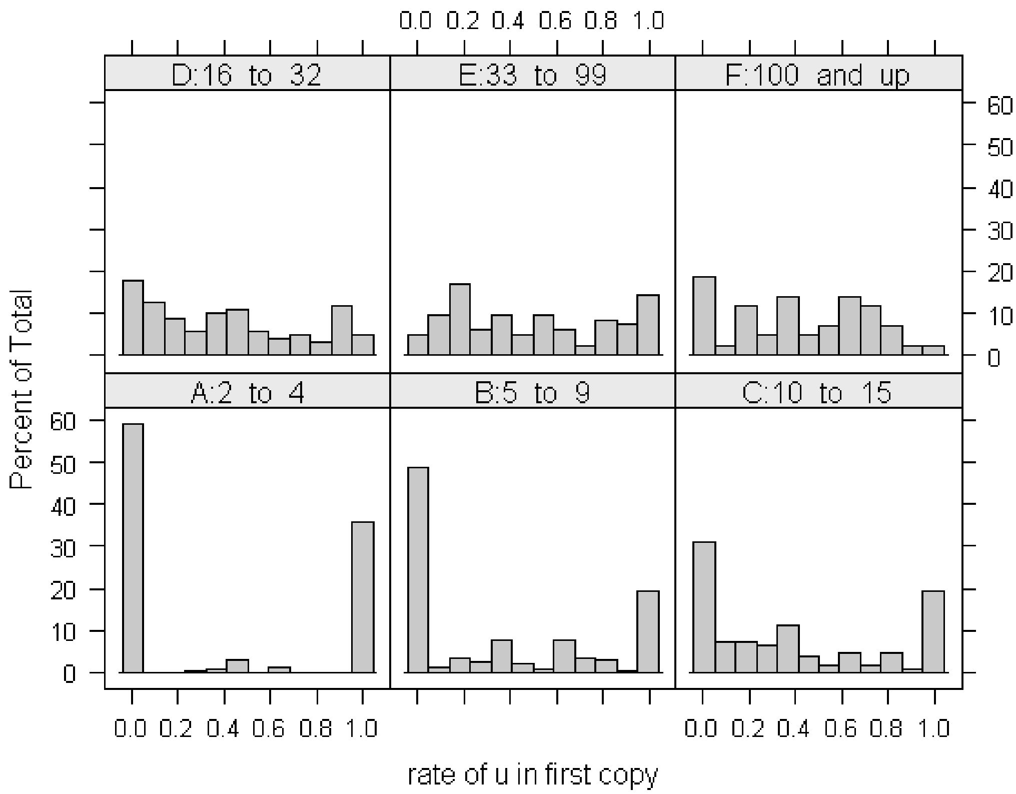


Figure 9: σσ-reduplication: rate of *u* use in first copy, grouped by frequency.

There is also a reduplicative identity effect. Two-syllable-reduplicated words can take suffixes, as in *ka-tapus-tapus-an* 'very last'. In these words, the second copy's vowel is nearly always *u*—this is to be expected in the grammar introduced above, where stem and suffix obligatorily belong to the same p-word, so that the last vowel of the second copy is not word-final. In these words, the first copy's vowel strongly tends to be *u* also, as shown in the histogram on the left side of Figure 10. We also see now that among unsuffixed words (on the right) there is a stronger tendency towards *o* than was apparent in Figure 8.

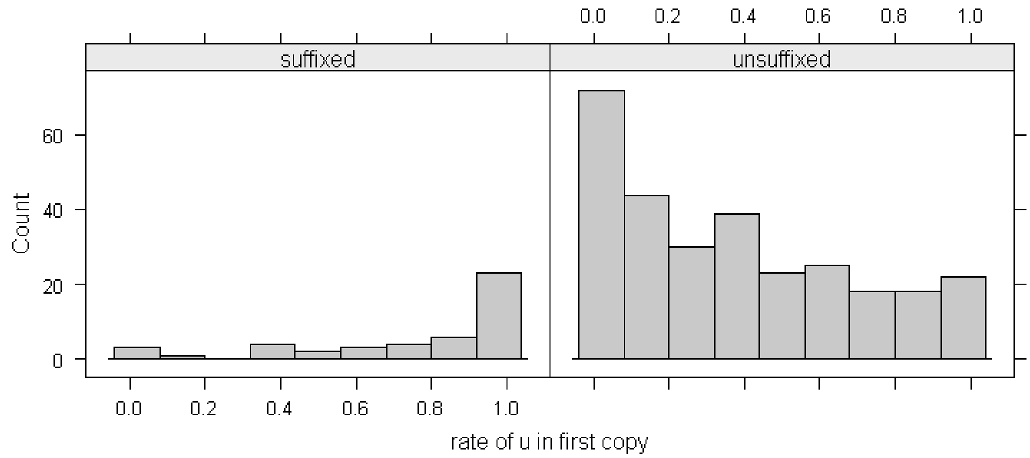


Figure 10: σσ-reduplication: rate of *u* in first copy, suffixed (left) vs. unsuffixed (right).

This reduplicative identity effect seems to be basically obligatory, wiping out any possible frequency effect. In Figure 11 we see that *u* predominates just as strongly for the low-frequency words as for the high-frequency words.

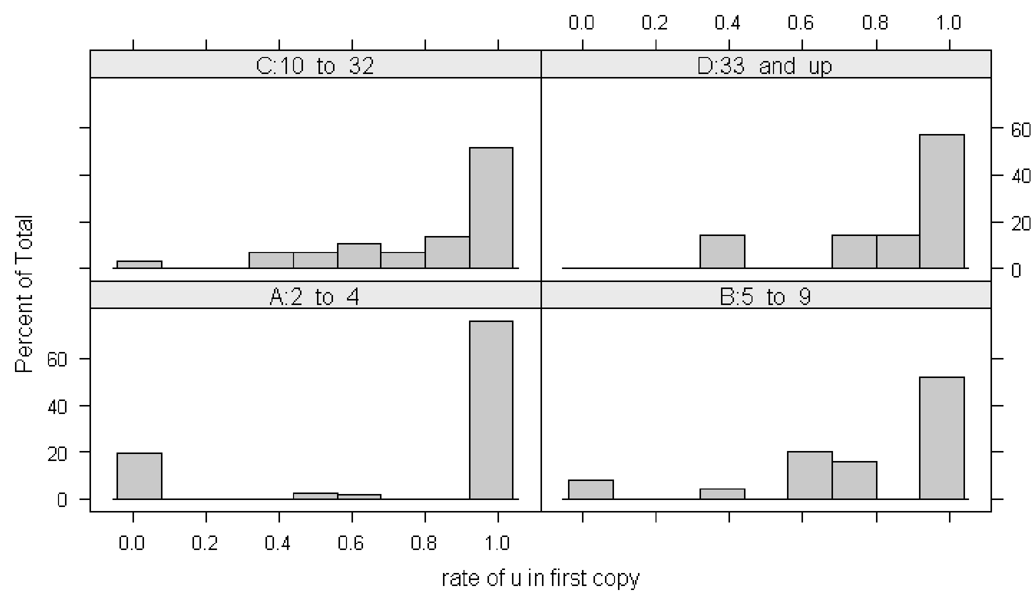
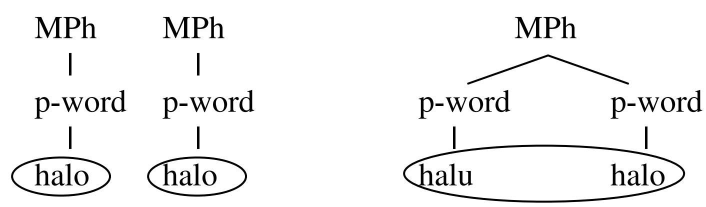


Figure 11: σσ-reduplication, suffixed words only: percent *u* in first copy, split by frequency.

Accounting for the vowel-raising facts requires a few additions to the grammar proposed above. First, we must deal with the variation in vowel height for compound-reduplicated words: the second syllable of the first copy behaves sometimes as domain-final—*halo-halo*—and sometimes as domain-medial—*halu-halo*—whereas with tapping, we saw that the beginning of the second copy always behaves as domain-initial. These facts on their own could be accounted for by letting the first copy either be adjoined to the second copy's p-word—*(halu(halo)$_{\omega}$)$_{\omega}$*—or head its own p-word—*(halo)$_{\omega}$(halo)$_{\omega}$*. The problem with that analysis is that it fails to distinguish the tapping behavior of prefixed words (which vary) and compound-reduplicated words (which resist tapping). In the grammar above, this was dealt with through a high-ranked constraint STEMISHEAD which requires each copy in compounding reduplication to head its own p-word. This constraint rules out *(halu(halo)$_{\omega}$)$_{\omega}$*. Instead, I propose that the domain of raising/lowering is bigger than that of tapping. I will refer to this domain as the minor phrase (abbreviated MPh), though this term may turn out not to be compatible with a fuller analysis of Tagalog intonation (see, e.g., Richards 2006). A compound-reduplicated word can have two different structures, depending on access mode (15): two separate minor phrases if the two copies are accessed separately or a single minor phrase if the word is accessed whole. In both structures, each copy forms its own p-word (preventing tapping), but they differ in whether the second syllable of the first copy is minor-phrase final ([o]) or minor-phrase-medial ([u]). It remains to be seen whether compound-reduplicated words with and without vowel raising have different intonational or durational properties to provide external support for the proposed prosodifications.

(15)



The new constraints needed are in (16).[10]

(16) a. *o/NONMPHFINAL: The vowel *o* is forbidden in syllables that are not final in some minor phrase.
b. *u]: The vowel *u* is forbidden in the last syllable of a minor phrase.
c. ALIGN(AccU,L,MPh,L): The left edge of any accessed unit must coincide with the left edge of some minor phrase.
d. ALIGN(MPh,L,AccU,L): The left edge of any minor phrase must coincide with the left edge of some accessed unit.

When decomposed access occurs, the constraint ALIGN(AccU,L,MPh,L) requires that each copy initiate a minor phrase:

10 IDENT-IO(hi) is not shown, because it is ranked lower than the other active constraints shown, and its violations will vary depending on what values for [high] are chosen for the input—in accordance with the principle of the rich base (Prince & Smolensky 1993/2004), the input could also have been *haluhalo*, *haluhalu*, or *halohalu*.

(17)

accessed: halo, halohalo	ALIGN(MPh,L,AccU,L)	*o/NONMPHFINAL	MAX-BR(hi)	STEMISHEAD	ALIGN(AccU,L,PWd,L)	ALIGN(AccU,L,MPh,L)	NORECURSION	*u]
☞ *a* [(halo)$_\omega$]$_\varphi$[(halo)$_\omega$]$_\varphi$								
b [(halu)$_\omega$]$_\varphi$[(halo)$_\omega$]$_\varphi$								*!
c [(halu)$_\omega$]$_\varphi$[(halu)$_\omega$]$_\varphi$								*!*
d [(halo)$_\omega$(halo)$_\omega$]$_\varphi$		*!				*		
e [(halu)$_\omega$(halo)$_\omega$]$_\varphi$						*!		
f [(halu)$_\omega$(halu)$_\omega$]$_\varphi$						*!		*
g [(halu(halo)$_\omega$)$_\omega$]$_\varphi$				*!		*	*	
h [((halu)$_\omega$ halo)$_\omega$]$_\varphi$				*!	*	*	*	
i [(haluhalo)$_\omega$]$_\varphi$				*!	*	*		

But when whole-word access occurs, the extra minor phrase is rule out by ALIGN(MPh,L,AccU,L):

(18)

accessed: halohalo	ALIGN(MPh,L,AccU,L)	*o/NONMPHFINAL	MAX-BR(hi)	STEMISHEAD	ALIGN(AccU,L,PWd,L)	ALIGN(AccU,L,MPh,L)	NORECURSION	*u]
a [(halo)$_\omega$]$_\varphi$[(halo)$_\omega$]$_\varphi$	*!							
b [(halu)$_\omega$]$_\varphi$[(halo)$_\omega$]$_\varphi$	*!							*
c [(halu)$_\omega$]$_\varphi$[(halu)$_\omega$]$_\varphi$	*!							**
d [(halo)$_\omega$(halo)$_\omega$]$_\varphi$		*!						
☞ *e* [(halu)$_\omega$(halo)$_\omega$]$_\varphi$								
f [(halu)$_\omega$(halu)$_\omega$]$_\varphi$								*!
g [(halu(halo)$_\omega$)$_\omega$]$_\varphi$				*!			*	
h [((halu)$_\omega$ halo)$_\omega$]$_\varphi$				*!			*	
i [(haluhalo)$_\omega$]$_\varphi$				*!				

Turning to the suffixed forms such as *ka-tapus-tapus-an*, where both copies tend strongly to have *u*, we can invoke McCarthy and Prince's (1995) base-reduplicant correspondence. Instead of using a symmetrical IDENT-BR(hi), which is violated any time corresponding base and reduplicant vowels fail to match for the feature [high, we will need the following asymmetrical constraint:[11]

(19) MAX-BR(hi): If a vowel in the base (assumed here to be the second copy) is [+high], then there must be a corresponding [+high] specification in the reduplicant (assumed here to be the first copy).[12]

As shown in the two tableaux below, when the word is suffixed the first copy must have *u* regardless of access mode.

(20)

accessed: halo, an, haloan	ALIGN(MPh,L,AccU,L)	*o/NONMPHFINAL	MAX-BR(hi)	ALIGN(AccU,L,PWd,L)	ALIGN(AccU,L,MPh,L)	NORECURSION	*u]
a [(halo)$_\omega$]$_\varphi$[(haloan)$_\omega$]$_\varphi$		*!		*	*		
b [(halo)$_\omega$]$_\varphi$[(haluan)$_\omega$]$_\varphi$			*!	*	*		
☞*c* [(halu)$_\omega$]$_\varphi$[(haluan)$_\omega$]$_\varphi$				*	*		*
d [(halo)$_\omega$]$_\varphi$[[(halo)$_\omega$]$_\varphi$ an]$_\varphi$				*	*	*!	
e [(halo)$_\omega$(haloan)$_\omega$]$_\varphi$		*!*		*	**		
f [(halo)$_\omega$(haluan)$_\omega$]$_\varphi$		*(!)	*(!)	*	**!		
g [(halu)$_\omega$(haluan)$_\omega$]$_\varphi$				*	**!		

[11] The MAX-BR(hi) solution adopted here is not the only possible one. Perhaps cyclic derivation is involved, with the reduplicant copied from the already-suffixed base. Whether this is tenable given the morphology remains to be seen.

[12] If a [+hi] vowel in the base lacks a correspondent altogether, this constraint will be violated, as in hypothetical *sunda-sundalu-hin* 'toy soldier' where the second base *u* is not copied. This will play no role, however, as higher-ranked constraints (needed independently) enforce a two-syllable size for the reduplicant regardless of vowel height in the base, and other constraints are needed independently to prevent corresponding features from associating to non-corresponding segments.

(21)

accessed: halohaloan	ALIGN(MPh,L,AccU,L)	*o/NONMPHFINAL	MAX-BR(hi)	ALIGN(AccU,L,PWd,L)	ALIGN(AccU,L,MPh,L)	NORECURSION	*u]
a [(halo)$_\omega$]$_\varphi$[(haloan)$_\omega$]$_\varphi$	*(!)	*(!)					
b [(halo)$_\omega$]$_\varphi$[(haluan)$_\omega$]$_\varphi$	*(!)		*(!)				
c [(halu)$_\omega$]$_\varphi$[(haluan)$_\omega$]$_\varphi$	*(!)						*
d [(halo)$_\omega$]$_\varphi$[[(halo)$_\omega$]$_\varphi$ an]$_\varphi$	*(!)					*	
e [(halo)$_\omega$(haloan)$_\omega$]$_\varphi$		*!*					
f [(halo)$_\omega$(haluan)$_\omega$]$_\varphi$		*(!)	*(!)				
☞ *g* [(halu)$_\omega$(haluan)$_\omega$]$_\varphi$							

The reason for rejecting IDENT-BR(hi) is that it would, incorrectly, force identity in non-suffixed words under whole-word access: in (22), candidate *f* would win instead of the correct *e*.

(22)

accessed: halohalo	ALIGN(MPh,L,AccU,L)	*o/NONMPHFINAL	MAX-BR(hi)	STEMISHEAD	ALIGN(AccU,L,PWd,L)	ALIGN(AccU,L,MPh,L)	NORECURSION	*u]
a [(halo)$_\omega$]$_\varphi$[(halo)$_\omega$]$_\varphi$	*!							
b [(halu)$_\omega$]$_\varphi$[(halo)$_\omega$]$_\varphi$	*!							*
c [(halu)$_\omega$]$_\varphi$[(halu)$_\omega$]$_\varphi$	*!							**
d [(halo)$_\omega$(halo)$_\omega$]$_\varphi$		*!						
☹ *e* [(halu)$_\omega$(halo)$_\omega$]$_\varphi$			*!					
💣 *f* [(halu)$_\omega$(halu)$_\omega$]$_\varphi$								*
g [(halu(halo)$_\omega$)$_\omega$]$_\varphi$			*(!)	*(!)			*	
h [((halu)$_\omega$ halo)$_\omega$]$_\varphi$			*(!)	*(!)			*	
i [(haluhalo)$_\omega$]$_\varphi$			*(!)	*(!)				

The grammar give so far correctly predicts that raising should be obligatory under suffixation: because the suffix must be part of the stem's p-word, a fortiori it must belong to the stem's minor phrase, and thus the last vowel of the stem is minor-phrase-medial:

(23) MPh
|
p-word

stem suffix
*tap***u***s* *in*

In native words, raising under suffixation does indeed seem to be obligatory (with the exception of some pseudo-reduplicated words: see Zuraw 2002). This is supported by the corpus data—nearly all the relevant words take *u* 95 to 100% of the time:

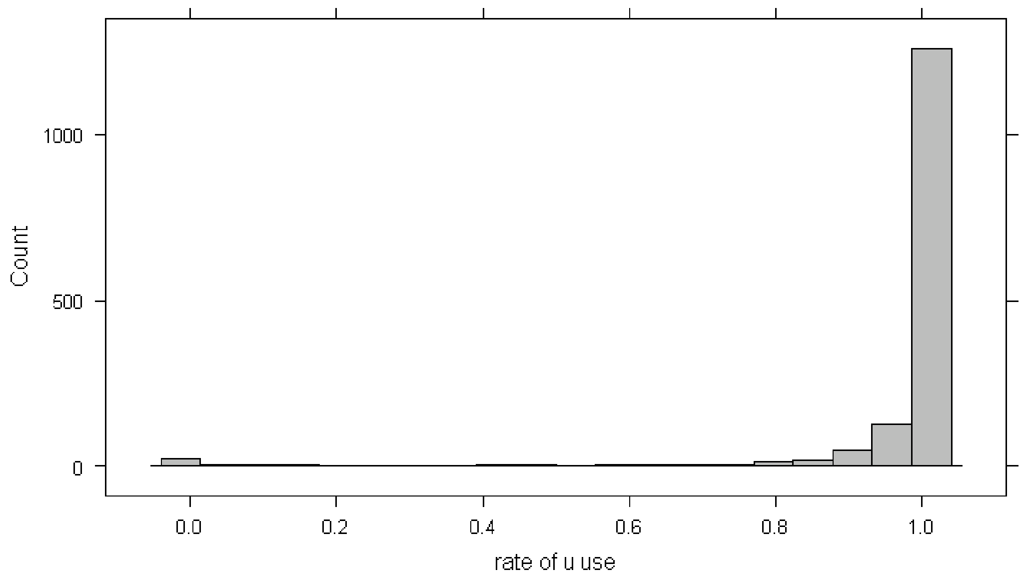


Figure 12: suffixation: rate of *u* in syllable preceding suffix.

Among loans, however, *o*-to-*u* raising under suffixation tends to be blocked when the preceding syllable has a mid vowel *e* or *o* (and more so to the extent that the two syllables are similar; this phenomenon is discussed in detail in Zuraw 2002). To take two extreme examples, *loko* 'crazy' (Spanish *loco*), whose penult and ultima both have *o*, tends not to undergo raising: *lok***o***-hin* 'be fooled'; by contrast, *prito* 'fry' (< Spanish *frito*), whose penult lacks a mid vowel, usually does raise: *prit***u***-hin* 'be fried'.

We might wonder if, in addition to this phonological effect, there is some frequency effect among the suffixed loans. The grammar so far predicts that there should not be: regardless of access mode, the stem and suffix belong to the same minor phrase. Whatever constraints are preventing raising in certain loans must be doing so despite the vowel in question's minor-phrase-medial position. This prediction is borne out by the corpus data. In Figure 13, we see that there is no consistent trend for words with lower frequency to have a greater proportion of *o*.

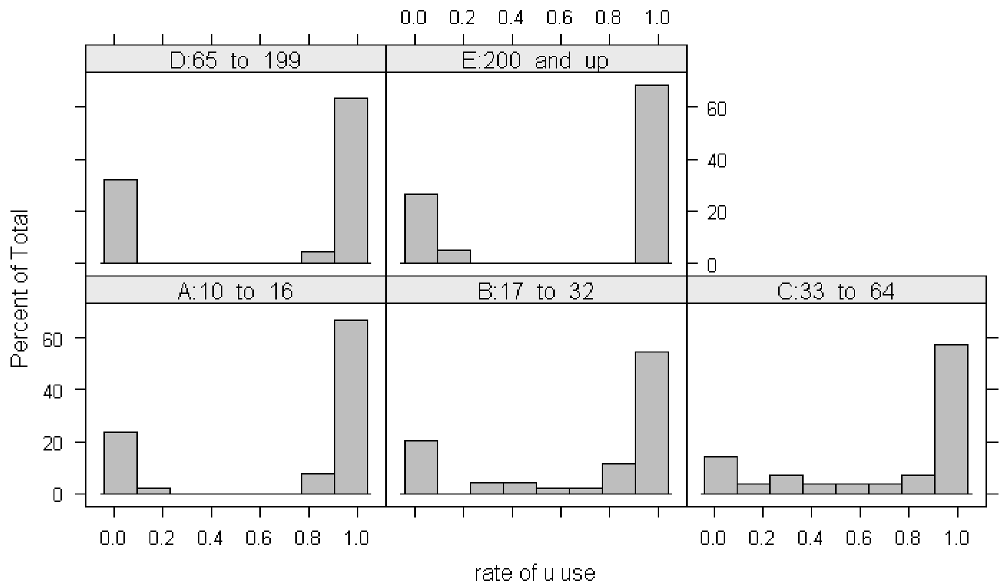


Figure 13: suffixation: rate of *u*, grouped by frequency.

4 Case study III: nasal substitution

Tagalog prefixes that end in a nasal (often taken to be underlyingly /ŋ/) can display three behaviors when they combine with an obstruent-initial stem (see Zuraw 2000 for much more detail on this phenomenon), as shown in (24). They can simply combine (a), or the nasal can assimilate in place to the following obstruent (b), or the nasal and obstruent can be replaced by a nasal with the place of articulation of the original obstruent (c). It is this third case that is known as nasal substitution; in /paŋ+pasko/, the nasal-substitution option is by far the most frequent, but for other words, other options may predominate.

(24)		/paŋ+pasko/	‘for Christmas’	
	a.	non-assimilation	paŋ-pasko	<pang-pasko>
	b.	assimilation	pam-pasko	<pam-pasko>
	c.	nasal substitution	pamasko	<pamasko>

As discussed in Zuraw 2000, whether a word characteristically undergoes nasal substitution is greatly influence by the stem-initial obstruent: voiceless *p*, *t*, *s*, *k* undergo nasal substitution at high rates, *b* at a lower rate, *d* at still a lower rate, and *g* almost never. The mosaic plot in Figure 14 shows data compiled from a dictionary, English 1986 (1,422 words). Each column shows, for words whose stems begin with a certain obstruent, the percentage of words that undergo nasal substitution (black), the percentage that vary (grey), and the percentage that do not undergo nasal substitution. The widths of the columns are scaled to reflect the number of items in the column—equal areas represent equal numbers of items; for example, the *d* and *g* columns are narrow because there are relatively few relevant words with *d*- or *g*-initial stems.

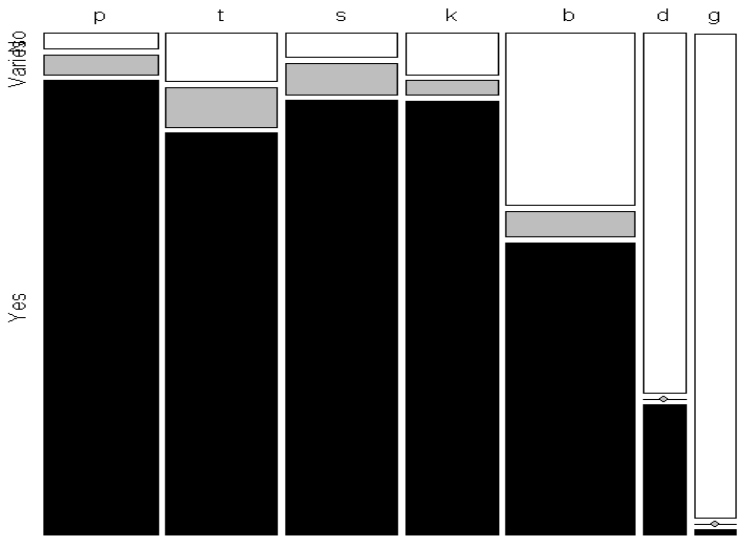


Figure 14: nasal substitution rates according to dictionary (English 1986)

In addition to the phonological effect—which wipes out most of the possibility for variation except in *b*- and *d*-initial stems—there seems to be a frequency effect. Figure 15 shows data for words with *b*-initial stems identified by the morphological segmenter. We can see that the highest-frequency words have the highest rate of nasal substitution (assimilation and non-assimilation are grouped together).

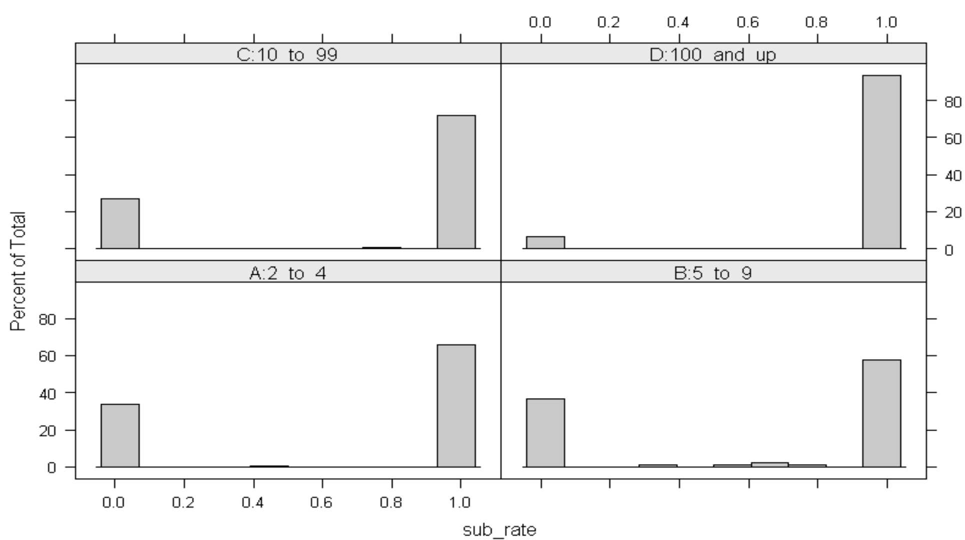


Figure 15: nasal substitution rates in corpus for words with *b*-initial stems

The frequency effect in nasal substitution is particularly interesting because it is not compatible with a model in which morphology-sensitive rules simply are blocked by morpheme boundaries, and applicable when boundaries have been erased because of, e.g., high frequency (perhaps diachronically). This is because nasal substitution is restricted to the prefix-stem boundary. There are plenty of nasal-obstruent sequences within roots, and they don't seem to show any tendency towards deletion of the obstruent. The rule or constraint driving nasal substitution must refer to the morpheme boundary, so the morpheme boundary must remain present, at least at the moment when a word acquires a nasal-substituted form. (Nasal substitution is still productive—Spanish and English loans can acquire nasal-substituted forms, as for example *pasko* in (24), from Spanish *pascua*.) I leave integration of the nasal-substitution case into the grammar set out above for future research.

5 Summary

The application of phonological processes has been argued here to be partly sensitive to distributional factors such as frequency, and partly governed by what we standardly think of as the grammar. This paper has explored three case studies of such dual conditioning in Tagalog. In the first case, tapping, we saw that morphology can override frequency: prefixed words show frequency-influenced variation, but suffixed words must undergo the rule, and compound-reduplicated words rarely do. In the second case, vowel height, we saw that a phonological effect, base-reduplicant identity, could override frequency: compound-reduplicated words show frequency-influenced variation, but (asymmetrical) base-reduplicant identity overrides this variation. No frequency effects were found for vowel height in suffixed words; there is variation among loans, but it seems to be conditioned by phonological factors only, consistent with the analysis of suffixed words that was given to account for the tapping facts—again, morphology overrides frequency. In the third case, nasal substitution, phonological factors made frequency effects invisible except in a restricted group of cases. In sum, the application of phonological processes can be sensitive to frequency, and morphology and phonology can override frequency.

As a result of these findings, it was proposed that the grammar has access not just to a set of retrieved morphemes, but to information about which lexical entries were accessed to retrieve those morphemes.

References

Anderson, S. To appear. Second position clitics in Tagalog. In *The Nature of the Word*, ed. by S. Inkelas & K. Hanson. MIT Press.

Baayen, R. H. & R. Schreuder. 1999. War and peace: morphemes and full forms in a non-interactive activation parallel dual route model. *Brain and Language* 68.27–32.

Baroni, M. 2000. *Distributional Cues in Morpheme Discovery: A Computational Model and Empirical Evidence.* UCLA dissertation.

Baroni, M. 2001. The representation of prefixed forms in the Italian lexicon: Evidence from the distribution of intervocalic [s] and [z] in northern Italian. In *Yearbook of Morphology 1999*, ed. by G. Booij & J. van Marle, 121-152. Dordrecht: Springer.

Baroni, M. 2005. PotaModule. Software.

Benua, L. 1997. *Transderivational Identity: Phonological Relations between Words*. University of Massachusetts, Amherst dissertation.

Burzio, L. 1996. Surface Constraints versus Underlying Representation. In *Current Trends in Phonology: Models and Methods*. Vol. 1, ed. by J. Durand & B. Laks, 97-122. Salford: University of Salford Publications.

Bybee, J. 2001. *Phonology and Language Use.* Cambridge: Cambridge University Press.

Christophe, A., S. Peperkamp, C. Pallier, E. Block & J. Mehler. 2004. Phonological phrase boundaries constrain lexical access. *Journal of Memory and Language* 51.523-547.

Crosswhite, K. 1998. Segmental vs. prosodic correspondence in Chamorro. *Phonology* 15.281-316.

Davis, M. H. & W. Marslen-Wilson. 2002. Leading up the lexical garden path: segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance* 28.218-244.

English, L. 1986. *Tagalog-English Dictionary*. Manila: National Book Store.

Ghani, R., R. Jones & D. Mladenić. 2004. Building minority language corpora by learning to generate Web search queries. *Knowledge and Information Systems* 7.56-83.

Guevara, R. C., M. Co, E. Espina, I. D. Garcia, E. Tan, R. Ensomo, & R. Sagum. 2002. Development of a Filipino speech corpus. In *Proc. 3rd National ECE Conference*.

Hay, J. & R. H. Baayen. 2002. Parsing and productivity. In *Yearbook of Morphology 2001*, ed. by G. Booij & J. van Marle, 203-235. Dordrecht: Kluwer Academic Publishers.

Hay, J. 2003. *Causes and Consequences of Word Structure.* New York & London: Routledge.

Hayes, B. 1989. The Prosodic Hierarchy in meter. In *Rhythm and Meter*, ed. by P. Kiparsky & G. Youman, 201-260. Orlando, FL: Academic Press.

Hayes, B., B. Tesar & K. Zuraw. 2003. OTSoft. Software.

Kenstowicz, M. 1996. Base identity and uniform exponence: alternatives to cyclicity. In *Current Trends in Phonology: Models and Methods*. Vol. 1, ed. by J. Durand & B. Laks, 363-393. Salford: University of Salford Publications.

Kiparsky, P. 1982. Lexical morphology and phonology. In *Linguistics in the Morning Calm*, ed. by I.-S. Yang, 3-91. Seoul: Hanshin.

Martin, A. 2005. Loanwords as pseudo-compounds in Malagasy. In *Proc. 12th Annual Conference of the Austronesian Formal Linguistics Association*, *UCLA Working Papers in Linguistics* 12, ed. by J. Heinz & D. Ntelitheos, 287-295.

McCarthy, J. & A. Prince 1993.Generalized alignment. In *Yearbook of Morphology 1993*, ed. by G. Booij & J. van Marle, 79-153. Dordrecht: Kluwer.

McCarthy, J. & A. Prince. 1995. Faithfulness and reduplicative identity. *Papers in Optimality Theory, UMass Occasional Papers in Linguistics* 18.249-348.

Mohanan, K.P. 1986. *The Theory of Lexical Phonology*. Dordrecht/Boston: D. Reidel Publishing Company.

Nespor, M. & I. Vogel 1986. *Prosodic Phonology.* Dordrecht: Foris.

Peperkamp, S. 1997. *Prosodic Words.* The Hague: Holland Academic Graphics.

Pesetsky, D. 1979. Russian morphology and lexical theory. Ms., MIT.

Pierrehumbert, J. 2002. Word-specific phonetics. In *Laboratory phonology VII*, ed. by C. Gussenhoven & N. Warner (eds.), 101-140. Berlin/New York: Mouton de Gruyter.

Prince, A. & P. Smolensky 1993/2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Malden, MA: Blackwell. Originally circulated (1993) as Technical Report TR-2 (Rutgers Center for Cognitive Science)/Technical Report CU-CS-696-93 (University of Colorado at Boulder Department of Computer Science).

Raffelsiefen, R. 1999a. Diagnostics for prosodic words revisited: the case of historically prefixed words in English. In *Studies on the Phonological Word*, ed. by T. A. Hall & U. Kleinhenz, 133-203. Amsterdam: Benjamins.

Richards, N. 2006. Beyond strength and weakness. Ms., MIT.

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

Selkirk, E. 1980. Prosodic domains in phonology: Sanskrit revisited. In *Juncture*, ed. by M. Aronoff & M.-L. Kean, 107-129. Saratoga, CA: Anma Libri.

Steriade, D. 2000. Paradigm Uniformity and the phonetics-phonology boundary. In *Papers in Laboratory Phonology* Vol. 5, ed. by M. Broe & J. Pierrehumbert, 313-334. Cambridge: Cambridge University Press.

Trubetzkoy, N. 1939. *Grundzüge der Phonologie*. Göttingen: Vandenhoek & Ruprecht.

Urbanczyk, S. 2001. *Patterns of Reduplication in Lushootseed*. New York: Garland.

Zuraw, K. 2000. *Patterned Exceptions in Phonology.* Ph.D. dissertation, UCLA.

Zuraw, K. 2002. Aggressive reduplication. *Phonology* 19.395-439.

Zuraw, K. 2006. Using the Web as a phonological corpus: a case study from Tagalog. In *Proceedings of the 2nd International Workshop on Web as Corpus (EACL06)*, ed. by A. Kilgarriff & M. Baroni, 59-66.