Nat Lang Linguist Theory (2010) 28: 417–472
DOI 10.1007/s11049-010-9095-z

A model of lexical variation and the grammar with application to Tagalog nasal substitution

Kie Zuraw

Received: 8 April 2008 / Accepted: 2 June 2009 / Published online: 29 April 2010
© The Author(s) 2010. This article is published with open access at Springerlink.com

Abstract This paper presents a case of patterned exceptionality. The case is Tagalog nasal substitution, a phenomenon in which a prefix-final nasal fuses with a stem-initial obstruent. The rule is variable on a word-by-word basis, but its distribution is phonologically patterned, as shown through dictionary and corpus data. Speakers appear to have implicit knowledge of the patterning, as shown through experimental data and loan adaptation. A grammar is proposed that reconciles the primacy of lexical information with regularities in the distribution of the rule. Morphologically complex words are allowed to have their own lexical entries, whose use is preferred to on-the-fly morphological concatenation. The grammar contains lower-ranked markedness constraints that govern the behavior of novel words. Faithfulness for lexicalized full words is ranked high, so that an established word will have a stable pronunciation. But when a word is newly coined through affixation, the outcome varies according the lexical trends. A crucial aspect of the proposal is that the ranking of the "subterranean" markedness constraints can be learned despite training data in which all words are pronounced faithfully, using Boersma's (1997, 1998) Gradual learning algorithm. The paper also shows, by summarizing the rule's behavior in related languages, that the same constraints, in different rankings, seem to be at work even in languages reported to lack variation.

Keywords Tagalog · Nasal substitution · Lexical variation · Exceptions

1 Introduction

In studying variation in phonology, researchers distinguish **free variation** from what we might call **lexical variation**. In free variation, each word has more than one pronunciation. For example, for many speakers of American English, an *nt* sequence

K. Zuraw (✉)
University of California, Los Angeles, Los Angeles, USA
e-mail: kie@ucla.edu

can optionally be pronounced roughly as a nasal tap when the following vowel is unstressed, as in [ˈwɪntɚ]∼[ˈwɪ̃ɾɚ] 'winter'. In implementations of free variation (see references below), words cannot idiosyncratically choose just one variant—variation applies uniformly across the lexicon.

In lexical variation, by contrast, most individual words have a fixed pronunciation (though some words may have more than one variant), and variation is seen mainly across the lexicon, not in pronunciations of a single word. An example is liaison in French: for some words, a final consonant appears only when a vowel-initial word follows within a phonological phrase (*peti[Ø]* 'small', *peti[t] enfant* 'small child'); for other words a final consonant always appears (*prétéri[t]* 'preterite'); and for others no final consonant ever appears (*joli[Ø] enfant* 'pretty child'). There are only a few words for which there is within- or across-speaker variation (*bu[Ø]* ∼ *bu[t]* 'goal'). This type of variation is sometimes called **lexical conditioning**, or, when one variant is much less frequent, **exceptionality**. Lexical variation can also occur in the absence of alternation. In English, for example, nasal-obstruent clusters usually agree in place of articulation morpheme-internally: compare typical *antler* to atypical *femtoliter*.

Various mechanisms have been proposed to deal with free variation, and a thorough review of this literature will not be attempted here. In rule-based frameworks a rule may be marked as applying optionally, with its rate of application potentially dependent on other linguistic or extralinguistic factors (Weinreich et al. 1968; Labov 1969). In Optimality Theory (OT; Prince and Smolensky 1993/2004) constraints have been designated as freely ranked (Anttila 1997, 2002 and others; Reynolds and Nagy 1994; Nagy and Reynolds 1997; Ross 1996) or probabilistically ranked (Boersma 1997, 1998; Hayes and MacEachern 1998; Boersma and Hayes 2001). In all these approaches, the grammar applies in the same way to every lexical item, so that the same variation is predicted to be available for every lexical item. This is probably an idealization. For example, see Coetzee and Pater's (2008) data on English *t*/*d* deletion, which shows word-specific effects even beyond those of usage frequency.

Lexical variation has less often been tackled. In the French case, it suffices to somehow make the lexical entries for the three types of word different, whether through a representational distinction (Tranel 1987 for French specifically), listing of allomorphs (Tranel 1996 for suppletive cases in French), exceptionality diacritics (Chomsky and Halle 1968; Zonneveld 1978), or some other means. In the English place-assimilation case, we could similarly mark *femtoliter* as an exception to place assimilation. Or, we could allow assimilation to apply only to nasals of unspecified place, letting the nasal of *antler* be underlyingly unspecified for place, but the nasal of *femtoliter* be underlyingly labial (following Inkelas et al. 1997)—or simply give up on capturing static generalizations and let the underlying form determine the surface pronunciation, with the prevalence of place agreement left unaccounted for, as a lexical accident that plays no role in the speaker's grammar.

This paper presents a case where an exception-marking approach to lexical variation is not sufficient, because the distribution of the "exceptions" themselves is phonologically patterned and speakers appear to have implicit knowledge of the patterning. In Zuraw 2000, this phenomenon was called **patterned exceptionality**. The case is Tagalog nasal substitution, a phenomenon in which a prefix-final nasal affects a stem-initial obstruent. The rule is lexically variable—some words undergo it

and some don't—but there are trends within that variation. Voiceless obstruents are more likely to undergo nasal substitution than are voiced, and obstruents with a more front place of articulation are more likely to undergo it, especially within the voiced obstruents.

In response, an OT grammar is proposed that reconciles the primacy of lexical information with regularities in the distribution of the rule. The key ingredients of the proposal are (i) that morphologically complex words can have their own lexical entries, (ii) that use of such lexical entries is preferred to on-the-fly morphological concatenation (Aronoff 1976; Kiparsky 1982), and (iii) that a grammar can contain lower-ranked markedness constraints that govern the behavior of novel words. In the Tagalog case, faithfulness for lexicalized full words is ranked high, so that an established word will have a stable pronunciation. But when a word is newly coined through affixation, different, lower-ranked faithfulness constraints apply; as discussed in Section 3, in this case the difference in faithfulness constraints results from the representation of the prefix with a floating feature. These faithfulness constraints are variably ranked with respect to various markedness constraints, so that the outcome varies, but with a tendency towards the lexical trends encoded by those markedness constraints and their rankings. A crucial aspect of the proposal is that the ranking of the "subterranean" markedness constraints can be learned despite training data in which all words are pronounced faithfully, using Boersma's (1997, 1998) Gradual learning algorithm.

The paper is structured as follows. Section 2 presents the basic nasal-substitution data, and uses dictionary and corpus data to illustrate the voicing effect and the place-of-articulation effect. Section 3 proposes constraints responsible for those trends, as well as a basic analysis of nasal substitution. Section 4, which includes additional data on lexical idiosyncrasy, presents the full model of how fixed pronunciations for known words coexist in the grammar with information about lexical trends, along with learning simulations. Section 5 presents the evidence, from experimental tasks and loan adaptation, that the lexical trends should actually be represented in the grammar. Finally, Section 6 briefly surveys the patterns of nasal substitution across Western Austronesian to show that the same constraints seem to be at work even in languages reported to lack variation, but that their ranking can vary.

The Tagalog data in this paper come, except where noted, mainly from English's 1986 Tagalog-English dictionary and from a corpus of approximately 20 million Tagalog words gathered from the web (see Section 2.3 for details). Unless otherwise noted, examples were included only if attested in the corpus; some examples appear in the corpus only, not in English's dictionary. Any frequencies given are, of course, from the corpus. Most data are given in broad IPA transcription (International Phonetic Association 1999); when spellings are given, they are enclosed in angled brackets (⟨⟩).

2 Nasal substitution in Tagalog

2.1 The alternation

The phoneme inventory of Tagalog is given in (1) (see, for example, Schachter and Otanes 1972). The d/ɾ, i/e, and u/o contrasts are robust only in loans. Loans have

introduced other sounds for some speakers, especially [f, t͡ʃ, ʃ, d͡ʒ, ɹ]. The coronal obstruents [s, t, d] are often palatalized before [j].

(1)

p	t		k	ʔ	i	u
b	d		g		e	o
	s			h		a
m	n		ŋ			
	l					
	ɾ					
w		j				

A process known as nasal substitution occurs frequently, though somewhat unpredictably, in Tagalog words, and is conditioned to a large extent by the initial obstruent of the stem. When certain prefixes attach to a stem beginning in a sonorant other than [l], they appear as *paŋ-*, *maŋ-*, or *naŋ-* (which is derived morphologically from *maŋ-*), as in (2a). Before [l] (2b) and, in loans, [ɾ] (2c), they usually appear instead as *pan-*, *man-*, or *nan-*, though this place assimilation is not obligatory, especially before [ɾ].

(2)

		stem		affixes	affixed form	
a.	*h*	**h**ukbó	‘army’	paŋ-	pa**ŋ-h**ukbó	‘military’
	m	**m**aɾká	‘mark’	paŋ-	pa**ŋ-m**aɾká	‘marker’
	n	**n**egósjo	‘business’	paŋ-	pa**ŋ-n**egósjo	‘for business’
	ŋ	**ŋ**álit	‘grinding of teeth’	paŋ-RED-	pa**ŋ-ŋ**a-**ŋ**álit	‘grinding of teeth’
	w	**w**isik-án	‘to sprinkle on’	paŋ-	pa**ŋ-w**isík	‘sprinkler’
	j	**j**amót	‘annoyance’	maŋ-	ma**ŋ-j**amót	‘to annoy’
b.	*l*	**l**abás	‘exterior’	paŋ-	pa**n-l**abás	‘external’
c.	*ɾ*	**ɾ**ehjón	‘region’	paŋ-	pa**n-ɾ**ehjón	‘regional’

But when these prefixes attach to an obstruent-initial stem (3), there are two options. First, they can behave as they do before sonorants, with place assimilation to the obstruent usually applying, so that they appear as *pam-/pan-/paŋ-*, *mam-/man-/paŋ-*, and *nam-/nan-/paŋ-* (e.g., ***p**oʔók* ‘district’, *pa**m-p**oʔók* or less typical *paŋ-**p**oʔók* ‘local’). This is shown in the first example for each consonant in (3). This paper does not attempt to describe or analyze variation in application of nasal assimilation, mainly because there are too few examples of dictionary-listed words showing up in the corpus (see Section 2.3) with an unassimilated variant (only 31 words out of 1,107).

The second option is for the final nasal of the prefix and the initial obstruent of the stem to be both replaced by a nasal that is homorganic to the original obstruent. This second option is known variously as nasal substitution (the term I will use), nasal replacement, nasal coalescence, and nasal fusion. There is lexical variation, with some words consistently displaying nasal substitution, some consistently not, and some varying.

(3)

	stem		affixes	affixed form	
p	**p**oʔók	‘district’	paŋ-	pa**m-p**oʔók	‘local’
	pighatíʔ	‘grief’	paŋ-RED-	pa-**m**i-**m**ighatíʔ	‘being in grief’
t	**t**abój	‘driving forward’	paŋ-	pa**n-t**abój	‘to goad’
	tiwálaʔ	‘faith’	ka-paŋ- -an	kà-pa-**n**iwálaʔ-an	‘traditional belief’
s	**s**úlat	‘writing’	paŋ-	pa**n-s**úlat	‘writing instrument’
	súlat	‘writing’	maŋ-RED-	mà-**n**u-**n**ulát	‘writer’
k	**k**úlam	‘sorcery’	maŋ-RED-	ma**ŋ-k**u-**k**úlam	‘witch’
	kamkám	‘usurpation’	ma-paŋ-	ma-pa-**ŋ**amkám	‘rapacious’
ʔ	**ʔ**ulól	‘silly’	maŋ-	ma**ŋ-ʔ**ulól	‘to fool someone’
	ʔisdáʔ	‘fish’	maŋ-	ma-**ŋ**isdáʔ	‘to fish’
b	**b**igkás	‘pronouncing’	maŋ-RED-	ma**m-b**i-**b**igkás	‘reciter’
	mag-**b**igáj	‘to give’	maŋ-	ma-**m**igáj	‘to distribute’
d	**d**iníg	‘audible’	paŋ-	pa**n-d**iníg	‘sense of hearing’
	daláŋin	‘prayer’	i-paŋ- -in	ʔi-pa-**n**aláŋinin	‘to pray’
g	**g**áwaj	‘witchcraft’	maŋ-RED-	ma**ŋ-g**a-**g**áwaj	‘witch’
	gindáj[1]	‘unsteadiness on feet’	paŋ-RED-	pa-**ŋ**i-**ŋ**indáj	‘unsteadiness on feet’

As (2) and (3) illustrate, there are several (impressionistically) productive morphological constructions that can participate in nasal substitution. In all of them, the prefix complex ends in *paŋ-*, *maŋ-*, or *naŋ-* . There are also some (again, impressionistically) unproductive constructions ending in a nasal. Their prefix complexes end in *taŋ-*, *tuŋ- siŋ-*, *hiŋ-*, *kaŋ-*, and *kuŋ-*, and some of them can trigger substitution, as illustrated in (4). Although no substituting examples were found for three of the prefixes, the number of total cases is so small that the gaps may be accidental. The fairly productive construction *mag-kaŋ-RED*, for verbs of accidental result (*dapáʔ* ‘face down’, *mag-kan-da-ɾápaʔ* ‘to fall on one’s face’), never produces substitution, despite containing *kaŋ-*.

(4)

taŋ-	*(no substituting examples found)*			
	bílaŋ	‘number’	ta**m-b**ílaŋ (not in corpus)	‘digit’
tuŋ-	*(no substituting examples found)*			
	balík	‘upside-down’	tu**m-b**alík	‘return’
siŋ-	púnoʔ	‘leader’	si-**m**únoʔ	‘grammatical subject’
	tábiʔ	‘move aside!’	pa-si**n-t**ábiʔ	‘respect; asking pardon’
hiŋ-	kúto	‘louse’	hi-**ŋ**utú-han	‘to pick out lice’
	túlot	‘permission’	pa-hi**n-t**úlot	‘permission’
kaŋ-	patáj	‘corpse’	ka-**m**àtáj-an	‘death’
	gatáʔ	‘coconut milk’	kà-ka**ŋ-g**atáʔ	‘first extraction of coconut milk; essence’
kuŋ-	*(no substituting examples found)*			
	babáʔ	‘descent’	mag-pa-ku**m-b**abáʔ	‘humble’

[1]Neither the bare root *gindáj* nor any of its derivatives appear in the corpus. This is the only instance of substitution of *g* found in English’s dictionary.

This exhausts the prefixes that end in *ŋ*, as far as I know. With the exception of *mag-kaŋ-RED*, then, all nasal-final prefixes can trigger nasal substitution (there are no prefixes ending in /m/ or /n/). There are some other morphemes ending in *ŋ* that are sometimes described as prefixes, but I believe they should instead be regarded as stems that can form compounds: *waláŋ-* 'not exist', *(ʔi)sáŋ-* 'one', *(ka)síŋ-* 'as X as', *pagigíŋ-* 'becoming', and *magíŋ-* 'become', illustrated in (5). The reasons for regarding these as compounding elements are that they are all at least two syllables long in their full forms, can bear their own stress,[2] produce semantically transparent words, never induce nasal substitution, and often fail to undergo nasal place assimilation. In addition, *waláŋ-* and *(ʔi)sáŋ-* are presumably derived from the freestanding words *walá ʔ* 'does not have/exist' and *ʔisá ʔ* 'one', plus the "linker" *-ŋ-*. Forms with *magíŋ-* are usually spelled as two separate words (e.g., ⟨maging abogado⟩ = [magíŋ-abogádo]).

(5)	bájad	'payment'	waláŋ-**b**ájad	'free'
	dáliʔ	'finger-width'	sa**n**-**d**áliʔ	'one finger width'
	ʔitím	'black'	kasíŋ-**ʔ**itím	'as black as'
	táʔo	'person'	pagigíŋ-**t**áʔo	'becoming a person'
	ʔabogádo	'lawyer'	magíŋ-**ʔ**abogádo	'to become a lawyer'

A few remarks on the examples above in (2) and (3) are needed before moving on: First, when nasal substitution occurs, the resulting nasal is part of the base of reduplication: /paŋ-RED-pighatíʔ/ becomes *pa-**m**i-**m**ighatíʔ* rather than **pa-**m**i-**p**ighatíʔ*, with a nasal occurring only adjacent to the triggering prefix. Various explanations have been proposed for this double application of substitution: that nasal substitution precedes reduplication, in a counterbleeding order (Bloomfield 1917; Carrier 1979; Raimy 2000); that both reduplicant and base select a nasal-substituted allomorph because of the morphological context (Marantz 1982; Inkelas and Zoll 2000, 2005); or that a special relationship between base and reduplicant forces nasal substitution to apply to both (Wilbur 1973; McCarthy and Prince 1995).

What is important about the reduplicated cases for our purposes is their bearing on the affiliation of the nasal resulting from substitution. The reduplicated forms suggest that when nasal substitution applies, the resulting nasal belongs to a stem that is used as the base of reduplication. Thus *pa-**m**ighatíʔ* has the stem ***m**ighatíʔ* and is reduplicated *pa-**m**i-**m**ighatíʔ*. If the structure were *pa**m-**ighatíʔ*, with a stem *ighatíʔ*, we would expect reduplicated **pa**m-**i-(ʔ)ighatíʔ*. Conversely, when nasal substitution doesn't apply, the nasal is not part of the base of reduplication: *ma**m**-bigkás* has the stem *bigkás* and is reduplicated *ma**m**-bi-bigkás*. If the structure were *ma-**m**bigkás*, with a stem ***m**bigkás*, we would expect reduplicated **ma-**m**bi-**m**bigkás*. This claim about the morphological structure of nasal-substituted words will figure in the analysis.

Second, it is not clear whether nasal substitution is possible on nasal-initial stems. Nasal-initial stems are rare to begin with, and among those that do exist, it is not always possible to tell what the prefix is. For example, in *ma-manhíd* 'to become

[2] I transcribe the stresses of these elements as primary rather than secondary because, impressionistically, they seem to be associated with pitch-accents, whereas the stresses transcribed here as secondary do not.

numb', from *manhíd* 'numb', it is not clear whether the prefix is *maŋ-*, with nasal substitution, or simply *ma-* (which can also form verbs, with similar semantics).[3] There do exist unambiguous constructions, but I have found no cases of nasal-initial stems in them. For these reasons, nasal-initial stems are not included in the data and analysis below.

Third, glottal stop is problematic. Word-final *ʔ* contrasts with zero in Tagalog (*bága* 'ember', *bágaʔ* 'lung'). Initial *ʔ* does not contrast with zero, however—there are no strictly vowel-initial words in citation form—so many researchers have treated *ʔ* as predictably inserted at the beginnings of vowel-initial words or perhaps phrases. The preservation of stem-initial glottal stop in prefixed words like *mag-**ʔ**áwaj* 'to fight' (or *ma**ŋ-ʔ**ulól* in (3)) would then be explained as a failure to resyllabify across the prefix-stem boundary, and apparent nasal substitution as in *ma**ŋ**isdáʔ* (3) would represent mere resyllabification (/maŋ-isdáʔ/ → [ma.**ŋ**-is.dáʔ]) rather than true nasal substitution. The variation would concern syllabification rather than nasal substitution. An additional mechanism would be necessary to explain copying of the nasal in reduplicated forms: *mà-**ŋ**i-**ŋ**isdáʔ* 'fisher' (see Bhandari 1997 on nasal substitution-reduplication interactions for vowel-initial stems cross-linguistically). Data for *ʔ* are included in the figures below, but *ʔ* is not included in the analysis because of its unclear status.[4] See French 1988; Ross 1996; Boersma 1998 (Chapter 9), and Halle 2001 for further discussion of initial *ʔ* in Tagalog.

Having laid out the basics of substitution and non-substitution, we can proceed to their distribution in the lexicon.

2.2 Nasal substitution in the dictionary

As was seen in (3), for every stem-initial obstruent there exist words that undergo nasal substitution and words that don't. The distribution of nasal substitution is far from even, however. Examining the non-loan words from English's (1986) dictionary[5] that have an obstruent-initial stem and a potentially nasal-substituting prefix (see Section 5.3 for loans), we find two trends. First, substitution is more likely if the stem-initial consonant is voiceless than if voiced. For example, as shown in Fig. 1, 253 out of 263 *p*-initial stems undergo substitution (96%), whereas 177 out of 277 of *b*-initial stems do (64%). Second, among the voiced consonants, substitution is most

[3]See Schachter and Otanes (1972) and Carrier (1979) for arguments bearing on this question. Schachter and Otanes argue that the prefix of a verb like *ma-manhíd* is indicated by the gerund form, but Carrier refutes the claim. Carrier argues that these nasal-initial stems are not nasal-substituted, because some of them are clearly unsubstituted when prefixed with *paŋ-* (*paŋ-noʔód* 'for watching'). But, as we will see below, a stem may show different nasal-substitution with different affixes.

[4]Carrier (1979) considers and rejects the idea that there is a contrast between underlyingly glottal-stop-initial and underlyingly vowel-initial stems, which determines whether nasal substitution will appear to occur. Some glottal-initial words of Tagalog derive diachronically from Proto-Malayo-Polynesian forms that have been reconstructed as **q*-initial, and others derive from PMP forms that have been reconstructed as vowel-initial (see, e.g., entries beginning with *i* in Zorc 1979). I have not investigated whether this etymological difference is predictive of nasal-substitution behavior.

[5]Figure 1 includes all the data from English 1986, not just those items that appear in the corpus.

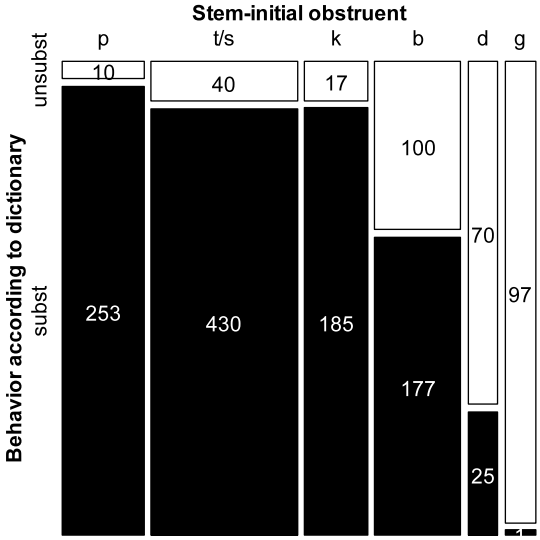


Fig. 1 Rates of nasal substitution for entire lexicon—dictionary data

likely with *b* and least likely with *g*.[6] Figure 1 combines data from all constructions (*t* and *s* are also combined, to better illustrate the two trends, and *ʔ* is omitted; in the more detailed figures that follow, *t* and *s* are separated and *ʔ* is included). The figure is a mosaic plot, made using the *mosaic()* function of the *vcd* package (Meyer et al. 2006, 2007) of the statistical computing program R (R Development Core Team 2007). The widths of the columns are scaled so that the area of each "tile" is proportional to the number of tokens in that tile. Thus, the columns for *d* and *g* are narrow because there are relatively few words in this set with *d*- or *g*-initial stems.

Different constructions have different overall substitution rates, but all follow—or at least do not contradict—the generalizations about voicing and place. Figure 2 through Fig. 7 show the dictionary data for the six most common affix patterns, accounting for 1,670 of the 1,736 words in the dictionary. The breakdown by affix is based in part on De Guzman (1978). She distinguishes adversative verbs, which are hostile to the patient (*ma-mató* ∼ *mam-bató* 'to throw stones at'), from other verbs, including inchoative, stative, professional, habitual, distributive, and repetitive verbs, and others. De Guzman also distinguishes instrumental adjectives (*pa-nítik* 'used for writing') from reservative adjectives (*pam-baŋkéte* 'appropriate for a banquet'). When a category has no members (a zero count), such as *p*-initial stems without substitution in Fig. 2, a line with a circle appears. The grey tiles represent words listed in the dictionary as variable.

The constructions illustrated in Fig. 2 through Fig. 7 are *paŋ-RED-*, which forms mainly gerunds (*tahíʔ* 'stitch', *pa-na-nahíʔ* 'sewing'), but also some less transparent

[6]Previous accounts of the lexical distribution of nasal substitution have stated, mostly in passing, that *g* never substitutes (Bloomfield 1917; Schachter and Otanes 1972); that *d* and *g* rarely substitute (Blake 1925); that voiceless consonants substitute more than voiced ones (De Guzman 1978); and that morphology matters (Schachter and Otanes 1972; De Guzman 1978, who gives detailed claims about various morphological constructions).

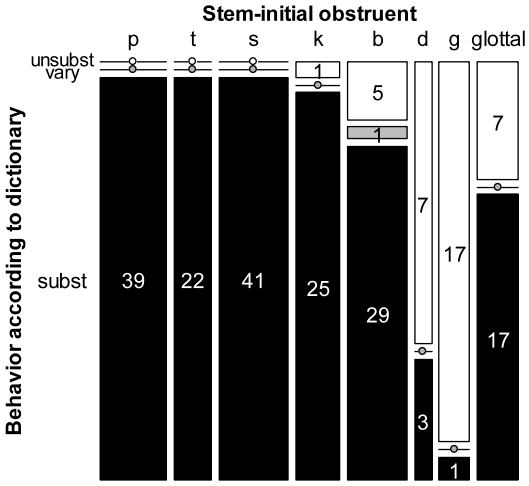


Fig. 2 Rates of substitution for *paŋ-RED-* construction

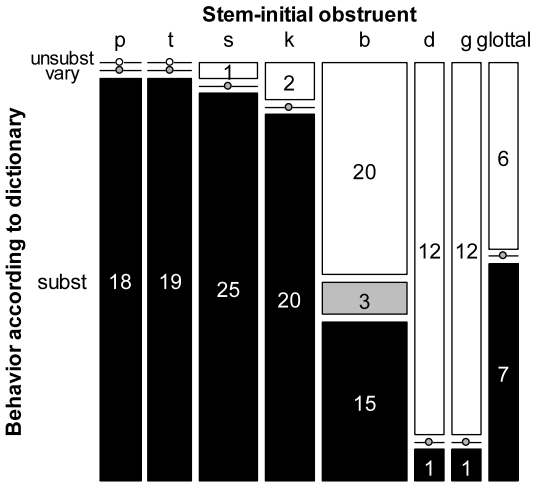


Fig. 3 Rates of substitution for *maŋ-RED-* construction

nominalizations; *maŋ-RED-*, which forms professional or habitual nouns (*bátas* 'law', *mam-ba-batás* 'legislator'); adversative-verb-forming *maŋ-*; non-adversative-verb-forming *maŋ-*; noun-forming *paŋ-* (instrumentals, gerunds, and other nominalizations, e.g., *gúgol* 'expense', *paŋ-gúgol* 'spending money'); and reservative-adjective-forming *paŋ-*. No other constructions had enough examples of each obstruent to make a chart meaningful. In Fig. 6, where overall rates of substitution are lower, there is a suggestion of a place effect among the voiceless obstruents as well as the voiced.

2.3 Nasal substitution in a written corpus

Relying on dictionary data has drawbacks. The dictionary may include archaic words, or omit newer words. And it is hard to know what to make of dictionary pronunciations. High-budget dictionaries of major world languages may have elaborate systems

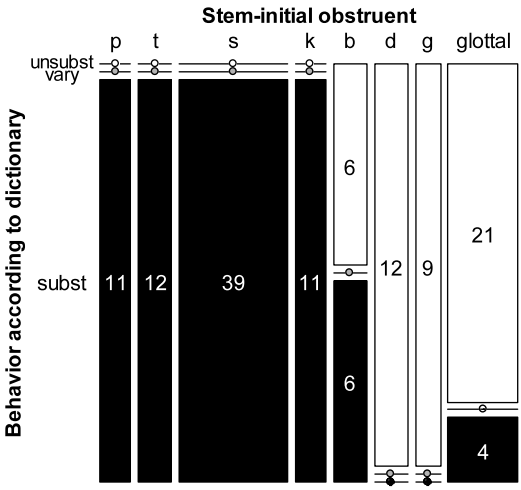


Fig. 4 Rates of substitution for *maŋ-* (adversative) construction

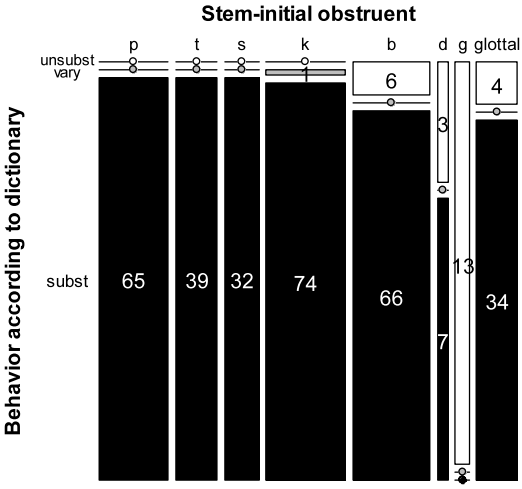


Fig. 5 Rates of substitution for *maŋ-* (other) construction

for tracking acceptable pronunciations; Merriam-Webster, for example, maintains a "pronunciation file" containing transcriptions collected by pronunciation editors from live and broadcast speech (Merriam-Webster 1994). Such methods are not feasible in dictionaries of most of the world's languages. English (1986) is the work of an English-speaking priest living in the Philippines and six or more Tagalog-speaking colleagues and assistants. The dictionary does not say how pronunciations were arrived at, but presumably they reflect the judgments of those persons. We might worry that this is too small a sample of speakers, or that the pronunciations given are biased towards normative rather than colloquial pronunciations, or simply that limited editing time allowed errors to slip through.

This section presents corroborating data drawn from a written corpus, which has drawbacks of its own, but complements and corroborates the dictionary data. The corpus was created by sending queries to the web search engine Google

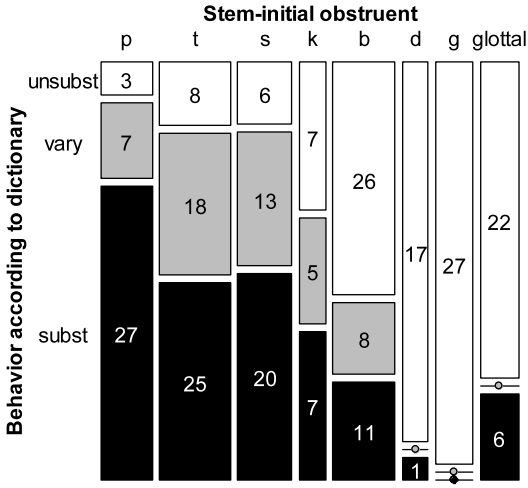


Fig. 6 Rates of substitution for *paŋ-* (noun) construction

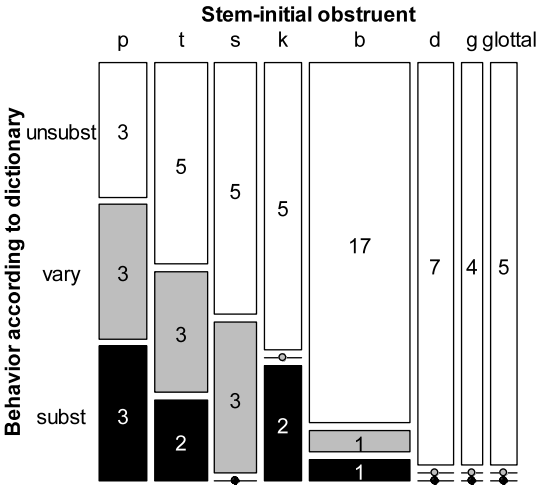


Fig. 7 Rates of substitution for *paŋ-* (reservative) construction

(www.google.com), using the Google APIs service and software written by Ivan Tam. Queries were designed to retrieve pages containing Tagalog language text. This was done by taking a smaller demonstration corpus consisting mainly of Tagalog, generously supplied by Rosie Jones (derived from Ghani et al. 2004, whose idea inspired the procedure used here) and finding Tagalog words with high frequency in that corpus. Queries were composed automatically by selecting words from that list (with probabilities in proportion to their frequencies) to form a string such as ⟨sa hindi ang⟩, which, when sent to Google as a query, finds pages that have all three of those words, not necessarily adjacent or in that order. Unlike in Ghani et al.'s approach, negative search terms from other languages were not used (e.g., exclude *the*), so as not to exclude pages written in a mixture of Tagalog and another language. The HTML contents of the web pages returned by the Google search were then automatically retrieved. The resulting corpus contains approximately 20 million words of Tagalog,

although of course most of these tokens are not relevant to nasal substitution. (See Zuraw 2006 for more details.)

This corpus has the advantage of being large and reflecting a variety of writing styles, from extremely colloquial to formal. It has the disadvantages of any written corpus in that writing allows time for reflection—and even editing by another person—so that what we see on the page may not be what the speaker would spontaneously utter. We may also wonder if spellings are an accurate reflection of pronunciations. In the case of nasal substitution, it seems highly unlikely that a Tagalog speaker would write ⟨pampook⟩ to represent the pronunciation [pa-moʔók], or ⟨pamimighati⟩ to represent [pam-pi-pighatíʔ], but still this might be a concern. There may also be typographical errors, especially since most of the text here is not edited. The writing reflects a mix of many authors, who may speak different dialects, and some of whom may not even be native speakers of Tagalog.

Clearly there are drawbacks to both the dictionary and the corpus data. If they agree on the patterns of nasal substitution, however, we can be more confident that those patterns are robust. In order to test agreement between the two sources, all potentially nasal-substituted items from the dictionary were paired with the alternative pronunciation's spelling and both were searched in the corpus. For example, if the dictionary contains *pam-poʔók*, the corpus was searched for both ⟨pampook⟩ (and unassimilated ⟨pangpook⟩) and hypothetical ⟨pamook⟩. Aspectual conjugations of verbs were also included, as well as forms with the "linker" added.[7] This was done instead of searching the corpus for all potentially nasal-substituted words, because identifying words with nasal-substituting affixes—as opposed to words that accidentally contain strings like *pam* and *man*—would require a prohibitive amount of hand-checking and examination in context, as would identifying the underlying stem-initial obstruent in words that have undergone nasal substitution (since *p* and *b* are neutralized, as are *t/d/s* and *k/g/ʔ*). Any words that were entirely absent from the corpus, occurring with neither the dictionary spelling nor the alternative spelling, were omitted from the results. Out of 1,715 dictionary words probed in this way (a few of the full 1,736 words were excluded because of problematic morphology), 1,107 were attested in the corpus in at least one variant, for a total of 195,513 tokens.

We can make a corpus-based chart to compare to the dictionary data by breaking up each word that appears in the corpus according to its corpus behavior: that is, if a *b*-initial word appears 30 times substituted and 30 times non-substituted in the corpus, it contributes 0.5 each to the substituted and non-substituted counts for *b*-initial stems. The corpus data are shown on the left in Fig. 8; on the right is repeated Fig. 1, except with *t* and *s* separated, and a column for the glottal stop. We see the voicing effect (more substitution for *p*, *t*, *s*, *k*) and, within the voiced obstruents, the place effect ($b > d >$ g). The agreement is very strong: correlating the percent-substituted for each of the six voicing/place categories in the corpus versus the dictionary, we obtain an R^2 of .988.

We can also examine the word-by-word agreement between the dictionary and the corpus. For words described in the dictionary as not substituting, 90% (287/318) of

[7] The linker ⟨ng/na⟩ is a morpheme that occurs between a noun and its modifier, and in some other contexts (see Schachter and Otanes 1972:118). Depending on the final segment of the first word, it may appear as a suffix or as an enclitic.

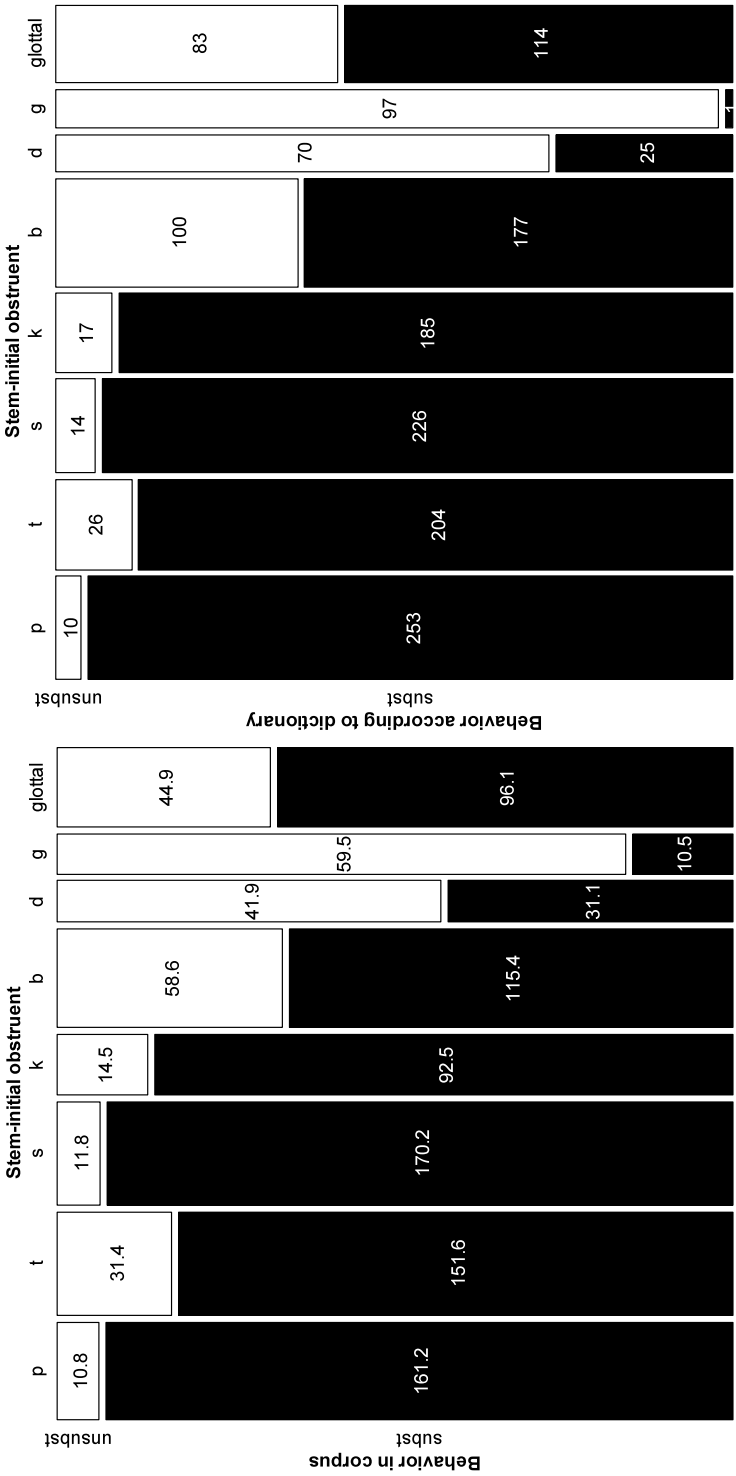


Fig. 8 Rates of nasal substitution in corpus vs. dictionary, native words only

Table 1 Dictionary claims non-substituted

consonant	total that appear in corpus	# that occur non-substituted		# that occur substituted	
p	40	40	100%	4	10%
t	20	18	90%	9	45%
s	7	6	86%	5	71%
k	18	15	83%	6	33%
b	57	48	84%	27	47%
d	50	45	90%	21	42%
g	72	67	93%	36	50%
ʔ	54	48	89%	29	54%

Table 2 Dictionary claims substituted

consonant	total that appear in corpus	# that occur non-substituted		# that occur substituted	
p	191	20	10%	190	99%
t	154	24	16%	152	99%
s	175	16	9%	174	99%
k	130	16	12%	124	95%
b	117	30	26%	109	93%
d	23	1	4%	23	100%
g	0				
ʔ	92	16	17%	92	100%

those that appear in the corpus at all have a non-substituted form that occurs with frequency of at least 1, but only 43% have a nasal-substituted form that appears in the corpus (the percentages do not sum to 100% because it is possible for both variants of a word to appear in the corpus). For words predicted by the dictionary to nasal-substitute, 98% (864/882) of those that appear in the corpus have a nasal-substituted form that appears, but only 14% have a non-substituted form that appears. The words listed in the dictionary as varying are in between, with 81% (45/58) occurring substituted and 76% appearing non-substituted. In all cases, the substituted figure is an overestimate, because of ambiguous nasal-substituted words like ⟨mamili⟩, which could represent *maŋ+bili* 'shop' or *maŋ+pili* 'choose'. (This overcount of nasal substitution also affects Fig. 8.)

Table 1 and Table 2 show that the accuracy of the dictionary's prediction is similar across consonants.

Since the two data sources agree on the voicing effect and the place effect within voiced obstruents, I will conclude that these patterns are genuinely present in the lexicon and move on to their analysis, after considering one last aspect of the dictionary and corpus data.

2.4 Nasal substitution and boundary strength

Nasal substitution seems to be negatively correlated with the degree to which a word is transparently prefixed.

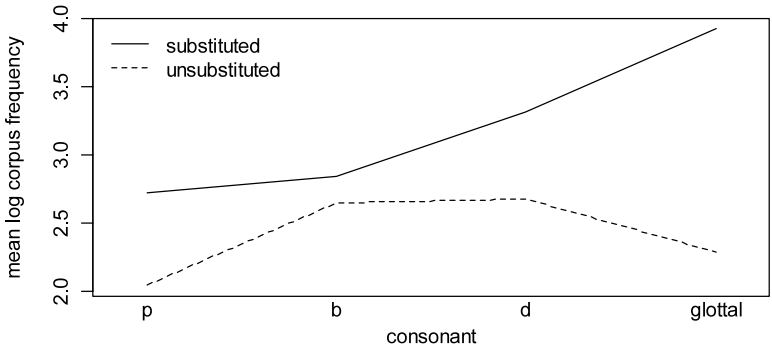


Fig. 9 Corpus frequencies for substituted and unsubstituted words

The first piece of evidence for this claim concerns meaning. Though I haven't conducted any systematic study of semantic opacity, we can see some trends in Fig. 2 through Fig. 7 that support a negative correlation. The lowest rates of substitution are found in the *paŋ-* (noun) and *paŋ-* (reservative adjective) constructions; the latter is almost always transparent, though the former varies. The highest rates of substitution are found in the miscellaneous-verb-forming *maŋ-* construction, whose semantics are very unpredictable, and the nominalizing *paŋ-RED-*, whose semantic transparency varies.

It seems plausible that higher-frequency words are more likely to be treated by speakers as whole units rather than as prefix-stem combinations. We can also use the corpus to compare frequencies in nasal-substituted versus unsubstituted words, and it turns out that nasal-substituted words have higher frequency on average. This can be seen in Fig. 9 for the four consonants that had at least 20 words in each of the two categories (substituted and not)—in all four cases, the mean frequency of the nasal-substituted words is higher. A linear regression with substitution status and initial consonant as independent variables finds that substitution status is a significant predictor of log corpus frequency ($p = .0005$).

Thus, although nasal substitution is a phenomenon that occurs only under prefixation, it seems to occur less often when the prefixation is transparent.

3 Analysis

This section presents an analysis of the voicing and place effects, and of nasal substitution itself, without yet tackling the question of lexical variation.

3.1 What drives nasal substitution?

Before turning to some questions about *NC̥, we must consider why nasal substitution applies at all. This section will present one analysis that is expositorily simple, then discuss some other possibilities.

There are two crucial properties, seen above, that any analysis of Tagalog nasal substitution must capture. First, nasal substitution occurs only with nasal-final prefixes;[8] nasal-obstruent clusters are very common morpheme-internally, with no obvious instability in their pronunciation.[9] Second, it seems nasal substitution is less likely to occur at "looser" morpheme boundaries—that is, with semantically transparent prefixes and low-frequency words.

An analysis based on an underlying floating feature (e.g., McCarthy 1983), in this case [+nasal], can capture both these properties: under this account, nasal substitution is due to the need to realize the prefix's nasal feature, with a dispreference for inserting an extra segment. This is illustrated in (6) (constraints discussed below).

(6)

		/p_1a_2[+nas]$_3$/+/$b_4i_5g_6a_7j_8$/	MAX(+nas)	DEP-C	*ASSOCIATE
☞	*a*	pa-$m_4$igaj | [+nas]$_3$			*
	b	pam$_9$-$b_4$igaj | [+nas]$_3$		*!	
	c	pa-$b_4$igaj	*!		

I adopt McCarthy and Prince's (1993, 1995) correspondence approach to faithfulness, with features treated autosegmentally, as in Zoll (1996). The correspondence indices on the winning candidate (*a*) are meant to indicate that the segment [m] corresponds to the segment /b/ of the input stem ("4"), but its nasal feature corresponds to the floating feature of the prefix ("3"). Candidate *b*'s nasal feature also corresponds to the underlying floating feature ("3"), but the segment itself has no input correspondent ("9"), in violation of DEP-C.

The faithfulness constraint violated by the winning candidate in (6), *ASSOCIATE$_{\text{hetero-morphemic}}$ (abbreviated *ASSOCIATE in all tableaux) is based on the more general *ASSOCIATE (e.g., Yip 2002, 2007), which penalizes addition of an association line. The more specific constraint used here penalizes adding an as-

[8] A possible diachronic account of how nasal substitution came to occur only at morpheme boundaries is as follows. It has not been established just when root-internal nasal-consonant clusters arose in Malayo-Polynesian languages (see Ross 1995:62–64, for an argument that in most cases these clusters were not present in proto-Austronesian). It is possible that before any clusters existed root-internally, there was a diachronic process of post-nasal deletion (**mam-bili* > **mamili*), stronger for some obstruents than others, whose environment happened to be met only at morpheme boundaries because of the lack of root-internal clusters. Only later did root-internal clusters enter the proto-language, through syncope, borrowing, and perhaps other means, leaving nasal substitution as a derived-environment rule. (This account must exempt pseudo-reduplicated roots from deletion, however.)

Alternatively, Herbert (1980) assumes that nasal substitution arose at a time when root-internal nasal-stop sequences did exist, and proposes that nasal substitution served to reinforce the morphological information supplied by the prefix. This might have been especially important in Proto-Malayo-Polynesian because of the danger of homophony among **{p,m}aN-*, **{p,m}aR-*, and **{p,m}a-*.

[9] Of course, phonetic study would be needed to confirm this. Herbert (1980) claims that there is some instability in nasal–voiceless-stop clusters in Malagasy, a related language that also has nasal substitution restricted to morphologically derived environments.

sociation line between material belonging to different morphemes—in this case, the association between the [+nas] feature of the prefix and the stem segment.

(7) *ASSOCIATE$_{\text{hetero-morphemic}}$: Do not associate new association lines between phonological units whose input correspondents belong to different morphemes.

*ASSOCIATE is not violated by the place assimilation of candidate (b) in (6)—even though an association line is presumably added between the stem-initial *b*'s [labial] feature and the nasal segment inserted into the prefix—because that inserted nasal segment has no input correspondent.

For the sake of legibility, I'll continue to refer to the prefixes as *paŋ-*, *maŋ-*, etc., rather than *pa*[+nas], *ma*[+nas], etc.

No nasal substitution is possible on sonorant-initial stems, because the resulting segment would be illegal (as discussed in Section 2.1 above, it is unknown whether nasal substitution can apply to nasal-initial stems). This is shown in (8).

(8)

/pa[+nas]/+/ligo/	* [+lateral, +nasal]	MAX(+nas)	DEP-C	*ASSOCIATE
a pa-l̃igo | [+nas]	*!			*
☞ *b* pan-ligo			*	
c pa-ligo		*!		

Under the OT concept of Richness of the Base (Prince and Smolensky 1993/2004), we must consider the possibility of an underlying floating [+nasal] feature within a root. According to the grammar here, this feature would nasalize a following obstruent (as in (9)), and if there was no following obstruent, it would surface as its own segment. But this would not appear, to the analyst or the learner, as a case of nasal substitution: there would simply be a non-alternating nasal segment in all realizations of that root morpheme.

(9)

/li[+nas]$_1$p$_2$a/ *(hypothetical)*	MAX(+nas)	DEP-C	*ASSOCIATE
☞ *a* lim$_2$a | [+nas]$_1$			
b lim$_6$p$_2$a | [+nas]$_1$		*!	
c lip$_2$a	*!		

If, on the other hand, a monomorpheme contains a nasal-obstruent cluster, it will surface faithfully even if doing so violates markedness constraints, such as NOCODA (Prince and Smolensky 1993/2004), because of high-ranking MAX-C and UNIFORMITY (McCarthy and Prince 1995), which forbids two distinct underlying segments from corresponding to a single surface segment:

(10)

	/bin$_1$t$_2$íʔ/ ‘calf’	MAX-C	UNIFORMITY	NOCODA
a	bin$_{1,2}$íʔ		*!	
☞ *b*	bin$_1$t$_2$íʔ			*
c	bit$_2$íʔ	*!		

Thus, the difference between morpheme-internal and prefix-stem-boundary environments here lies in the underlying representations of the prefixes, whose floating features can dock to another segment without violating UNIFORMITY. It might be desirable to adopt a more general solution for processes that apply only in derived environments, such as McCarthy’s (2002, 2003) Comparative Markedness, but the analysis above will be retained here for the sake of simplicity. If the lack of a timing slot is viewed as a form of underspecification, then there is a link to a more general principle proposed in Kiparsky (1993), where structure-building rules can apply only to underspecified representations

To account for nasal substitution’s reluctance to apply across a “loose” morpheme boundary, symbolized with # in the tableau below (though this should not be taken literally as the # boundary type of Chomsky and Halle 1968[10]), we can introduce the constraint MorphemeCohesion. This constraint is violated if the floating [+nasal] is separated from the rest of the prefix material by a # boundary.

(11) MORPHEMECOHESION: If X and Y are phonological units belonging to a single morpheme in the input, X corresponds to output X, and Y corresponds to output Y, then X and Y should not be separated by a # boundary.

Tableau (12) illustrates how this constraint forces the nasal feature to be associated with an inserted prefix segment.

(12)

	/pa[+nas]$_3$/ # /p$_4$ulítika/ ‘political’	MORPHEME COHESION	MAX(+nas)	DEP-C	*ASSOCIATE
a	pa # m$_4$ulítika | [+nas]$_3$	*!			*
☞ *b*	pam$_{12}$# p$_4$ulítika | [+nas]$_3$			*	
c	pa # p$_4$ulítika		*!		

I will assume that the locus of phonologically conditioned variation in nasal substitution is the words with + boundaries. That is, words with # boundaries are excluded from nasal substitution, but within words with + boundaries, the voicing and place effects (constraint for which are proposed below) apply.

[10]It may even be that boundary looseness or tightness is a continuum. In that case, the “#” referred to by MORPHEMECOHESION could represent boundaries at or above some threshold value of looseness; other implementations are imaginable also.

Let us return briefly to the prefixation of sonorant-initial stems. As noted in Section 2.1, the prefix nasal is assimilated in place to a following *l* (and, variably, *ɾ*), but otherwise, it shows up as *ŋ* before sonorants. In the case of *w*, *j*, *h*, or *ʔ*, this is plausibly because there is no legal nasal in the language that fully shares the place features of the sonorant, and velar is the default place of articulation for coda nasals in this language. A velar default is consistent with the (rarer) unassimilated variants before obstruents, such as ⟨mambabasa⟩ ∼ ⟨mangbabasa⟩ 'reader', from *b-um-ása* 'to read', which might reflect either variability in the ranking of assimilation versus default nasal-coda place, or the variable presence of a boundary strong enough to block assimilation (i.e., even stronger than the boundary that blocks nasal assimilation).

The lack of place assimilation before *m* and *n* is more problematic. Perhaps it is due to the general prohibition on geminate consonants within the language (geminates are never found within a morpheme in Tagalog).[11] That prohibition can be violated when there is no better option, as in *pa**ŋ-ŋ**a-ŋálit*: Tableau (13) assumes a # boundary that prevents the [+nas] from docking to the stem consonant. As noted in Section 2.1, it's hard to tell whether nasal-initial stems ever do undergo nasal substitution.

(13)

/pa[+nas]/#RED+/ŋálit/	IDENT (place)	*m]$_\sigma$	*n]$_\sigma$	*GEMINATE	NASASSIM	*ŋ]$_\sigma$
☞ *a* pa**ŋ**-ŋa-ŋálit				*		
b pa**m**-ŋa-ŋálit		*!			*	
c pa**n**-ŋa-ŋálit			*!		*	

IDENT(place) is included in the tableau above to emphasize that because there is no underlying nasal segment, only a floating feature, there is no IDENT(place) violation. The *N]$_\sigma$ constraints penalize only a coda nasal with its own place; candidate (a) has no violation of *ŋ]$_\sigma$ because I assume that the prefix *ŋ* shares its place with the following onset *ŋ* (see Itô 1986 on coda-place licensing). Compare (13) to the prefixation of an *m*-initial or obstruent-initial stem—again, there is no violation of *m]$_\sigma$ in either of the (b) candidates below, because the *m* shares its place with the following onset consonant:

(14)

/pa[+nas]/#/marká/	IDENT (place)	*m]$_\sigma$	*n]$_\sigma$	*GEMINATE	NASASSIM	*ŋ]$_\sigma$
☞ *a* pa**ŋ**-marká					*	*
b pa**m**-marká				*!		
c pa**n**-marká			*!		*	

[11] Besides words like *pa**ŋ-ŋ**a-ŋálit*, the other way that geminates can be created within a word is for a *g*-initial word to be prefixed with *pag-*, *mag-*, or *nag-*, e.g. *mag-gamót* 'to treat'.

(15)

/pa[+nas]/#/poʔók/	IDENT (place)	*m]$_\sigma$	*n]$_\sigma$	*GEMINATE	NASASSIM	*ŋ]$_\sigma$
a paŋ-poʔók					*!	*
☞ *b* pam-poʔók						
c pan-poʔók			*!		*	

To summarize the proposal, nasal substitution is motivated after certain prefixes, where the choice between realizing a floating nasal feature on a stem consonant and giving it its own segmental slot must be made. Nasal substitution is disfavored across loose morphological boundaries by the constraint MORPHEMECOHESION. It would also be possible to analyze boundary-strength effects in terms of derivational levels (Allen 1971; Siegel 1974; Kiparsky 1982, and others): we would say that a word like /paŋ+bigáj/ → [pa-migáj] has its prefix added early, feeding nasal substitution, but *pam-pulítika* has its prefix added at a later level when nasal substitution is no longer active. The two analyses seem to make the same empirical predictions here.

The grammar so far predicts that, as long as the prefix-stem boundary is tight enough, nasal substitution will occur. Clearly this is not the case, since there is variation. As a first approximation, we could say that DEP-C and *ASSOCIATE are variably ranked—notated by the jagged line—allowing both (12a) and (12b) to surface:

(16)

/p_1a_2[+nas]$_3$/+/$b_4i_5g_6a_7j_8$/	MAX(+nas)	DEP-C	*ASSOCIATE
☞ *a* pa-$m_4$igaj | [+nas]$_3$			*
☞ *b* pam$_9$-$b_4$igaj | [+nas]$_3$		*	

As the next two subsections discuss, additional constraints concerning voicing and place of articulation will cause the two variants to have different probabilities depending on the stem-initial consonant.

3.2 The voicing effect

Following Pater's (1999) analysis of nasal substitution in Indonesian, I attribute the higher rate of substitution on voiceless-initial stems to a constraint *NC̥, which forbids a sequence of a nasal and a voiceless obstruent:

(17) *NC̥: A [+nasal] segment must not be immediately followed by a [−voice, −sonorant] segment.

Hayes (1999) and Hayes and Stivers (1995) propose a phonetic motivation for *NC̥, supported by a simulation of vocal tract aerodynamics. In an *NC̥* sequence, voicing must be "on" for the nasal, but "off" for the following obstruent. Voicing can be turned off passively, by allowing supraglottal pressure to build up so that airflow across the glottis ceases, and/or actively, by abducting the vocal folds to make them

less likely to vibrate (abducting the vocal folds also increases transglottal airflow, accelerating the buildup of supraglottal air pressure and thereby further promoting voicelessness). In a nasal-to-oral transition, two factors slow the buildup of supraglottal pressure and thus inhibit passive loss of voicing. The first is what Hayes and Stivers (1995) call 'nasal leak': as the velum rises to cut off nasal airflow, it reaches the point of insufficient velar opening for a percept of nasality before the velar port is fully closed. Although the perceptually oral portion of the cluster has begun, air continues to leak out into the nasal cavity, impeding the build-up of supraglottal pressure. The second factor is 'velar pumping': the velum may, and often does, continue to rise past the point of full closure of the velar port. This expands the oral cavity, further hampering the buildup of supraglottal pressure. The *C̥* of an *NC̥* sequence is thus likely to be realized as at least partially voiced, unless extra effort (such as glottal abduction) is exerted to turn off voicing. Hayes and Stivers propose that the articulatory difficulty of *NC̥* clusters drives postnasal voicing. Pater (1999) discusses *NC̥ as the motivation for Indonesian nasal substitution (which applies only to voiceless obstruents), and for postnasal voicing, nasal deletion, and denasalization in various languages.

*NC̥ favors substitution in voiceless-initial stems. A word of the form *pan-tabój*, without substitution, violates *NC̥, but *pa-nabój*, with substitution, does not. *NC̥ is irrelevant for voiced-initial stems, since it is violated by neither non-substitution (*pan-diníg*) nor substitution (hypothetical *pa-niníg*). Thus, for voiceless-initial stems, there is an additional constraint favoring the nasal-substituted candidate.

Although *NC̥ is apparently ranked high enough to produce a voicing effect in nasal substitution, it is violated quite freely root-internally (*sampál* 'slap', *bintíʔ* 'calf', *taŋkád* 'tallness'). The faithfulness constraints that would be violated in repairing an NC̥ cluster, such as DEP(+voice), MAX-C, MAX(+nasal), and UNIFORMITY, must outrank *NC̥:

(18)

/hin$_1$t$_2$aj/	DEP(+voice)	MAX-C	MAX(+nasal)	UNIFORMITY	*NC̥
a hin$_{1,2}$aj				*!	
☞ *b* hin$_1$t$_2$aj					*
c hin$_1$d$_2$aj	*!				
d hin$_1$aj		*!	*!		
e hil$_1$t$_2$aj			*!		

3.3 The place effect

We have seen that nasal substitution is more common on "fronter" obstruents. That is, it is more common on labials than coronals, and more common on coronals than on dorsals. I propose that this is due to the markedness of the resulting stem-initial nasal: $[_{stem}ŋ$ is more marked than $[_{stem}n$, which is more marked than $[_{stem}m$. This translates into the *[NASAL constraint family given in (19).

(19) *[ŋ (*[n, *[m): A stem must not begin with *ŋ* (*n*, *m*) or a fronter nasal.

The constraints are stated as a stringency hierarchy (Prince 1997; de Lacy 2002), so that, for example, a stem-initial *n* violates both *[n and *[m.

It was argued in Section 2.1 that a nasal produced by nasal substitution is stem-initial, as diagnosed by its behavior in reduplication. The *[NASAL family of constraints therefore disfavors substitution. For example, *pa-nabój*, with substitution, violates *[n (and *[m), because the *n* that results from substitution is stem-initial. But *pan-tabój*, without substitution, does not violate *[n, because the *n* belongs to the prefix only. For example, if *[ŋ, *[n ≫ DEP-C ≫ *[m, then, all else being equal, substitution would occur on a labial-initial stem, but not on a coronal- or velar-initial stem:

(20) *[ŋ, *[n ≫ DEP-C ≫ *[m (hypothetical)

		/maŋ/+/bala/	*[ŋ	*[n	DEP-C	*[m
☞	*a*	ma-mala				*
	b	mam-bala			*!	

		/maŋ/+/dala/	*[ŋ	*[n	DEP-C	*[m
	c	ma-nala		*!		*
☞	*d*	man-dala			*	

		/maŋ/+/gala/	*[ŋ	*[n	DEP-C	*[m
	f	ma-ŋala	*!	*		*
☞	e	maŋ-gala			*	

There is some other support for the idea that backer stem-initial nasals are more marked. Looking at Tagalog roots, we find that there are few root-initial nasals in native roots, both overall and as a proportion of nasals in all positions—but, among the nasals, *m* is better represented root-initially than *n* or *ŋ*. This consonantal distribution suggests that (and would provide evidence to the learner that) root-initial nasals are disfavored, but the fronter ones less so.

Cross-linguistically, the constraint *[ŋ can be widely observed. Stem-initial *ŋ* is prohibited in English, for example, and untrained English speakers have difficulty producing initial *ŋ* even as a phonetic exercise. McCarthy and Prince (1995) propose that *[ŋ is responsible for blocking *g* → *ŋ* lenition non-postvocalically in certain dialects of Tokyo Japanese. Flack (2007) presents a typology in which *ŋ* can be banned from syllable onsets, from word onsets, or from utterance onsets. Although initial *ŋ* does occur in Tagalog, the avoidance of nasal substitution on velar obstruents, and the relative scarcity of *ŋ*-initial roots, suggests that initial *ŋ* is nevertheless disfavored.

What might be the motivation for a constraint *[ŋ, and can it be extended to a weaker prohibition on initial *n*?[12] The resonating cavity during production of the backest nasal, uvular [N] (which does not occur in Tagalog), is approximately a single tube from the glottis, through the pharynx and nasal cavity, to the nostrils; the oral

[12] I'm very grateful to Dan Silverman for suggesting and discussing this acoustic explanation; any remaining errors in it are mine, and he would probably not agree with the appeal to sonority.

cavity is blocked off by the uvular closure. This results in a vowel-like formant structure (Fujimura 1962; Johnson 1997). The resonating cavity for the frontest nasal, labial [m], is the glottis-to-nostrils tube plus an oral 'side tube' from the uvula to the closed lips. This closed side tube resonates at certain frequencies, according to its length; in the acoustic output, however, antiformants (frequencies at which amplitude is decreased) are created at these resonant frequencies of the oral side tube. Fujimura (1962) found the lowest antiformant for [m] to be between 750 Hz and 1250 Hz—low enough to interfere greatly with the vowel-like formants contributed by the main glottis-to-nostrils tube. In addition, frequencies above the antiformant are reduced in amplitude. The result is a sound that is not vowel-like. In between the extremes of [ɴ] and [m], velar [ŋ] has only a very short oral side tube, resulting in antiformants at high frequencies (the lowest is above 3000 Hz), where amplitude is already low because of the absorptive nasal cavity. Thus, the antiformants associated with the oral side tube in [ŋ] cause little interference with the vowel-like formant structure contributed by the main tube—the result is a sound that is much more vowel-like than *m*. Alveolar [n]'s antiformants are lower than [ŋ]'s (the lowest is between 1450 Hz and 2200 Hz), but still higher than [m]'s, so [n] is somewhere in between in vowel-like-ness. See Narayan 2006 on the acoustics of Tagalog/Pilipino nasals in particular.

If syllable-onset consonants are preferably of low sonority (e.g., Dell and Elmedlaoui 1985; Clements 1990; Prince and Smolensky 1993/2004—and see de Lacy 2001 and Flack 2007 for word-initial onsets in particular), and sonority is correlated with loudness and/or vowel-like-ness, then the backer a nasal consonant is, the more sonorant it is, and therefore the less suited to onset position. If these acoustic properties are the motivation for *[ŋ, then they are shared, to a lesser degree, by onset *n*, so we have motivation for the harmony scale [*m* ≻ [*n* ≻ [ŋ ([*m* is more harmonic than [*n*, which is more harmonic than [*ŋ*).

To summarize, in addition to the *NC̥ constraint that favors the nasal-substituted constraint for voiceless stem-initial consonants, there are also constraints disfavoring the nasal-substituted candidate for velars as compared to coronals, and for coronals as compared to labials. Excluding words that cannot undergo nasal substitution because of a strong # boundary, various rankings of these constraints would produce corresponding invariant patterns of nasal substitution in the remaining words: for example, we have seen, in (20), how to produce substitution on labials only. But, what we want is variation, with nasal substitution more likely for some consonants than for others. After some alternatives to the constraints just proposed are considered in Section 3.4, Section 4 describes how these conflicting constraints interact with each other and with lexical specifications to produce the desired pattern of variation.

3.4 Alternative analyses

The floating-feature account of what drives nasal substitution has at least two viable alternatives. For example, it would be reasonable to make nasal substitution a stipulative part of the morphology. This could be implemented as a FIATSTRUC constraint (MacBride 2004) requiring certain morphosyntactic features to be realized phonologically as nasalization of a stem-initial consonant. Or a more general

and more stipulative constraint could simply require a prefix-final nasal to fuse with the following consonant. We could also adopt Kaufman's (2005) proposal that the prefix nasal underlyingly has a timing slot, but this skeletal position is lost, leaving the [+nasal] feature in need of a new home. The rest of the paper will use the floating-feature analysis introduced above, but these other approaches could work.

There are also some alternatives that seem not to work for Tagalog: *NC̥, CRISPEDGE, and *CC. Starting with *NC̥, unlike in Malay/Indonesian (Pater 1999), it clearly cannot be the driving constraint for Tagalog nasal substitution, because it applies to both voiced and voiceless obstruents.

In Pater's 2001 reanalysis of Indonesian, the driving constraint is one that requires a "crisp edge" at prefix-root boundaries (because they are prosodic-word boundaries). This is violated by a non-substituted candidate such as hypothetical *pan-tinig*, because the prefix-final nasal and the root-initial obstruent share place features; a candidate where they fail to share place features, such as **paŋ-tinig*, is ruled out by a language-wide requirement of nasal place assimilation. The crisp-edge constraint is satisfied if nasal substitution occurs, on the assumption (shared here) that the resulting segment belongs only to the stem: *pa-ninig*. The crisp-edge idea is supported by other phenomena in Indonesian (Cohn and McCarthy 1994/1998), and attractively explains why, in Indonesian, substitution occurs only at prefix-root boundaries (not root-internally, and not at prefix-prefix boundaries).

Can a crisp-edge approach work for Tagalog? Nasal substitution can occur at prefix-prefix boundaries in Tagalog: [ʔíbig] 'love', [**k**a-ʔibíg-an] 'friend', [ma-**ŋ**a-ʔibíg-an] 'make friends' (/maŋ-ka-ʔibíg-an/), [pa-**ŋ**a-**ŋ**a-ʔibíg-an] 'friendly attitude' (/paŋ-RED-ka-ʔibíg-an/). But this could merely reflect differences between Tagalog and Indonesian in prosodic-word structure; or, it could be that in cases like this the inner prefix (here, *ka-*) has been lexicalized as part of the root. So data like these don't pose a grave threat to a crisp-edge analysis.

More problematic is that the crisp-edge approach treats nasal substitution as a symptom of a prefix-stem boundary, which should predict that if a word's prefixed status is semantically opaque, the phonology is less likely to treat the word as prefixed. See, for example, Baroni's (2001) findings for *s*-voicing in prefixed words in Italian. In other words, semantic opacity should therefore suppress nasal substitution: the more the prefix is treated as "melded" with the root, the less substitution should apply. But as noted above, the opposite seems to be true: nasal substitution represents a tight integration, not a sharp separation, of the prefix and stem.

Finally, Archangeli et al. (1998) propose that nasal substitution is driven by simple avoidance of consonant clusters, *CC. The Tagalog prefix inventory is phonologically quite restricted, with productive prefixes ending in *ŋ*, *g*, or a vowel, so the only other relevant case is the *g*-final prefixes, principally *mag-*, *nag-*, and *pag-*. These prefixes also produce consonant clusters (and, with velar-initial stems, possibly non-crisp edges: *mag-kilatís-an* 'to appraise each other' from *kilátes* ∼ *kilátis* 'carat'). But they do not induce anything like deletion or coalescence, even though the faithfulness violations would be no worse than those incurred in nasal substitution. A reviewer points out that *mag-*, *nag-*, and *pag-* (and also *tig-*) could simply be exceptions to a fusion

requirement. With only two consonants (*ŋ* and *g*), it seems impossible to determine which one's behavior should be viewed as regular and which as exceptional.

Two alternative explanations for the place effect, related to voicing, are unpromising. First, among voiceless obstruents, the place effect could be seen as a fine-tuned version of *N̥C̥. Recall that the phonetic motivation proposed by Hayes and Stivers (1995) for *N̥C̥ is that the expansion of the oral cavity during velum raising encourages voicing. Their model also found that frontness of the obstruent encourages voicing, because there is a greater expanse of flexible cheek wall that can expand outward and reduce supraglottal pressure. This would explain why *p* substitutes slightly more often than *k*. But it does not explain the larger effect in which *b* substitutes more often than *d*, since turning off voicing is not necessary in *mb*, *nd*, and *ŋg* clusters—indeed, the frontness of *b* would make voicing easier to maintain,[13] and thus the cluster *mb* would be less marked (and so less subject to repair by coalescence) along this dimension than *nd* or *ŋg*. A second possibility, extrapolating from Pater (2001), is that faithfulness violations are greater when substituting a backer consonant. Pater proposes that the reason voiced obstruents do not substitute in Indonesian is that if they did, IDENT-IO(pharyngeal expansion) would be violated: voiced obstruents are [+pharyngeal expansion]—they require active expansion of the pharynx, or some other exertion, to maintain voicing—but nasals are [–pharyngeal expansion], because voicing is maintained by venting air out the nose. Fronter consonants should require less pharyngeal expansion, because more cheek area is available for passive expansion, and so nasalizing a *b* is less of a violation of (some gradient version of) IDENT-IO(pharyngeal expansion) than nasalizing *g*. The (admittedly small) place effect among voiceless consonants is then a puzzle, though, because voiceless consonants require no pharyngeal expansion; there is no violation of Ident-IO(pharyngeal expansion) when nasalizing a voiceless obstruent.

Another explanation for the place asymmetry is offered by Kaufman (2005): nasal substitution among bilabials collapses a two-way contrast between *p* and *b*. Among coronals it threatens a three-way contrast (*t*, *s*, *d*), and among velars and glottal stop it threatens a three-way contrast (*k*, *g*, and *ʔ*). Kaufman proposes an analysis based on prohibition of neutralization that captures the *b* vs. *d*, *g* place effect in Tagalog—two-way but not three-way neutralization permitting *b* to substitute but not *d* and *g*—as well as some interesting effects in related languages (Kapampangan, Mori, Mukah Melanau, and Tombonuwo). The proposal does not, however, address the difference between *d* and *g* in Tagalog.

4 The model and how it applies

Section 2 illustrated phonological trends within Tagalog nasal substitution, and Section 3 presented constraints that can produce absolute versions of the trends seen. This section gives evidence that many words' pronunciations are lexically determined, and

[13] Ohala and Riordan (1979) found that passive cavity expansion maintained voicing longer for *b* than for *d* or *g*.

then addresses the question of how the lexicon and grammar interact to produce lexical variation.

4.1 Lexical idiosyncrasy

There are several ways in which words that take *paŋ-* and *maŋ-* prefixes can be idiosyncratic. First, despite the lexical trends described in Section 2, it is not completely predictable which words will undergo substitution—the combination of one of these prefixes with a stem that begins with *b*, the most variable initial consonant, has a good chance of displaying and a good chance of not displaying nasal substitution. Substitution is not even consistent among derivatives of the same stem, as illustrated in (21), so it does not suffice to have a single diacritic for each stem. Walther and Wiese (1999) have suggested that each stem and each prefix could be given a diacritic: when a prefix and stem are combined, if their diacritics are both [+substitute], substitution occurs; if both diacritics are [−substitute], or if the diacritics disagree, no substitution occurs. But this approach too would require listing of exceptions. For example, if reservative-adjective-forming *paŋ-* is [−substitute], then all cases where substitution does occur with *paŋ-* must be listed as exceptions—as seen in Fig. 7, there are 3 such cases. And if *paŋ-* is [+substitute], then cases where substitution fails to apply to a stem that otherwise does undergo substitution (e.g., *pam-búhaj* in (21)) must be listed as exceptions. Most cases of mixed behavior within a stem do involve the prefix *paŋ-*, so it might be possible to apply a simple diacritic approach for other constructions, but at the very least, behavior with *paŋ-* would have to be separately specified for many stems.

(21)

prefixes	*nas. sub.?*	*(freq. no ; freq. yes)*		
			búhaj	‘life’
paŋ-	no	(10 ; 0)	pam-**b**úhaj	‘vivifying’
maŋ-	yes	(0 ; 652)	ma-**m**úhaj	‘to live’
paŋ-RED-	yes	(1; 1975)	pa-**m**u-**m**úhaj	‘manner of living’
			batás	‘law’
paŋ-	no	(30 ; 0)	pam-**b**atás	‘legal’
paŋ- -an	yes	(1 ; 47)	pa-**m**àtás-an	‘legislative’
maŋ-RED-	no	(766 ; 0)	mam-**b**a-**b**atás, mam-**b**a-**b**átas	‘legislator’

The second type of idiosyncrasy is semantic. Although the semantic connection between stem and derivative is typically apparent, as seen in (22) exact meanings can be unpredictable. This is especially true of verb-forming *maŋ-*. Semantic idiosyncrasy is found in both substituted and non-substituted words. This means that the lexicon must additionally, at least in some cases, specify the meaning of a potentially nasal-substituted word.

(22)

ʔabáŋ	'watcher'	ma-ŋabáŋ	'to wait near people who are eating, hoping to get food'
babáʔe	'woman'	mam-babáʔe	'to have a mistress'
siʔíl	'oppressed by ruler'	ma-niʔíl	'to strangle to death'
ʔibábaw	'surface'	paŋ-ʔibábaw	'veneer'
kíta	'visible'	pà-ŋitáʔ-in, pà-ŋitaʔ-ín	'apparition, omen'
túbig	'water'	ma-nubíg	'to urinate'
balík	'return'	pa-malík	'hand rudder'
gánt͡ʃo	'hook'	maŋ-ga-gánt͡ʃo	'con man'

Third, certain affixes can cause unpredictable stress/length shifts. This idiosyncrasy too occurs in both substituted and non-substituted words (though more often for substituted). This adds a third piece of information that must be lexically specified for at least some potentially nasal-substituted words. (Because stress is not indicated in regular orthography, stress data here come solely from the dictionary, English 1986, though all items are attested in the corpus.)

(23)

		maŋ-RED- with stress shift	
tahíʔ	'sewing'	mà-na-n**á**hiʔ	'seamstress'
		maŋ-RED- without stress shift	
pun**á**	'remark'	mà-mu-mun**á**	'critic'
ʔ**á**wit	'song'	maŋ-ʔa-ʔ**á**wit	'singer'
		maŋ- with stress shift	
t**ú**big	'water'	ma-nub**í**g	'to urinate'
		maŋ- without stress shift	
k**í**kil	'carpenter's file'	ma-ŋ**í**kil	'to chisel; to ask for money'
		paŋ- with stress shift	
s**í**pit	'claws'	pan-sip**í**t	'(type of) rat-trap'
		paŋ- without stress shift	
t**ú**koj	'mention'	pan-t**ú**koj	'article [grammar]'

For many words with nasal-substituting affixes, then, a speaker must know and lexically encode a number of facts not predictable from other words containing the same stem. There are various ways this could be done, including elaborate systems of diacritics. A simple method, which will be adopted here, is to allow at least some morphologically complex words their own lexical entries, such as /mamigáj/ 'to distribute', which is formed from /bigáj/ by prefixation with /maŋ-/, or perhaps /ma-migáj/, with the morpheme boundary encoded in the lexical entry.[14] This lexical entry would contain any unpredictable information about meaning and stress. Most important for our purposes, the lexical entry determines whether the word is nasal-substituted with respect to its stem. Even if some form of output-output (O-O) faithfulness to related

[14] I leave open whether these lexical entries encode the word's morphological complexity, and, if so, if they do it through morphological boundaries, morphological bracketing, or merely relationships to related words with the stem or affix(es).

mag-bigáj ‘to give’ favors “undoing” nasal substitution in *ma-migáj*, higher-ranked input-output faithfulness will not allow it:

(24)

		/mamigáj/ related to [mag-bigáj]	I-O FAITH	O-O FAITH
☞	*a*	mamigáj		*
	b	mambigáj	*!	

It may be that speakers lexicalize all potentially nasal-substituted words (as long as they are frequent enough), or that speakers lexicalize only words that buck the trends. For example, we could imagine that speakers lexicalize only words that do undergo nasal substitution (or have other idiosyncrasies), or that they lexicalize only words that belong to the minority pattern for their construction, or to the minority pattern for their construction and initial consonant. What is important is that any word that is in danger of being assigned an incorrect pronunciation by the grammar be protected by lexical information.[15]

There is potentially a three-way distinction to be made: there are (i) words that are lexicalized as undergoing nasal substitution; (ii) words lexicalized as not undergoing nasal substitution; and (iii) words not yet lexicalized. Examples of words plausibly of types (i) and (ii) are *mà-ma-mahála?* ‘responsibility’ (< *bahála?* ‘manager’) and *mam-ba-básа* ‘reader’ (< *bása* ‘reading’) and. Both have *b*-initial stems and belong to the *maŋ-RED-* construction; both are frequent (725 for ⟨mambabasa⟩ and 81 for ⟨mamamahala⟩) and consistent in their behavior (corpus frequencies of 0 for *⟨mamamasa⟩ and *⟨mambabahala⟩). Likely examples in the corpus of type (iii) words include presumably nonce or fairly recent coinages based on English loans such as those in (25), sampled from the *maŋ-* construction.

(25)

spelled form	*frequency*	*presumed English source*
mangkonjugation	1	conjugation
mam-bird-repel	1	bird-repel
mamblog	11	blog
man-takeover	1	takeover
ma-mrospect	1	prospect
mangareer	5	career (⟨ng⟩ = [ŋ])

If a word is lexicalized without nasal substitution, various input-output faithfulness constraints prevent substitution—just as it is prevented within a root—even if some markedness constraint such as NOCODA (or, if relevant *NC̥) favors substitution. Illustrated here is just one of the McCarthy-Prince faithfulness constraints that would

[15] Why do some words have these idiosyncratic properties? The answers are presumably different from the diachronic point of view and from the point of view of the learner who must replicate the ambient pronunciations and meanings. Hay (2003) discusses a two-way relationship between lexical representation/access and idiosyncrasy: lexical idiosyncrasy can cause learners to treat the affected words more as wholes than as morphologically composed; conversely, if some other factor causes learners to treat certain words as wholes, then those words have a better chance of developing idiosyncrasies over time.

be violated, UNIFORMITY-IO, which forbids two distinct underlying segments from corresponding to a single surface segment (McCarthy and Prince 1995).

(26)

		/mam$_1$b$_2$abása/	MAX-IO	UNIFORMITY-IO	NOCODA
☞	*a*	mam$_1$b$_2$abása			*
	b	mam$_1$amása	*!		
	c	mam$_{1,2}$amása		*!	

For a word whose behavior is not yet established, any faithfulness constraint violated by nasal substitution, such as *ASSOCIATE, should be variably ranked with respect to DEP-C, as indicated by the jagged line, on the assumption that both non-substituted and substituted outcomes are possible at this stage in the word's life:

(27)

		ma[+nas]$_1$ + /b$_2$lóg/	*ASSOCIATE	DEP-C
☞	*a*	ma-m$_2$lóg | [+nas]$_1$	*	
☞	*b*	mam$_8$-b$_2$lóg | [+nas]$_1$		*

Faithfulness to a lexical entry such as /mambabása/ can't be enforced, however, if that lexical entry is not used. For example, if the input to (26) were *maŋ+RED+básа* instead, nasal substitution might occur, because of the variability in ranking between *ASSOCIATE and DEP-C. To prevent this, there must be some preference for using the input *mambabása* if it can be accessed, rather than synthesizing *maŋ+RED+básа*. The idea that lexicalized whole words are preferred over novel syntheses has a long history, as in Aronoff's (1976) blocking principle, or Kiparsky's (1982) Elsewhere Condition, with listed words viewed as specific rules that pre-empt more general ones. It is beyond the scope of this work to decide whether this preference should be enforced by a constraint in the phonological grammar (such as Zuraw's 2000 USELISTED) or by a morphological principle or processing effect that determines the input to the phonological grammar in the first place. I will simply assume in the tableaux below that the lexical entry is used if available.[16]

The idea that some words' behavior is determined by their lexical entries and others' is free to be determined by the grammar has a precedent in Inkelas et al. 1997, where the distinction is between fully specified segments and segments with underspecified features, for which Ident is irrelevant. There is also similarity to the proposal of Becker (2009) that known words can be indexed to a particular faithfulness constraint, whose ranking determines their behavior (see Pater 2006), but novel words must be assigned an indexation.

[16]The leap during word-learning from unknown word to fully available lexical entry is likely not instantaneous, and presumably the likelihood that a lexical entry will be available for use on a given occasion is a function of the strength/activation of that entry in the speaker's lexicon.

4.2 Faithfulness *vs*. the subterranean grammar

Because of the high ranking of input-output faithfulness constraints that preserve lexical information, the constraints proposed in Section 3 come into play only when a full lexical entry is not available, and a word is synthesized from a prefix plus a stem. These "subterranean" constraints must be variably ranked, because a fixed ranked such as that in (28) would incorrectly require, in synthesized words, that nasal substitution apply to all voiceless obstruents (p, t, s, k) and to b, but not to d and g.

(28) *NC̥ ≫ *[ŋ ≫ *[n ≫ MAX(+nas) ≫ *[m ≫ *ASSOCIATE

To encode the voicing effect gradiently, *ASSOCIATE (which forbids nasal substitution) must be variably ranked with respect to *NC̥ (which is prefers nasal substitution on a voiceless consonant) so that substitution is optional for voiceless obstruents. To encode the place effect gradiently, DEP-C—which is violated by non-substitution—must be variably ranked with respect to *ASSOCIATE and the *[N constraints, so that substitution is possible though decreasingly likely with backer place, for all voiced obstruents.

In order to encode all these preferences, I adopt stochastic constraint ranking, in the sense of Boersma (1997, 1998) and Boersma and Hayes (2001).[17] In this framework, constraints are assigned "ranking values" on a continuous scale. Any time the grammar is used, these ranking values are randomly perturbed; the resulting perturbed values are used to derive a linear ranking of the constraints, and evaluation proceeds in the usual way. Perturbation of ranking values is achieved by adding to them a Gaussian random variable with mean of 0, so that each constraint is associated with a bell-curve probability density function centered on its ranking value. The more two constraints' tails overlap, the more variably ranked they are. If the ranking values of two constraints C_1 and C_2 are exactly the same, the curves overlap exactly, and the frequency with which C_1 outranks C_2 in evaluation is 50%.

Boersma's (1997, 1998) Gradual Learning Algorithm was used to simulate the process of learning a (fragment of a) grammar from the real Tagalog lexicon, and then applying that grammar to newly coined words.

The algorithm was run using Hayes et al.'s (2005) OTSoft. The training data given to the learner are as summarized in (29). The training items are treated as whole,

[17] In a system such as Anttila's (1997 and elsewhere), some constraint rankings are specified as obligatory, and others are left unspecified. The rate of substitution for a given obstruent depends on the number of linear rankings consistent with the specified rankings that produce substitution, compared to the number that do not. If the 6 constraints DEP-C, *NC̥, *ASSOCIATE, *[ŋ, *[n, and *[m are all freely ranked, then there are $6! = 720$ possible rankings. Of those, 50% produce substitution on p (those where, out of DEP-C, *NC̥, *ASSOCIATE, and *[m, DEP-C or *NC̥ is ranked topmost), 40% produce substitution on t and s (those where DEP-C or *NC̥ is ranked topmost out of DEP-C, *NC̥, *ASSOCIATE, *[n, and *[m), 33% produce substitution on k (those where DEP-C or *NC̥ is ranked topmost), 33% produce substitution on b (those where DEP-C is ranked topmost out of DEP-C, *ASSOCIATE, and *[m), 25% produce substitution on d (those where DEP-C is ranked topmost out of DEP-C, *ASSOCIATE, *[n and *[m, and 20% produce substitution on g (those where DEP-C is ranked topmost out of constraints DEP-C, *ASSOCIATE, *[ŋ, *[n, and *[m). This approach would capture the voicing and place effects, but because there doesn't exist a learning algorithm for crucially non-linear constraint rankings, it wouldn't be possible to carry out the learning simulations here.

listed words that contain *{m/p/n}aŋ-* and a stem. Each item has a faithful winning candidate and an unfaithful losing candidate.

Some of the words are underlyingly non-substituted, so fusion of the nasal and obstruent in the losing, unfaithful candidate results in a UNIFORMITY-IO violation. There is also a resulting FAITH-OO violation, because the surface form will differ from other words with same stem, such as the bare stem itself. I assume that in most cases the derivational base to which the candidates are being compared by FAITH-OO is the bare stem, such as *poʔók* for *pam-poʔók* (3), although a further research would be needed to establish this. None of the constraints in the simulations favor an unfaithful realization of the stem-initial consonant in /poʔók/, so it's reasonable to exclude tableaux for /poʔók/ and other bases from the learning data—that is, the output form being used for comparison can be regarded as fixed, regardless of the ranking. Some nasal-substituted words lack bare-stem bases, and there we might run into a learning problem if the base has a *g*-final prefix: in /mag-bigáj/, for example, if NOCODA ≫ MAX-C the candidate *[mag-igáj] might win if something prefers it over *[ma-bigáj]—and thus the evaluation of FAITH-OO would change as the ranking values change during learning. The simulation here ignores cases of this type. I assume that, in the real Tagalog learning situation, these cases would be few enough that learners could establish the high ranking of MAX-C early on and potentially delay evaluating O-O faithfulness between pairs like [mag-bigáj] and [ma-migáj] until that time. In a fuller simulation of the Tagalog lexicon and grammar, there would presumably be a FAITH-OO constraint specific to each morphological construction, such that some constructions are more permissive of nasal substitution than others.

Other words are underlyingly substituted—in comparison to other words formed from the same stem, that is. The unfaithful output candidate for a substituted input "undoes" nasal substitution by splitting the underlying nasal into a nasal and an obstruent, violating INTEGRITY-IO (McCarthy and Prince 1995). The frequencies were taken from the lexical data in Fig. 1. For example, 10 words were given to the learner with non-substituted /p/ and 253 words were given with substituted /p/.

Because in these training data the input is always a full, listed word, rather than a prefix-stem concatenation, there are no floating [+nasal] features, and thus no violations of *ASSOCIATE[18] (violated when the prefix [+nas] associates to the stem segment), DEP-C-IO (violated when a segment is added to the prefix to support the [+nas] feature), or MAX(+nas).

[18]Ensuring that these forms don't violate *ASSOCIATE requires a detailed consideration of exactly what "do not insert new association lines" (Yip 2002:79) means—I assume that the change from

$$\begin{array}{ccc} /\ldots m_1 b_2 \ldots/ & & [\ldots m_{1,2} \ldots] \\ \quad | & \text{to} & \quad | \\ [+\text{nas}]_3 & & [+\text{nas}]_3 \end{array}$$

does not involve insertion of a new association line, since items 1 and 3 were already associated in the input, even though 2 and 3 were not.

(29)

training items				*number of constraint violations*									
# of items	input		output	INTEGRITY-IO	UNIFORMITY-IO	*NC̥	*[ŋ	*[n	*[m	FAITH-OO	NOCODA	*ASSOCIATE	DEP-C-IO
10	/Cam$_1$p$_2$V…/	☞	[…mp…]			1					1		
			[…m$_{1,2}$…]		1				1	1			
253	/Cam$_1$V…/		[…mp…]	1		1					1		
		☞	[…m$_{1,2}$…]						1	1			
26	/Can$_1$t$_2$V…/	☞	[…nt…]			1					1		
			[…n$_{1,2}$…]		1			1	1	1			
430	/Can$_1$V…/		[…nt…]	1		1					1		
		☞	[…n$_{1,2}$…]					1	1	1			
17	/Caŋ$_1$k$_2$V…/	☞	[…ŋk…]			1					1		
			[…ŋ$_{1,2}$…]		1		1	1	1	1			
285	/Caŋ$_1$V…/		[…ŋk…]	1		1					1		
		☞	[…ŋ$_{1,2}$…]				1	1	1	1			
100	/Cam$_1$b$_2$V…/	☞	[…mb…]								1		
			[…m$_{1,2}$…]		1				1	1			
177	/Cam$_1$V…/		[…mb…]	1							1		
		☞	[…m$_{1,2}$…]						1	1			
70	/Can$_1$d$_2$V…/	☞	[…nd…]								1		
			[…n$_{1,2}$…]		1			1	1	1			
25	/Can$_1$V…/		[…nd…]	1							1		
		☞	[…n$_{1,2}$…]					1	1	1			
97	/Caŋ$_1$g$_2$V…/	☞	[…ŋg…]								1		
			[…ŋ$_{1,2}$…]		1		1	1	1	1			
1	/Caŋ$_1$V…/		[…ŋg…]	1							1		
		☞	[…ŋ$_{1,2}$…]				1	1	1	1			

The algorithm was run on the training data for 5,000,000 cycles, with an initial plasticity of 0.02 and a final plasticity of 0.002. The resulting ranking values are given in Table 3.

The first thing to note about Table 3 is that the two faithfulness constraints, INTEGRITY-IO and UNIFORMITY-IO, are ranked well above any other constraint. The probability UNIFORMITY-IO's being outranked by the next-highest constraint,

Table 3 Ranking values returned by learner

112.213	INTEGRITY-IO
112.176	UNIFORMITY-IO
102.799	*NC̥
100.538	*[ŋ
100.251	*[n
100.038	NOCODA
100.000	*ASSOCIATE
100.000	DEP-C-IO
99.962	FAITH-OO
99.962	*[m

*NC̊, is vanishingly low. Thus, as desired, and as in the learning data, lexical information is preserved faithfully: If a word is lexicalized as not undergoing nasal substitution, it will not undergo it; and if a word is lexicalized with nasal substitution, substitution will not be undone in order to satisfy a lower-ranked constraint.

However, before establishing the top ranking of these faithfulness constraints—and thus ceasing to make the errors that drive this learning algorithm—the learner acquired a ranking of the lower-ranked constraints in accordance with the lexical statistics fed into it. For example, because there were 253 learning items of the type $/Cam_1V\ldots/ \rightarrow [\ldots m_{1,2}\ldots]$, and only 10 of the type $/Cam_1p_2V\ldots/ \rightarrow [\ldots mp\ldots]$, the learner promoted *NC̊ and demoted *[m far more often than it did the reverse, and *NC̊ ended up with a higher ranking value than *[m.

If a word has not been lexicalized, however, these top two faithfulness constraints are irrelevant. As discussed in Section 3, the prefix is assumed to have a floating [+nasal] feature; if it docks to the stem consonant (substitution), there is a violation of *ASSOCIATE, and if it docks instead to an inserted prefix consonant (non-substitution), there is a violation of DEP-C-IO. There is no question of fusing or splitting two segmental root nodes, and thus no violation of INTEGRITY-IO or UNIFORMITY-IO. It's up to the lower-ranked markedness constraints to decide.

The lower-ranked constraints are close together in their ranking values, so the outcomes will vary: a novel prefix+stem combination might be produced with substitution on one occasion, and without substitution on another. But there is a preference for the higher-ranked constraints to be obeyed at the cost of the lower-ranked constraints. To test the rates of substitution predicted, the grammar in Table 3 was applied to the items shown in (30). These items represent newly coined words, concatenations of prefix and stem.

(30)

testing items		*number of constraint violations*									
input	output	INTEGRITY-IO	UNIFORMITY-IO	*NC̊	*[ŋ	*[n	*[m	FAITH-OO	NOCODA	*ASSOCIATE	DEP-C-IO
/Ca[+nas]/ + /pV…/	[…mp…]			1					1		1
	[…$m_{1,2}$…]						1	1		1	
/Ca[+nas]/ + /tV…/	[…nt…]			1					1		1
	[…$n_{1,2}$…]					1	1	1		1	
/Ca[+nas]/ + /kV…/	[…ŋk…]			1					1		1
	[…$ŋ_{1,2}$…]				1	1	1	1		1	
/Ca[+nas]/ + /bV…/	[…mb…]								1		1
	[…$m_{1,2}$…]						1	1		1	
/Ca[+nas]/ + /dV…/	[…nd…]								1		1
	[…$n_{1,2}$…]					1	1	1		1	
/Ca[+nas]/ + /gV…/	[…ŋg…]								1		1
	[…$ŋ_{1,2}$…]				1	1	1	1		1	

Each item was tested 10,000 times to determine the rate at which, for each input, the substituted output candidate wins, and the rates of substitution obtained are shown in Fig. 10.

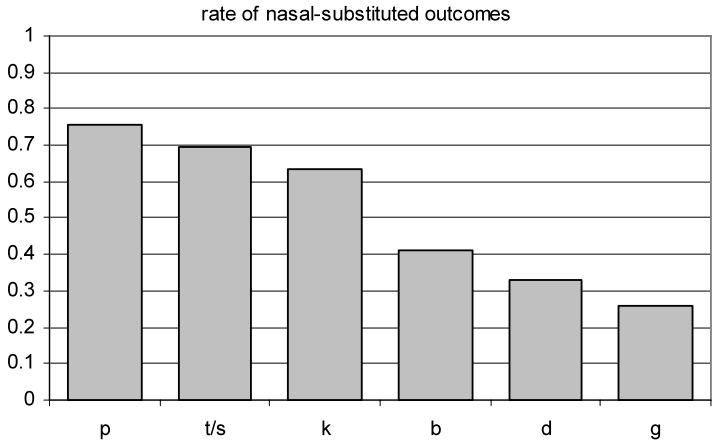


Fig. 10 Substitution rates for novel items produced by the grammar of Table 3

There is a certain smoothing here in comparison to the input learning data (compare to Fig. 1). In particular, substitution for *g*-initial stems is not as infrequent as in the learning data, presumably because *[ŋ, which penalizes substitution on *g*, has its ranking depressed by the prevalence of substitution on *k*—and, in accordance with the lexical statistics, there were more training items given for *k* than for *g*.[19] Still, to a great extent, the learning algorithm has acquired the lexical pattern and projects it onto new items, despite the fact that all the training data simply displayed faithfulness to the underlying form.

The above simulation is uninformative as to the necessity of learning, however, because the proposed constraint set will favor the voicing and place effects even if no learning occurs. If, as been argued above, the voicing and place effects are driven by constraints with a phonetic basis, and if the constraint set available to the learner does not include counter-phonetic opposing constraints, then even without any learning the voicing and place effects will tend to emerge. Because of the apparent phonetic naturalness of the pattern in the Tagalog case, it is difficult to test directly whether learning is needed or not. With a grammar like that in Table 3, but with all constraints ranked equally, the result is as in Fig. 11: the place and voicing effects are still there, though the voicing effect is muted as compared to Fig. 10.

We can, however, test whether learning will succeed when the constraint set is not phonetically biased. A second simulation was conducted with a more agnostic constraint set, as shown in (31). The difference is in the markedness constraints. First, *NC̥, which penalizes a sequence of nasal followed by voiceless obstruent, is accompanied by a *NC̬, which penalizes a nasal followed by a voiced obstruent.[20]

[19]Even with a specific constraint for each consonant (*mp, *nt, etc.) instead of the four markedness constraints used here, and even with the number of training items for each consonant made constant, the result is still quite smoothed: the *ŋg constraint, which is the lowest-ranked constraint in the resulting grammar, simply can't get a very low ranking before UNIFORMITY and INTEGRITY climb so high that errors are no longer made and learning stops.

[20]For a debate about the possibility of *NC̬ in Tswana, see Coetzee 2000; Hyman 2001; Zsiga et al. 2006; Coetzee et al. 2007.

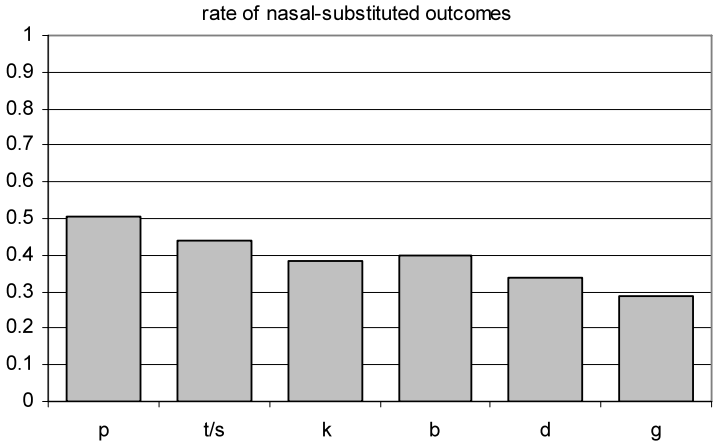


Fig. 11 Applying uniform rankings values of the constraints in Table 3 to novel items

And second, the *[N constraints, rather than reflecting a markedness scale, are each atomic—for example, *[n penalizes only stem-initial [n], not both [n] and the less-bad [m].

(31)

testing items				*number of constraint violations*										
# of items	input		output	INTEGRITY-IO	UNIFORMITY-IO	*NC̥	*NC̬	*[ŋ_atomic	*[n_atomic	*[m_atomic	FAITH-OO	NOCODA	*ASSOCIATE	DEP-C-IO
10	/Cam$_1$p$_2$V…/	☞	[…mp…]			1						1		
			[…m$_{1,2}$…]		1					1	1			
253	/Cam$_1$V…/		[…mp…]	1		1						1		
		☞	[…m$_{1,2}$…]							1	1			
26	/Can$_1$t$_2$V…/	☞	[…nt…]			1						1		
			[…n$_{1,2}$…]		1				1		1			
430	/Can$_1$V…/		[…nt…]	1		1						1		
		☞	[…n$_{1,2}$…]						1		1			
17	/Caŋ$_1$k$_2$V…/	☞	[…ŋk…]			1						1		
			[…ŋ$_{1,2}$…]		1			1			1			
285	/Caŋ$_1$V…/		[…ŋk…]	1		1						1		
		☞	[…ŋ$_{1,2}$…]					1			1			
100	/Cam$_1$b$_2$V…/	☞	[…mb…]				1					1		
			[…m$_{1,2}$…]		1					1	1			
177	/Cam$_1$V…/		[…mb…]	1			1					1		
		☞	[…m$_{1,2}$…]							1	1			
70	/Can$_1$d$_2$V…/	☞	[…nd…]				1					1		
			[…n$_{1,2}$…]		1				1		1			
25	/Can$_1$V…/		[…nd…]	1			1					1		
		☞	[…n$_{1,2}$…]						1		1			
97	/Caŋ$_1$g$_2$V…/	☞	[…ŋg…]				1					1		
			[…ŋ$_{1,2}$…]		1			1			1			
1	/Caŋ$_1$V…/		[…ŋg…]	1			1					1		
		☞	[…ŋ$_{1,2}$…]		1		1	1						

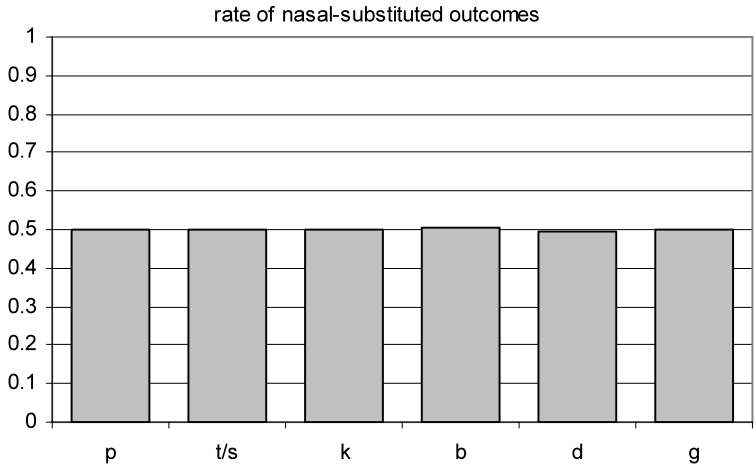


Fig. 12 Substitution rates under an impartial constraint set with no learning

We can see that this constraint set is neutral by first testing a grammar composed of these constraints, but all having the same ranking value, as though no learning had occurred. When such a grammar is tested on the novel items in (32), the result is the flat substitution rates shown in Fig. 12. There are no differences, beyond noise, among the different obstruents.

(32)

testing items		*number of constraint violations*										
input	output	INTEGRITY-IO	UNIFORMITY-IO	*NC̥	*NC̬	*[ŋ	*[n	*[m	FAITH-OO	NOCODA	*ASSOCIATE	DEP-C-IO
/Ca[+nas]/ + /pV…/	[…mp…]			1						1		1
	[…$m_{1,2}$…]							1	1		1	
/Ca[+nas]/ + /tV…/	[…nt…]			1						1		1
	[…$n_{1,2}$…]						1		1		1	
/Ca[+nas]/ + /kV…/	[…ŋk…]			1						1		1
	[…$ŋ_{1,2}$…]					1			1		1	
/Ca[+nas]/ + /bV…/	[…mb…]				1					1		1
	[…$m_{1,2}$…]							1	1		1	
/Ca[+nas]/ + /dV…/	[…nd…]				1					1		1
	[…$n_{1,2}$…]						1		1		1	
/Ca[+nas]/ + /gV…/	[…ŋg…]				1					1		1
	[…$ŋ_{1,2}$…]					1			1		1	

On the other hand, if the learning data in (31) are subjected to the same regime as in the first simulation, the grammar in Table 4 results. As before, the faithfulness constraints are top-ranked, so listed items are faithfully reproduced, and the ranking of the other constraints is similar to what it was in the first simulation. The three *[N constraints are in the correct order—but with bigger differences in their ranking

Table 4 Ranking values with learning

112.239	UNIFORMITY-IO
112.115	INTEGRITY-IO
102.710	*NC̥
100.875	*[ŋ_atomic
100.161	*[n_atomic
100.124	FAITH-OO
100.000	*ASSOCIATE
100.000	DEP-C-IO
99.876	NOCODA
99.088	*[m_atomic
97.166	*NC̯

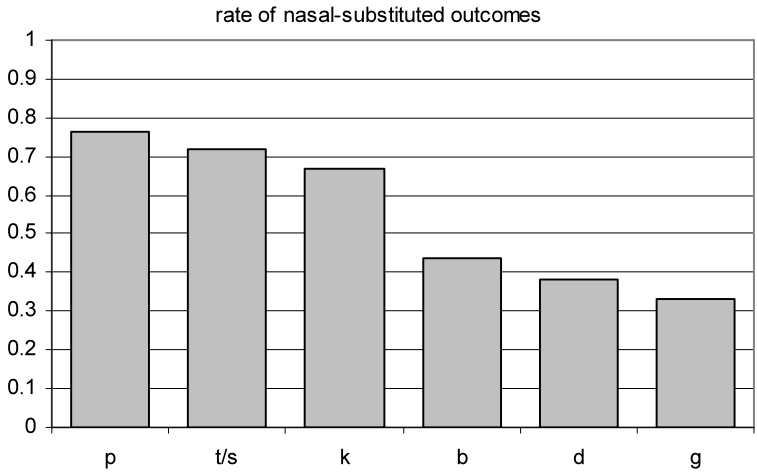


Fig. 13 Substitution rates under an impartial constraint set with learning

values, since there is no implicational relationship here to enhance the effects of *[ŋ and *[n—and the constraint *NC̯ is ranked far below its opposing counterpart *NC̥.

When this grammar is applied to the test data, the result is as shown in Fig. 13, with the voicing and place effects intact.

We can conclude from this section that even if the constraint set is unbiased, and even if every input learning datum reflects perfect input-output faithfulness, the resulting grammar will still reflect the rates at which different markedness constraints are violated by existing lexical items—as long as the learning algorithm is one that, like the Gradual Learning Algorithm, ranks constraints in a way that is sensitive to *how often* they are violated by winning vs. losing candidates. Such a learning process produces a grammar that is faithful to listed items' lexicalized behavior, but will treat newly synthesized items—to which the high-ranking faithfulness constraints are inapplicable—according to lower-ranked markedness constraints, whose ranking reflects lexical frequencies.

As mentioned in Section 3.1, the reason that established and novel words are treated differently here lies in the underlying representations of the prefixes with floating [+nasal] features, which makes the two types of word subject to different constraints. As a general treatment of lexical variation, this approach relies on the existence of some difference in the faithfulness constraints to which listed words and fresh morpheme concatenations are subject

A final note on the choice of learning algorithm: Some phonologists modeling variation have been moving away from the Gradual Learning Algorithm because of convergence problems (e.g., Pater 2008), and a Maximum Entropy version of OT has become popular (Goldwater and Johnson 2003). In this case, however, the Gradual Learning Algorithm is preferable for its ability to "overlearn" lexical trends—that is, to acquire the subterranean ranking despite also learning that faithfulness to lexicalized words is inviolable. When Maximum Entropy grammars are learning with a substantial Gaussian prior, the penalty against giving a constraint a large weight grows as the square of the weight, so there is a tendency to distribute responsibility for the data over several constraints rather than loading all the weight on to just a few constraints—see Martin 2007 for interesting consequences for phonotactic learning. But in the Tagalog case, this weight-distributing tendency was not enough for the voicing or place effects to be learned more than slightly. With either the biased or the unbiased constraint set, *N̥C̥ receives enough weight to cause a difference in rate of substitution between the voiceless and voiced consonants of only 6 percentage points, and the difference in *[N weights is negligible, causing place differences of less than .01 percentage points. (Maximum Entropy simulations performed using Hayes et al. 2005.)

5 Evidence that speakers know the distribution

Having established that certain lexical trends exist, and that they can be learned and represented in the grammar without endangering the pronunciations of individual words, we now consider whether those trends should be represented in the grammar. That is, do Tagalog speakers know the distribution of nasal substitution, or is it a diachronic accident that goes unnoticed? This section presents two forms of evidence for speakers' implicit knowledge of the voicing and place effects. In both cases, the evidence is stronger for the voicing effect than for the place effect.

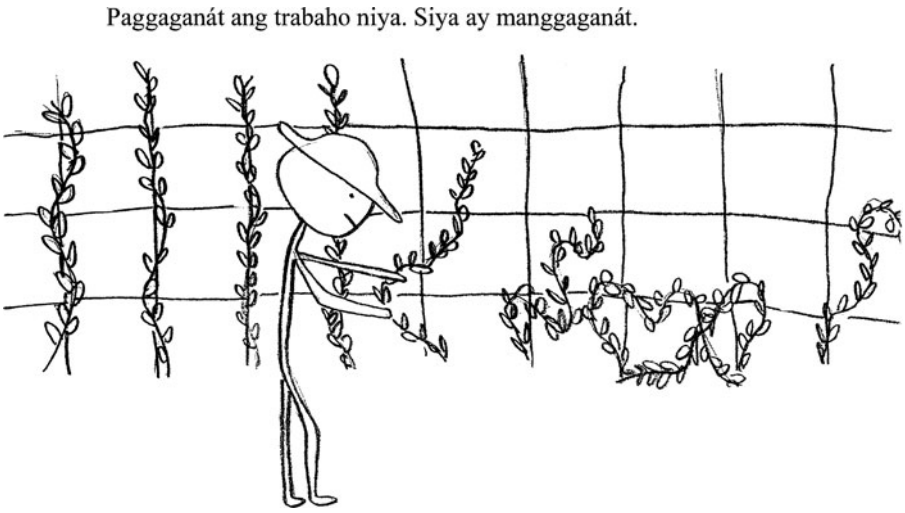
5.1 Evidence from an acceptability-judgment task

In a study previously reported in Zuraw 2000, nine native speakers of Tagalog living in the United States participated in study of nasal substitution in invented words. The participants ranged in age from 18 to 69, and had emigrated from the Philippines 3 to 20 years earlier. In a production task, which will not be discussed in detail here, participants produced nasal substitution at low (but nonzero) rates, but did not reliably display the voicing or place effects.

In the second task, participants were shown 50 cards, each with a cartoon-like drawing of a person performing a farming or craft activity, with two sentences printed

at the top. A sample card is shown in (33). The sentences were printed in normal Tagalog orthography, except that accent marks—which are optional and not commonly used—were employed to indicate stress (including penultimate stress, which even when accent marks are used is left as the default). The 50 stimuli were arrived at by constructing three novel roots each for *p*, *t*, *s*, *k*, *b*, and *g*, and four for *d* (two with intervocalic tapping in the *pag-RED-* form and two without). There were two cards for each of the 25 resulting roots, one with substitution and one without (same picture for the two cards).

(33) Sample card for Task II



The sentences were designed as an acceptability-judgment version of a "wug" test (Berko 1958) for the *maŋ-RED-* construction, which forms professional and habitual nouns (similarly to English *-er*). The *pag-RED-* construction, which does not permit nasal substitution, presents the novel root (⟨bugnát⟩, in the sentence shown in (34)), and the *maŋ-RED-* construction is applied either with or without nasal substitution.

(34) Sample stimulus

Pagbubugnát	ang	trabaho	niya.	Siya	ay	mamumugnát.
to-*bugnat*	TOPIC	job	his/her	he/she	INVERSION	*bugnat*-er

'His/her job is to *bugnat*. He/she is a *bugnat*-er.'

Each participant was given a stack of cards that started with four novel-word practice items (substituted and non-substituted for each of two stems), and then presented each root twice (but not consecutively; order was randomized), once substituted and once non-substituted. The participant read the sentences aloud, then stated his or her rating of the sentence pair, on a scale from 1 (bad) to 10 (good).

Participants' acceptability judgments generally reflected lexical frequencies. Figure 14 shows the combined average for each segment of the rating given to a substituted stimulus minus the rating given to the corresponding non-substituted stimulus.

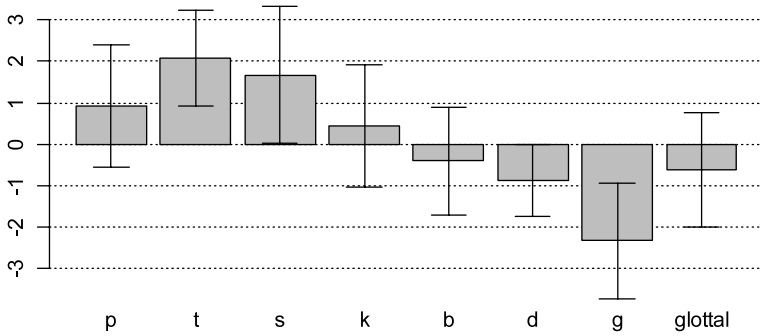


Fig. 14 Acceptability judgments: substituted–non-substituted; error bars indicate 95% confidence interval

A positive number means that over all, participants rated the substituted stimulus higher; a negative number means that over all, participants rated the non-substituted stimulus higher.

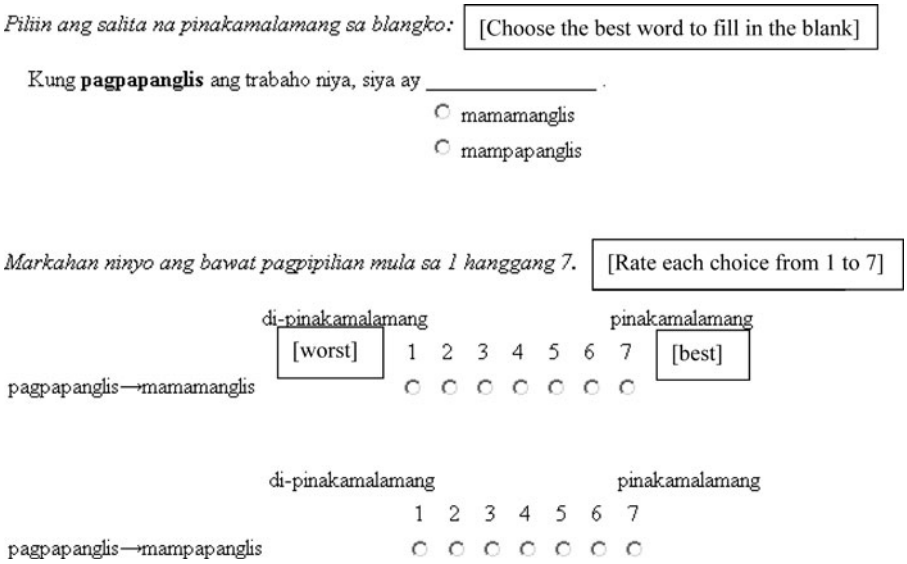
The positive numbers for voiceless-initial roots and negative numbers for voiced-initial roots mean that over all, participants preferred the substituted stimulus for voiceless-initial roots and preferred the non-substituted stimulus for the voiced-initial roots, reflecting the voicing effect. Looking at individual voiceless-voiced pairs, all the differences were significant (for each planned comparison, an unpaired, one-tailed t-test on rating differences was performed within each subject, and the resulting 9 p values were combined using Fisher's method: $p > b$ ($p < .05$), $t > d$ ($p < .001$), $s > d$ ($p < .005$), $k > g$ ($p < .05$). Acceptability judgments also did not contradict the place effect (the unexpectedly low ratings for p are not significantly different from t or s). Within the voiceless category, significant differences were $t > k$ ($p < .05$), $s > k$ ($p < .01$). Within the voiced category, the significant differences were $b > d$ ($p < .05$) and $b > g$ ($p < .001$). A drawback of Fisher's method for combining probabilities is that, in this case, a very low p-value for one subject can have a strong effect on the overall p-value. As a check on this, we can simply pool all the results and perform t-tests, ignoring subject information. In that case, the significant voicing differences are $t > d$ ($p < .0001$), $s > d$ ($p < .005$), $k > g$ ($p = .005$), and the significant place differences are $t > k$ ($p < .05$), $b > g$ ($p < .05$), $d > g$ ($p < .05$).

Another way of looking at the data is by fitting a mixed-effects model using the *lmer()* function in the *lme4* package of R (Bates et al. 2008), with *pvals.fnc()* of the *languageR* package (Baayen 2008) to assess significance. The model was fitted with participant as a random effect, voicing and place as fixed effects, and rating difference as the dependent variable. Voiceless consonants get a higher rating difference than voiced by 2.4 points ($p < .0001$). As for the place effect, labials get a higher rating difference than velars by 1.3 points ($p = .048$), and dentals get a higher rating difference than velars by 1.5 points ($p = .01$). As we have already seen, counter to prediction labials get a slightly *lower* rating difference than dentals—by 0.3 points—but this is not significant ($p = .65$).

5.2 Evidence from a binary-choice task

In order to gather data from a larger sample of speakers and to minimize item-specific effects, another experiment was performed. Participants were recruited over the web. After completing a brief form requesting demographic information, participants were shown a series of 12 screens, each with a fill-in-the-blank stimulus as follows, and two choices as to how to complete the second sentence and a request to rate each choice on a 1–7 scale.

(35) Sample stimulus, with translations shown



If a participant chose one option but rated the other higher, the response for that item was excluded as inconsistent.

The important differences in method between this survey and the study reported in Section 5.1 are as follows. First, because the experiment was conducted over the web, participants might be located anywhere in the world. Most of the participants were living in the Philippines. Given Zhang et al.'s (2010) finding that frequency effects were attenuated in Taiwanese speakers living abroad, the responses used were restricted to those from participants who reported both a birthplace and a current residence in the Philippines. A few participants were also excluded for giving incoherent responses to the demographic questions, or for choosing, from a drop-down menu in the demographic questionnaire, a frequency of Tagalog use that was less than daily. Second, each participant saw a different set of stimuli. For each participant, 12 stems were randomly generated beginning with *p*, *t*, *s*, *k*, *ʔ* (spelled with a hyphen), *b*, *d*, g, *l*, *w*, *j* (⟨y⟩), and *h* respectively. Each stem was of the shape CV(C)CV(C), and was checked against a list of real disyllabic roots collected from English's (1986) dictionary, and against an inventory of illegal CC clusters and illegal VC sequences. The order of the stimuli was randomized, and the order of the two options (substituted and non-substituted) was also randomized for each stimulus. Third, the inclusion of

sonorant-initial stems allowed a sort of check. Recall that real sonorant-initial stems do not undergo nasal substitution. Yet, some participants chose a nasal-substituted option (e.g., ⟨manunupik⟩ for ⟨lupik⟩) for at least one of the sonorant-initial stems. This might indicate a lack of attention to the base word (i.e., the participant was merely judging the *maŋ-RED-* prefixed form, not the correspondence between the *pag-RED-* form and the *maŋ-RED-* form). These participants were excluded. Participants had to give at least 11 consistent answers to be included. After all these criteria were applied, there were 21 usable participants. (Requiring a perfect 12 consistent answers would have reduced usable participants to 16.)

Some minor differences are that accent marks were not used, to avoid possible font problems, and illustrations were not used, to avoid slow download times. Rather than being paid for their time, participants were "rewarded" with interesting language facts in question-and-answer form: every second screen, a new question was presented, and in order to see the answer the participant would have to complete two more items.

Because participants could supply up to one inconsistent response or non-response without being excluded, there are 20 or 21 usable responses for each obstruent. Figure 15 shows, for each stem-initial consonant, the proportion of participants who selected the nasal-substituted option. The results clearly reflect the voicing effect. By Fisher's exact test (1-tailed), there is a significant difference between the number of substitution responses for p vs. b ($p = .019$), t vs. d ($p = .013$), and s vs. d ($p = .041$), though not k vs. g ($p = .260$). As for the place effect, differences are suggestive but not significant. The one unexpected result is for g, which almost never nasal-substitutes in the real lexicon, but for which the substituted option was chosen by nearly 30% of participants. A possible explanation lies in Tagalog orthography, which uses the digraph ⟨ng⟩ for the velar nasal. Thus, for a hypothetical stem *gibat*, the substituted form *maŋiŋibat* would be spelled ⟨mangingibat⟩ and the non-substituted form *maŋgigibat* would be spelled ⟨manggigibat⟩. I had hoped that the use of the reduplicated form would minimize the visual confusion between the ⟨ng⟩ and ⟨ngg⟩ sequences, but perhaps this was not successful for all participants.

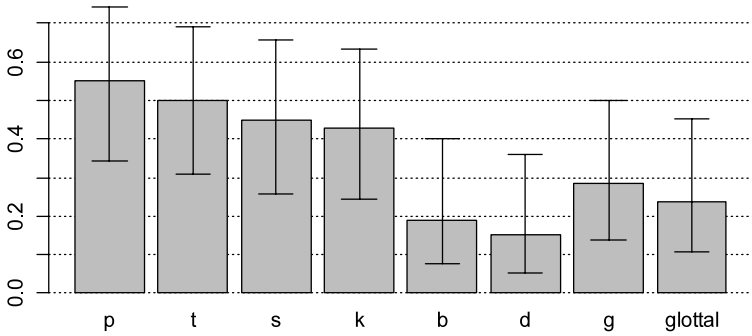


Fig. 15 Rates at which subjects selected nasal-substituted option in web survey; error bars indicate 95% confidence intervals[21]

[21] Using the *binconf()* function of the *Hmisc* package of R (Harrell 2008), default Wilson method.

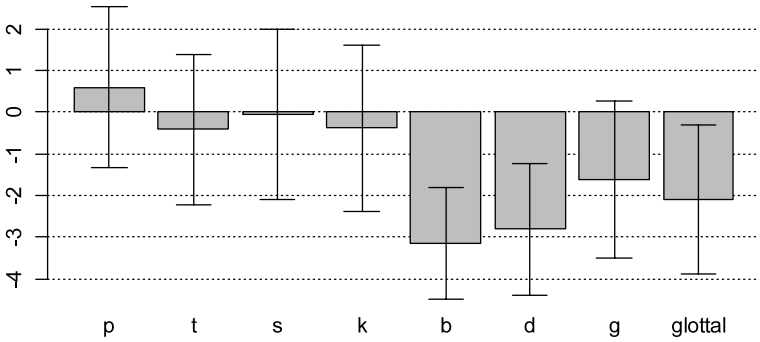


Fig. 16 Acceptability-judgment differences in the web survey; error bars indicate 95% confidence intervals

As in the first experiment, we can fit a linear mixed-effects model (this time using logistic regression, since the dependent variable is binary). The voiceless consonants are significantly more likely to prompt a nasal-substituted response ($p = .0001$), but the place differences are not significant.

Results for acceptability judgments in this experiment are, not surprisingly, similar to the results for the binary choices—see Fig. 16. When a linear mixed-effects model of the ratings differences is fitted, the voicing effect is significant ($p = .0001$), but not the place effect.

5.3 Evidence from loans

Another source of evidence for implicit knowledge of the voicing and place effects is loans. Tagalog was in close contact with Spanish from about the mid sixteenth to early twentieth century, and with English since the early twentieth century, and has adopted a large number of words from both languages. Enough stems from Spanish (though not English) have entered into potentially nasal-substituting constructions that we can examine the distribution of nasal substitution in these words. Figure 17 shows, for dictionary and corpus data respectively, the rates of substitution for each initial obstruent in these words, combining all affixal constructions. The voicing effect has been clearly perpetuated in these new words. The place-of-articulation effect is less clear, but there does seem to be a difference between *b* on the one hand and *d* and *g* on the other (though the number of *d*- and *g*-stems is surprisingly small).

I conclude from the data in this section that the voicing and place effects should be represented in the grammar.

6 Cross-linguistic data

This section discusses the voicing and place effects in related languages. We will see that although these two effects are almost universally respected, the exact pattern varies. This suggests that even if the available constraint set is biased (e.g., *NC̥ but no *NÇ), speakers still must learn how to rank the markedness constraints against each other and against faithfulness.

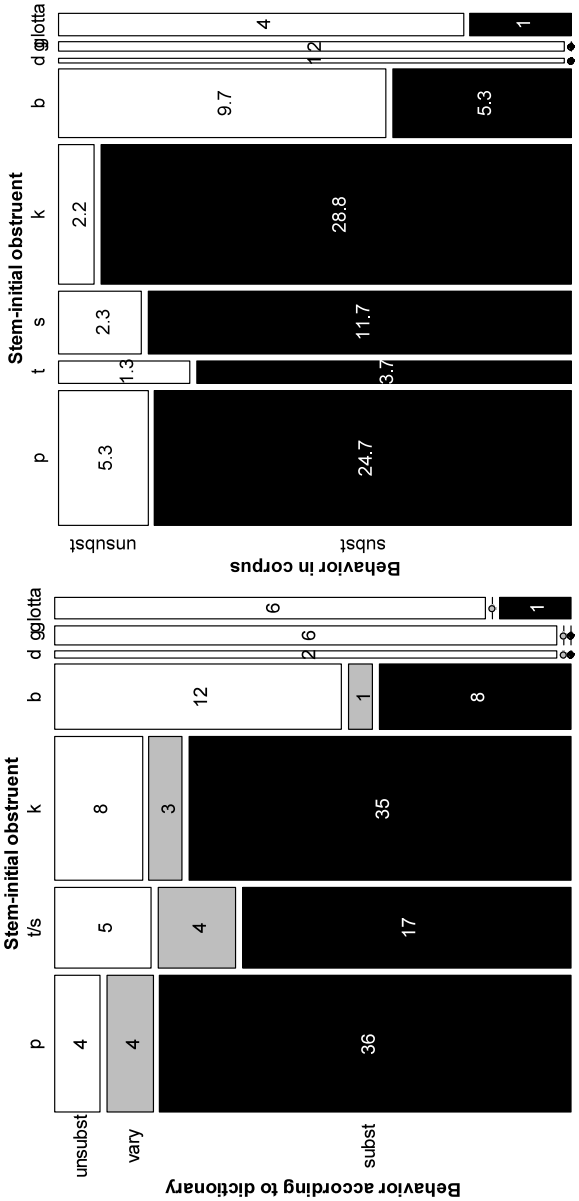


Fig. 17 Substitution rates for Spanish stems, all affixal patterns combined

Newman (1984) surveys nasal substitution in several Western Austronesian languages whose nasal substitution is reported to be less variable—that is, languages where the stem-initial consonant (almost) entirely predicts whether nasal substitution applies, so that we can say of a given consonant that it either does or does not undergo nasal substitution. Newman finds that in his sample, if nasal substitution applies to a voiced obstruent, then it applies to the corresponding voiceless obstruent. And if nasal substitution applies to a stop, then it applies to any fronter stop of the same voicing (fricatives may not fit the place pattern). Blust (2004), in a survey of 48 languages, replicates Newman's findings, except for in the case of Kapampangan (see below).

It is not known what the pattern of nasal substitution was in the proto-language (see Blust 2004 for some speculations).[22] Depending on the pattern's starting point, nasal substitution has either retreated from less susceptible segments in some daughter languages, spread to more susceptible segments in some daughter languages, or some of each. But whatever the case, as illustrated in this section, there seems to be great cross-linguistic consistency in what stem-initial consonants are more or less susceptible. This suggests that the voicing and place effects shape the diachronic development of a language's lexicon.[23]

Re-ranking the constraints proposed in Section 2 yields 10 language types, shown in Table 5, if we restrict our attention to voiced and voiceless stops at three places of articulation, ignore the possibility of deleting nasality altogether, and ignore the possibility of variation (see Blust 2004's survey for information on other stem types and on cases of nasal deletion). The typology checked with the help of OTSoft (Hayes et al. 2005), using exactly the items and violations in (32), except that UNIFORMITY, INTEGRITY, FAITH-OO, and NOCODA were omitted because they are redundant for these items.

Pattern (a), with no nasal substitution at all, is represented by the Sulawesi[24] languages Da'a (Barr 1995) and Wolio (Anceaux and Grimes 1995), where the descendants of nasal-substituting prefixes induce prenasalization, not substitution; and by Bugis (Sulawesi; Abas and Grimes 1995), where they produce gemination (/maŋ-tunu/ → [mattunu] 'burn s.th., bake s.th.'). If we accept Ross's (1988) evidence that "[c]ases of nasal substitution are preserved sporadically in Oceanic languages"

[22] Some possible evidence comes from Ross's (1988) discussion of fossilized forms in the Eastern Malayo-Polynesian family, which has lost productive nasal substitution, but retains some fossilized cases. Ross gives examples of substitution inherited for *t,*s,*k,*b,*D. For inherited non-substitution, he gives examples with *d,*D.

[23] Malay/Indonesian presents one of the few cases where change can be observed in the written record. Currently, Malay/Indonesian has a system in which nasal substitution applies to all the voiceless obstruents and none of the voiced (Lapoliwa 1981; though see Delilkan 2002 for prosodic and morphological complications). But, as Newman (1984) points out, Brakel (1973) claims that substitution can be found on voiced obstruents in 16th and 17th-century Malay manuscripts, with some such words "maintain[ing] themselves as archaic forms till well into the 19th C." (Brakel 1973:4). It is not clear from Brakel's discussion whether substitution was the norm on (at least some) voiced obstruents in these manuscripts, but we can at least say that the lexicon of Malay has been reshaped over the last few hundred years to reflect a different grammar of nasal substitution.

[24] All language-family information is from Gordon 2005.

Table 5 Factorial typology

	languages	*substituted?*						*sample ranking*
		p	t	k	b	d	g	
a	Da'a, Wolio, Bugis	−	−	−	−	−	−	*[ŋ, *[n, *[m, *ASSOC ≫ DEP-C, *NC̥
b	similar to Balantak	+	−	−	−	−	−	*[ŋ, *[n ≫ *NC̥ ≫ *ASSOC, *[m ≫ DEP-C
c	?	+	−	−	+	−	−	*[ŋ, *[n ≫ *NC̥, DEP-C ≫ *ASSOC, *[m
d	similar to Yami	+	+	−	−	−	−	*[ŋ ≫ *NC̥ ≫ *[n, *[m, *ASSOC ≫ DEP-C
e	sim. to Toba Batak	+	+	−	+	−	−	*[ŋ ≫ *NC̥ ≫ *[n ≫ DEP-C ≫ *ASSOC, *[m
f	?	+	+	−	+	+	−	*[ŋ ≫ *NC̥, DEP-C ≫ *[n, *[m, *ASSOC
g	Malay/Indonesian	+	+	+	−	−	−	*NC̥ ≫ *[ŋ, *[n, *[m, *ASSOC ≫ DEP-C
h	Sama-Badjau, Dibabawon Manobo	+	+	+	+	−	−	*NC̥ ≫ *[ŋ, *[n ≫ DEP-C ≫ *[m, *ASSOC
i	Cebuano, Isnag, Sarangani Manobo	+	+	+	+	+	−	*NC̥ ≫ *[ŋ ≫ DEP-C ≫ *[n, *[m, *ASSOC
j	Kalinga	+	+	+	+	+	+	*NC̥ or DEP-C ≫ *[ŋ, *[n, *[m, * ASSOC

notation adapted from Newman 1984: 10

(p. 41), then nasal substitution has also died out in the entire Central/Eastern Malayo-Polynesian branch of Malayo-Polynesian (rather than being an innovation confined to Western Malayo-Polynesian).

Pattern (b), with substitution on *p* only, is represented by Balantak (Sulawesi; Busenitz and Busenitz 1991; Busenitz 1994), where nasal substitution applies to *p*-initial stems, unless the next syllable also begins with *p* (Busenitz 1994:3). Pattern (c), with substitution on both labials, seems not to be attested. Yami (Northern Philippine; West 1995) almost exemplifies pattern (d): it distinguishes *p*, *t* from the rest, but the difference is that *p*, *t* are reported to undergo nasal substitution uniformly, and the other stops vary. Pattern (e) is similar to Toba Batak (see below), and pattern (f) does not seem to be attested.

Patterns (g), (h), and (i), where *NC̥ is respected, are robustly attested. Pattern (g) occurs in Indonesian/Malay (Sundic; Lapoliwa 1981) and many others; pattern (h) is found in Sama-Bajau (Sama-Bajaw; Verheijen 1986) and Dibabawon Manobo (S. Philippine; Forster 1970). Pattern (i) is found in Cebuano (Meso-Philippine; Wolff 1962)[25] and Isnag (N. Philippine; Vanoverbergh 1972).[26] Pattern (j) is exemplified by

[25] van Ogijk 1959, a description aimed at missionaries, appears to claim that application of nasal substitution is variable in Cebuano, but the passage (p. 44) is difficult to interpret because Odijk appears to be describing the distribution of the prefixes *maŋ-/naŋ-/paŋ-* vs. *mag-/nag-/pag*-rather than the distribution of nasal substitution *vs*. non-substitution.

[26] There are very few *g*-initial stems in Vanoverbergh's (1972) dictionary, and none takes a relevant prefix. The word-initial *g* of other Philippine languages seems to correspond to Isnag orthographic ⟨x⟩, ([h] in some dialects and [ɣ] in others). When this consonant takes *maŋ-* or *paŋ-*, it behaves as a non-substituted *g*: ⟨xabí⟩ 'night', ⟨ma[ŋ]-gabí⟩ 'to abstain from rice and taro while in mourning' (p. 245), with one exception, ⟨ma[ŋ]-xakkí⟩ 'to have one's skin open piecemeal' (p. 248).

Table 6 Languages with variation

	p	t	k	b	d	ɖ	g	*sample ranking*
Yami	+	+	∼	∼		∼	∼	*NC̥ > *[ŋ ∼ DEP-C > *ASSOC ≫ *[n ∼ *[m
Sasak	+	+	+∼	∼	−		−	*NC̥ > *[ŋ ≫ *[n ≫ *[m ∼ DEP-C ≫ *ASSOC
T. Batak	+	+∼	−	−∼	−		−	*[ŋ ≫ *NC̥ > *[n ≫ *[m > DEP-C ≫ *ASSOC
K. Batak N^2-	+	−∼	−∼	−∼	−		−	*[ŋ ∼ *[n > *NC̥ ≫ *[m > DEP-C ≫ *ASSOC
K. Batak N^1-	+	+	∼	+∼	−		−	*NC̥ ∼ *[ŋ ≫ *[n ≫ DEP-C > *[m ≫ *ASSOC
K. Batak $N^{3/5}$-	+	+	+	−	−		−	*NC̥ ≫ *[ŋ ∼ *[n ∼ *[m ∼ *ASSOC ≫ DEP-C
Palawan	+	+	+	∼	−		−	*NC̥ ≫ *[ŋ ∼ *[n ≫ *[m ∼ DEP-C ≫ *ASSOC
Kapampangan	+	+	+	+	−∼		+∼	*NC̥ ≫ *[ŋ ∼ *[n ∼ DEP-C ≫ *[m ∼ *ASSOC

Limos Kalinga (N. Philippine; Ferreirinho 1993), Ginaang Kalinga (N. Philippine; Gieser 1970), and Sarangani Manobo (S. Philippine; DuBois 1976).

There are various languages that, because of variation, look like a hybrid of more than one of the simple language types listed above. They still respect the voicing and place effects. Some of these languages are shown in Table 6, where "∼" means that variation is reported, "+ ∼" that variation is reported but with a preference for substitution to apply, and "− ∼" that variation is reported with a preference for substitution not to apply. A cell is shaded if the language represented on that row is lacking the consonant. In Toba Batak, for example (Sundic; Nababan 1981; Percival 1981; van der Tuuk 1867/1971), it is reported that *p* always substitutes, *t* (and *s*) usually do, *b* usually doesn't, and *d* and *g* never do. In Karo Batak (Sundic), Woollams (1996) reports that nasal substitution applies differently with three different prefixes: N^1-, which marks active voice; N^2-, which forms intransitive verbs; and $N^{3/5}$-, which forms certain adjectives. A variable constraint ranking that would produce each system is shown in Table 6. The symbol ">" indicates that the constraint on the left tends to outrank that on the right, but with some variation; "∼" indicates seemingly equal ranking. (Information on Sasak, a Bali-Sasak language, is from Goris 1938. Information on Palawan, a Meso-Philippine language, is from Revel 1995. Information on Kapampangan, a N. Philippine language, is from Forman 1971a, 1971b; del Corro 1980.)

In all the languages included in Tryon (1995), the languages surveyed by Newman (1984) and by Blust (2004), and others whose descriptions I have encountered, there is only one clear exception to Newman's implicational generalizations about voicing and place. In Kapampangan, looking at Forman's (1971b) dictionary, both *d* and *g* vary, but with *d* non-substitution is more common, while with *g* substitution is more common. See Kaufman (2005) for a treatment of this case in terms on contrast preservation.

Besides the phonological constraints concerning voicing and place, there are additional phonological regularities in many languages' nasal substitution discussed in Newman (1984) and Blust (2004), including special treatment for pseudo-reduplicated stems, monosyllabic stems, or stems that contain a nasal+obstruent sequence. Thus, phonological regulation of the distribution of nasal substitution in the lexicon is cross-linguistically common. Even if these factors now have categorical effects, these languages must have gone through stages in which what are now

regularities were merely tendencies. Nomoto (2009) gives an intriguing example from Malay, where stem-initial *t͡ʃ* shows variation between substitution and not, as in *mən-t͡ʃinta* ∼ *mə-ɲinta* 'to love', from *t͡ʃinta*. Nomoto uses web data to show that substitution is much more frequent when the stem contains a nasal+obstruent cluster than when it does not. (He attributes this effect, as well as similar categorical effects in other languages, to a conjoined constraint that prevents a word from having two nasal+obstruent sequences if one can be eliminated by nasal substitution.)

There are some languages where the nasal+obstruent effect is reported to be categorical, such as Timugon Murut (Prentice 1971), where non-substitution is not an option if the stem contains a nasal+obstruent sequence—the prefix nasal must either substitute or delete. In order for a language to pass from a Malay-like state to a Murut-like state, probabilistic phonological effects on nasal substitution must not be a mere artifact of the lexicon, unnoticed by speakers. Rather they must be learned and able to shape the treatment of new and even existing words. Thus, I take the phonological regulation of nasal substitution cross-linguistically to support the idea that lexical regularities can become encoded in the grammar.

7 Conclusion

This paper has argued, using data from Tagalog nasal substitution, for a model of lexical variation in which existing words' pronunciations are determined by their lexical entries, but new items' pronunciations are determined by a grammar that—if appropriate constraints are available to the learner—reflects the lexical pattern. Thus, the lexical pattern can be perpetuated as new items enter the language. In the Tagalog case, this was seen in the rates of nasal substitution on Spanish loanwords in Fig. 17.

This final section will consider some further questions and areas for future research.

7.1 Other models of lexical variation

There are other approaches to incorporating lexical patterns into the grammar. One prominent approach is that of Bybee (e.g., Bybee 2001), in which schemas of varying strengths are learned from the existing lexicon and compete for application to new words. To test this framework on the Tagalog case, it would be necessary to implement an algorithm for learning schemas that is able to partition words according to their stem-initial consonant (or perhaps just according to the consonant's place and voicing features), and to implement the quantitative competition between the resulting schemas.

An approach that is implementationally much closer to the one here is that of indexed constraints, as developed in Pater (2006, 2008); Coetzee and Pater (2008); and Becker (2009). When, in the course of learning, a ranking contradiction occurs, a constraint is chosen to be "cloned", or split into two differently-ranked versions that apply to different sets of words. This approach belongs to a family of frameworks in which lexical items are indexed either to particular constraints or to constraint rankings: see Inkelas et al. (1997); Inkelas and Zoll (2007); Itô and Mester (1995, 1999); and Anttila (1997, 2002), among others.

For example, the contradictory behavior of /pa[+nas] + búhaj/ → [pam-**b**úhaj] and /pa[+nas] + balík/ → [pa-**m**alík] (see (21) and (22)) could lead to the creation of two *[m constraints, one ranked high and indexed to /búhaj/, one ranked low and indexed to /balík/. Other markedness constraints would be similarly cloned. (See Becker 2009 for an algorithm that chooses which constraint to clone.) As additional words were learned, they would be added to the appropriate constraint's list of forms to which it applies. Constraints would then be able to attract newly coined words in proportion to how many existing words they were associated with. These assignments would then determine, perhaps probabilistically, the new word's behavior.

With the constraints used here, the indexation approach doesn't capture the voicing effect, only the place effect. The reason is that whenever a contradiction is reached, a *[N constraint is cloned, because it is the remaining constraint that discriminates among the fewest winner-loser pairs (see Becker 2009). Once all three *[N constraints have been cloned, all the learning data are accounted for by the ranking MAX(+nas) ≫ *[$\text{ŋ}_{\text{certain words}}$ ≫ *[$\text{m}_{\text{certain words}}$ ≫ *[$\text{n}_{\text{certain words}}$ ≫ DEP-C ≫ {*ASSOC, *N$\underset{\circ}{\text{C}}$, *[$\text{ŋ}_{\text{other words}}$, *[$\text{n}_{\text{other words}}$, *[$\text{m}_{\text{other words}}$}. A new word is probabilistically assigned to one of the two relevant *[N constraints—either the one ranked above DEP-C or the one ranked below it—and this assignment fully determines the word's fate, with *N$\underset{\circ}{\text{C}}$ playing no role. For example, a *p*-initial and a *b*-initial novel stem have the same probability of being assigned to the higher-ranked *[m clone, and thus of resisting nasal substitution. Implementing the constraint-indexation approach would thus require a different set of constraints, perhaps one for each of the six consonants (see footnote 19).

7.2 Construction effects

As shown in Section 2.2, different morphological constructions display different overall rates of nasal substitution. Unfortunately, there is not enough data on Spanish loans to see if these differences are replicated there, and the experiments reported here used only one construction. It remains to be seen whether the differences in substitution rate between constructions must be directly learned, such as through the rankings of different FAITH-OO constraints, or whether they can emerge by some other means. For example, it was proposed above that some words might have a sharper prefix-stem boundary than others, preventing the merger of prefix material and stem material. If learners use semantic and/or distributional information to assign these boundaries, then plausibly some constructions would show a higher percentage of sharp boundaries than others. The work of Hay 2003 and Hay and Baayen 2005 suggests that the relative retrievability of affixes, stems, and whole words (as influenced mainly by frequency) can affect word structure probabilistically without the need for explicit categories of boundary symbols.

7.3 Lexicalization, and the coexistence of lexical and free variation

The question of how novel words go from the free variation predicted by the grammar to the lexical variation seen in, for instance, the Spanish loans in Tagalog is beyond the scope of this paper, but my assumption is when a speaker utters a newly coined

word, there is a chance that a listener will begin to lexicalize that word, and that words' pronunciations stabilize as they become lexicalized throughout the speech community. See Zuraw (2000) for some modeling of the process by which a new word attains a stable pronunciation through repeated interactions between speakers and listeners.

Although the variation pattern documented here is mainly lexical—most words have a fixed pronunciation—there is also some free variation even in some frequent words and words that are established enough to be listed in a dictionary. As we saw, English's (1986) dictionary reports many words to be variable in their pronunciation (the grey-shaded cells in Fig. 1 through Fig. 8 and Fig. 17). And there were a similar number of items whose spelling in the corpus was variable.

The model as implemented relies on morphologically complex words' being allowed to have their own lexical entries. It is also essential to the model that these complex lexical entries be used in preference to freshly concatenated affixes and stems. Under a dual-route model such as that assumed by Baayen and Schreuder 1999 and Hay 2003, when the complex words are frequent enough their lexical entries should be accessed and sent as input to the grammar. If a word is infrequent, however, there is the likelihood that morphological concatenation will be used instead—just as though the word were new. This potential variation in how a word is accessed could account for some words' variable pronunciation. Of course, if different speakers lexicalize a word differently there will also be inter-speaker variation.

7.4 Limits of learnability

The model presented here predicts that whether a lexical pattern can be learned depends on whether it can be represented by the constraints available. If the constraint set is biased so that *ŋ* onsets are worse than *n* onsets, which are worse than *m* onsets, then a hypothetical Tagalog-like language whose lexicon came to have the reverse of the place effect would not cause the learner to acquire a subterranean grammar reflecting the pattern.

To test this directly, a simulation like those in Section 4.2 was carried out, but with the lexical statistics reversed, so that *p*, *t/s*, *k*, *b*, *d*, and *g* have the nasal-substitution rates of *g*, *d*, *b*, *k*, *t/s*, and *p* respectively—see the white bars in Fig. 18. The biased constraint set of (29) was used. The resulting grammar, whose behavior on new words is shown by the grey bars in Fig. 18, ranks *NC̥ low enough that there is no voicing difference at all; it also ranks the *[N constraints fairly low, but their implicational definition still produces a slight place effect in the real-Tagalog direction, counter to the pattern of the training data.

In this scenario, new words would not be assimilated into the unnatural pattern—their rates of nasal substitution would be about equal for all stem-initial consonants—so the pattern would tend to be diluted over time through the addition of these new words, the loss of existing words, and perhaps changes in behavior of existing words (see Martin 2007 on the diachronic maintenance of lexical patterns). The descriptive power of the model depends on the content of the constraints, and the ability, if any, that learners have to construct novel constraints.

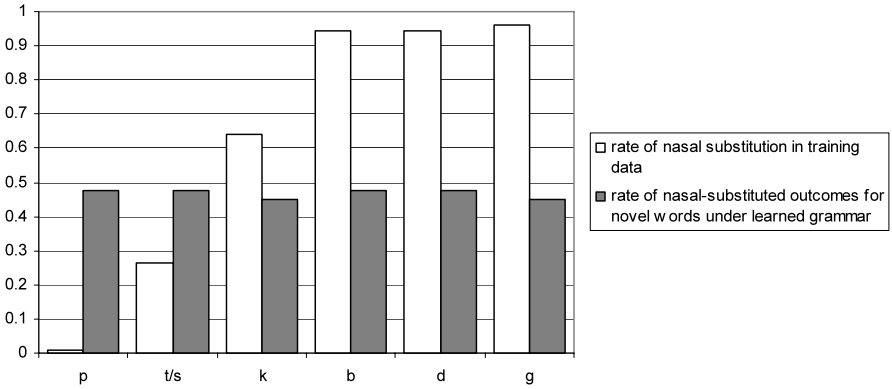


Fig. 18 Simulated (non-)learning of anti-Tagalog pattern

Acknowledgements This paper is a greatly revised version of portions of my 2000 UCLA dissertation. It has benefited from the comments and input of many people, who of course should not be held responsible for its errors or assumed to agree with its claims: Adam Albright, Marco Baroni, Katherine Crosswhite, Nenita Pambid Domingo, Bruce Hayes, Michael Kenstowicz, András Kornai, Peggy MacEachern, Donka Minkova, Joe Pater, Carson Schütze, Dan Silverman, Donca Steriade, and several reviewers. Ivan Tam wrote most of the software used in creating the web corpus, which took as its seed a corpus generously supplied by Rosie Jones from her own project, and Kevin Ryan did post-processing on the corpus. Thanks to Xiao Chen, Philip B. Ender, Michael Mitchell and Christine Wells for their advice on statistics for the somewhat unusual design of the survey. And thanks to all the anonymous survey participants who gave their judgments of novel words.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

Abas, Husen, and Charles E. Grimes. 1995. Bugis. In *Comparative Austronesian dictionary: an introduction to Austronesian studies, Part 1: Fascicle 1*, ed. Darrell T. Tryon, 549–561. Berlin/New York: de Gruyter.

Allen, Margaret. 1971. Morphological investigations. University of Connecticut, PhD dissertation.

Anceaux, J.C., and Charles E. Grimes. 1995. Wolio. In *Comparative Austronesian dictionary: an introduction to Austronesian studies, Part 1: Fascicle 1*, ed. Darrell T. Tryon, 573–584. Berlin/New York: de Gruyter.

Anttila, Arto. 1997. Deriving variation from grammar: a study of Finnish genitives. In *Variation, change, and phonological theory*, eds. F. Hinskens, R. van Hout, and L. Wetzels, 35–68. Amsterdam: John Benjamins.

Anttila, Arto. 2002. Morphologically conditioned phonological alternations. *Natural Language & Linguistic Theory* 20: 1–42.

Aronoff, Mark. 1976. *Word formation in generative grammar*. Cambridge: MIT Press.

Archangeli, Diana, Laura Moll, and Kazutoshi Ohno. 1998. Why not *NC̄? In *Proceedings of the 34th annual meeting of the Chicago Linguistic Society, Part I: the main session*, eds. M. Catherine Gruber, Derrick Higgins, Kenneth S. Olson, and Tamra Wysocki, 1–26. Chicago: Chicago Linguistic society.

Baayen, R. Harald. 2008. LanguageR: Data sets and functions with "Analyzing Linguistic Data: a practical introduction to statistics". R package version 0.953.

Baayen, R. Harald, and 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, Marco. 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*, eds. Geert Booij and Jaap van Marle, 121–152. Dordrecht: Kluwer.

Barr, Donald F. 1995. Da'a. In *Comparative Austronesian dictionary: an introduction to Austronesian studies, Part 1: Fascicle 1*, ed. Darrell T. Tryon, 529–537. Berlin/New York: de Gruyter.

Bates, Douglas, Martin Maechler, and Bin Dai. 2008. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-28. http://lme4.r-forge.r-project.org/

Becker, Michael. 2009. Phonological trends in the lexicon: the role of constraints. University of Massachusetts, Amherst, PhD dissertation.

Berko, Jean. 1958. The child's learning of English morphology. *Word* 14: 150–177.

Bhandari, Rita. 1997. Alignment and nasal substitution strategies in Austronesian languages. In *Recent papers in Austronesian linguistics*, ed. Matthew Pearson. Vol. 21 of *UCLA occasional papers in linguistics*, 59–69. Los Angeles: UCLA Department of Linguistics.

Blake, Frank. 1925. *A grammar of the Tagalog language*. New Haven: American Oriental Society.

Bloomfield, Leonard. 1917. *Tagalog texts with grammatical analysis*. Urbana: University of Illinois.

Blust, Robert. 2004. Austronesian nasal substitution: a survey. *Oceanic Linguistics* 43: 73–148.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. *Proceedings of the institute of phonetic sciences of the university of Amsterdam* 21: 43–58.

Boersma, Paul. 1998. *Functional phonology: formalizing the interactions between articulatory and perceptual drives*. The Hague: Holland Academic Graphics.

Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32: 45–86.

Brakel Lode F. 1973. Some notes on Malay morphology. *Linguistic Communications* 11: 1–10. Working Papers of the Linguistic Society of Australia

Busenitz, Robert L. 1994. Marking focus in Belantak. In *Studies in Sulawesi linguistics III*, ed. René van den Berg, 1–15. Jakarta: Universitas Katolik Indonesia Atma Jaya.

Busenitz, Robert L., and Marilyn J. Busenitz. 1991. Balantak phonology and morphophonemics. In *Studies in Sulawesi linguistics II*, ed. James N. Sneddon, 29–47. Jakarta: Universitas Katolik Indonesia Atma Jaya.

Bybee, Joan. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.

Carrier, Jill. 1979. The interaction of morphological and phonological rules in Tagalog. Massachusetts Institute of Technology, PhD dissertation.

Chomsky, Noam, and Morris Halle. 1968. *The sound pattern of English*. Cambridge: MIT Press.

Clements, George N. 1990. The role of the sonority cycle in core syllabification. In *Papers in laboratory phonology I: between the grammar and the physics of speech*, eds. John Kingston and Mary Beckman, 283–333. Cambridge: Cambridge University Press.

Coetzee, Andries. 2000. Post-nasal neutralization phenomena in Tswana: more on *N... constraints. Manuscript, University of Massachusetts, Amherst.

Coetzee, Andries, Susan Lin, and Rigardt Pretorius. 2007. Post-nasal devoicing in Tswana. In *Proceedings of the 16th international congress of phonetic sciences*, eds. Jürgen Trouvain and William J. Barry, 861–864.

Coetzee, Andries, and Joe Pater. 2008. The place of variation in phonological theory. Manuscript, University of Michigan and University of Massachusetts, Amherst. Draft chapter for 2nd edition of the *Handbook of phonological theory*, eds. John Goldsmith, Jason Riggle, and Alan Yu.

Cohn, Abigail, and John McCarthy. 1994/1998. Alignment and parallelism in Indonesian phonology. Manuscript, University of Massachusetts, Amherst. Published in *Working Papers of the Cornell Phonetics Laboratory* 12, 53–137.

De Guzman Videa. 1978. A case for nonphonological constraints on nasal substitution. *Oceanic Linguistics* 17: 87–106.

de Lacy, Paul. 2001. Markedness in prominent positions. In *HUMIT 2000, MIT Working Papers in Linguistics* 40, eds. Ora Matushansky, Albert Costa, Javier Martin-Gonzalez, Lance Nathan, and Adam Szczegielniak, 53–66.

de Lacy, Paul. 2002. The formal expression of markedness. University of Massachusetts, Amherst, PhD dissertation.

del Corro, Anicia. 1980. *Kapampangan morphophonemics*. Quezon City: Cecilio Lopez Archives of Philippine Languages and the Philippine Linguistics Circle. Publication 1 of *The Archive*.

Delilkan, Ann. 2002. Fusion and other segmental processes in Malay: the crucial role of prosody. New York University, PhD dissertation.

Dell, François, and Mohamed Elmedlaoui. 1985. Syllabic consonants and syllabification in Imdlawn Tashliyt Berber. *Journal of African Languages and Linguistics* 7: 105–130.

DuBois, Carl D. 1976. *Sarangani Manobo: an introductory guide*. Manila: Linguistic Society of the Philippines.

English, Leo. 1986. *Tagalog-English dictionary*. Manila: Congregation of the Most Holy Redeemer. Distributed by Philippine National Book Store.

Ferreirinho, Naomi. 1993. *Selected topics in the grammar of Limos Kalinga, the Philippines*. *Pacific Linguistics Series B* No. 109. Canberra: Australian National University.

Flack, Kathryn. 2007. The sources of phonological markedness. University of Massachusetts, Amherst, PhD dissertation.

Forman, Michael L. 1971a. *Kapampangan grammar notes*. Honolulu: University of Hawaii Press.

Forman, Michael L. 1971b. *Kapampangan dictionary*. Honolulu: University of Hawaii Press.

Forster, Jannette. 1970. Morphophonemic changes in Dibabawon. *Pacific Linguistics A* 24: 63–70.

French, Koleen Matsuda. 1988. *Insights into Tagalog: reduplication, infixation and stress from nonlinear phonology*. Dallas: Summer Institute of Linguistics and University of Texas at Arlington.

Fujimura, Osamu. 1962. Analysis of nasal consonants. *Journal of the Acoustical Society of America* 34: 1865–1875.

Ghani, Rayid, Rosie Jones, and Dunja Mladenić. 2004. Building minority language corpora by learning to generate Web search queries. *Knowledge and Information Systems* 7: 56–83.

Gieser, C. Richard. 1970. The morphophonemic system of Guininaang Kalinga. *Philippine Journal of Linguistics* 1: 52–68.

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a Maximum Entropy model. In *Proceedings of the Stockholm workshop on variation within Optimality Theory*, eds. Jennifer Spenader, Anders Eriksson, and Östen Dahl, 111–120.

Gordon, Raymond G. Jr., ed. 2005. *Ethnologue: languages of the world.*, 15th edn. Dallas: SIL International. Online version: http://www.ethnologue.com/.

Goris, R. 1938. *Beknopt Sasaksch–Nederlandsch woordenboek*. Singradja: Kirtya Liefrinck-van der Tuuk.

Halle, Morris. 2001. Infixation versus onset metathesis in Tagalog, Chamorro, and Toba Batak. In *Ken Hale: a life in language*, ed. Michael Kenstowicz. 153–168. Cambridge: MIT Press.

Harrell, Frank E. Jr. 2008. Hmisc: Harrell Miscellaneous. R package version 3.5-2. http://biostat.mc.vanderbilt.edu/s/Hmisc.

Hay, Jennifer. 2003. *Causes and consequences of word structure*. New York and London: Routledge.

Hay, Jennifer, and Harald Baayen. 2005. Shifting paradigms: gradient structure in morphology. *Trends in Cognitive Sciences* 9: 342–348.

Hayes, Bruce. 1999. Phonetically-driven phonology: the role of Optimality Theory and inductive grounding. In *Functionalism and formalism in linguistics*, eds. Michael Darnell, Edith Moravcsik, Frederick Newmeyer, Michael Noonan, and Kathleen Wheatly. Vol. I of *General papers*, 243–285. Amsterdam: John Benjamins.

Hayes, Bruce, and Margaret MacEachern. 1998. Quatrain form in English folk verse. *Language* 74: 473–507.

Hayes, Bruce, and Tanya Stivers. 1995. The phonetics of postnasal voicing. Manuscript, University of California, Los Angeles.

Hayes, Bruce, Bruce Tesar, Colin Wilson, and Kie Zuraw. 2005. OTSoft 2.3. Software package, www.linguistics.ucla.edu/people/hayes/otsoft/.

Herbert, Robert K. 1980. The role of perception in restructuring and relexicalization: two case histories. In *Papers from the 4th international conference on historical linguistics*, eds. Elizabeth Closs Traugott, Rebecca Labrum, and Susan Shepherd, 211–220. Amsterdam: John Benjamins.

Hyman, Larry. 2001. On the limits of phonetic determinism in phonology: *NC revisited. In *The role of speech perception in phonology*, eds. Elizabeth Hume and Keith Johnson, 141–185. San Diego: Academic Press.

Inkelas, Sharon, Orhan Orgun, and Cheryl Zoll. 1997. The implications of lexical exceptions for the nature of grammar. In *Derivations and Constraints in Phonology*, ed. Iggy Roca, 393–418. New York: Oxford University Press.

Inkelas, Sharon, and Cheryl Zoll. 2000. Reduplication as morphological doubling. Manuscript, University of California, Berkeley and Massachusetts Institute of Technology.

Inkelas, Sharon, and Cheryl Zoll. 2005. *Reduplication: doubling in morphology*. Cambridge: Cambridge University Press.

Inkelas, Sharon, and Cheryl Zoll. 2007. Is grammar dependence real? A comparison between cophonological and indexed constraint approaches to morphologically conditioned phonology. *Linguistics* 45: 133–171.

International Phonetic Association. 1999. *Handbook of the international phonetic association: a guide to the use of the international phonetic alphabet*. Cambridge: Cambridge University Press.

Itô, Junko. 1986. Syllable theory in prosodic phonology. University of Massachusetts, Amherst, PhD dissertation.

Itô, Junko, and Armin Mester. 1995. The core-periphery structure of the lexicon and constraints on reranking. In *University of Massachusetts occasional papers in linguistics 18: papers in Optimality Theory*, eds. Jill Beckman, Suzanne Urbanczyk, and Laura Walsh Dickey, 181–209.

Itô, Junko, and Armin Mester. 1999. The structure of the phonological lexicon. In *The handbook of Japanese linguistics*, ed. Natsuko Tsujimura, 62–100. Malden, Oxford: Blackwell.

Johnson, Keith. 1997. *Acoustic and auditory phonetics*. Cambridge: Blackwell.

Kaufman, Daniel. 2005. Austronesian nasal substitution and contrast neutralization. Manuscript, Cornell University.

Kiparsky, Paul. 1982. Lexical phonology and morphology. In *Linguistics in the Morning Calm*, ed. In-Seok Yang, 3–91, Seoul: Hanshin.

Kiparsky, Paul. 1993. Blocking in non-derived environments. In *Studies in lexical phonology*, eds. Sharon Hargus and Ellen Kaisse, 277–313. San Diego: Academic Press.

Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45: 715–762.

Lapoliwa, Hans. 1981. *A generative approach to the phonology of Bahasa Indonesia*. Canberra: Department of Linguistics, Research School of Pacific Studies, Australia National University.

MacBride, Alex. 2004. A constraint-base approach to morphology. University of California, Los Angeles, PhD dissertation.

Marantz, Alec. 1982. Re reduplication. *Linguistic Inquiry* 13: 435–482.

Martin, Andrew. 2007. *The evolving lexicon*. University of California, Los Angeles, PhD dissertation.

McCarthy, John. 1983. Consonantal morphology in the Chaha verb. In *Proceedings of the meeting of the west coast conference on formal linguistics 2*, eds. Michael Barlow, Daniel P. Flickinger, and Michael T. Wescoat, 176–188. Stanford: Stanford Linguistics Department.

McCarthy, John. 2002. Comparative markedness [long version]. In *University of Massachusetts occasional papers in linguistics 26: papers in Optimality Theory II*, eds. Angela C. Carpenter, Andries W. Coetzee, and Paul de Lacy, 171–246.

McCarthy, John. 2003. Comparative markedness [short version]. *Theoretical Linguistics* 29: 1–51.

McCarthy, John, and Alan Prince. 1993. Generalized alignment. In *Yearbook of morphology 1993*, eds. Geert Booij and Jap van Marle, 79–153. Dordrecht: Kluwer.

McCarthy, John, and Alan Prince. 1995. Faithfulness and reduplicative identity. In *University of Massachusetts occasional papers in linguistics 18: papers in Optimality Theory*, eds. Jill Beckman, Suzanne Urbanczyk, and Laura Walsh Dickey, 249–384.

Merriam-Webster. 1994. *Merriam-Webster's collegiate dictionary*, 10th edn. Springfield: Merriam-Webster. Frederic C. Mish, Editor in Chief.

Meyer, David, Achim Zeileis, and Kurt Hornik. 2006. The strucplot framework: visualizing multi-way contingency tables with vcd. *Journal of Statistical Software* 17: 1–48.

Meyer, David, Achim Zeileis, and Kurt Hornik. 2007. vcd: Visualizing Categorical Data. R package version 1.0-6. Software program.

Nababan, P.W.J. 1981. *A grammar of Toba-Batak. Materials in languages of Indonesia No. 6. Pacific Linguistics Series D* No. 37. Canberra: Australian National University.

Nagy, Naomi, and Bill Reynolds. 1997. Optimality theory and variable word-final deletion in Faetar. *Language Variation and Change* 9: 37–56.

Narayan, Chandan. 2006. Acoustic-perceptual salience and developmental speech perception. University of Michigan, PhD dissertation.

Newman, John. 1984. Nasal replacement in Western Austronesian: an overview. *Philippine Journal of Linguistics* 15: 1–17.

Nomoto, Hiroki. 2009. Distantly and prosodically conditioned nasal substitution in Austronesian languages. Talk presented at AFLA [Austronesian Formal Linguistics Association] XVI, University of California, Santa Cruz.

Ohala, John J., and Carol J. Riordan. 1979. Passive vocal tract enlargement during voiced stops. In *Speech communication papers*, eds. Jared J. Wolf, and Dennis H. Klatt, 89–92. New York: Acoustical Society of America.

Pater, Joe. 1999. Austronesian nasal substitution and other NC effects. In *The prosody morphology interface*, eds. René Kager, Harry van der Hulst, and Wim Zonneveld, 310–343. Cambridge: Cambridge University Press. [Edited version in *Optimality theory in phonology: a reader*, ed. John McCarhy. Oxford and Malden, Blackwell.]

Pater, Joe. 2001. Austronesian nasal substitution revisited: what's wrong with *NC̥. and what's not. In *Segmental phonology in Optimality Theory: constraints and representations*, ed. Linda Lombardi, 159–182. Cambridge: Cambridge University Press.

Pater, Joe. 2006. The locus of exceptionality: Morpheme-specific phonology as constraint indexation. In *UMOP: Papers in Optimality Theory III*, eds. Leah Bateman and Adam Werle, 1–36.

Pater, Joe. 2008. Gradual learning and convergence. *Linguistic Inquiry* 39: 334–345.

Percival, W.K. 1981. *A grammar of the urbanised Toba-Batak of Medan*. *Pacific Linguistics Series B* No. 76. Canberra: Australian National University.

Prentice, D.J. 1971. *The Murut languages of Sabah*. *Pacific Linguistics Series C* No. 18. Canberra: Australian National University.

Prince, Alan. 1997. Stringency and anti-Paninian hierarchies. Handout from lecture given at the LSA institute, 6/26/1997.

Prince, A., and P. Smolensky. 1993/2004. *Optimality Theory: constraint interaction in generative grammar*. Malden: 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.

R Development Core Team. 2007. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Software package, www.R-project.org.

Raimy, Eric. 2000. *The phonology and morphology of reduplication*. Berlin: Mouton de Gruyter.

Revel, Nicole. 1995. *Le palawan: phonologie, catégories, morphologie*. Paris: SELAF.

Reynolds, Bill, and Naomi Nagy. 1994. Phonological variation in Faetar: an Optimality account. In *Chicago Linguistic Society 30-II: Papers from the parasession on variation and linguistic theory*, eds. Katharine Beals, Jeannette Denton, Robert Knippen, Lynette Melnar, Suzuki Hisami, and Erica Zeinfeld, 277–292. Chicago: Chicago Linguistic Society.

Ross, Kie. 1996. Floating phonotactics: infixation and reduplication in Tagalog loanwords. University of California, Los Angeles, MA thesis.

Ross, Malcolm. 1988. *Proto Oceanic and the Austronesian languages of Western Melanesia*. Canberra: Pacific Linguistics.

Ross, Malcolm. 1995. Some current issues in Austronesian linguistics. In *Comparative Austronesian dictionary: an introduction to Austronesian studies, Part 1: Fascicle 1*, ed. Darrell T. Tryon, 45–120. Berlin/New York: de Gruyter.

Schachter, Paul, and Fe Otanes. 1972. *Tagalog reference grammar*. Berkeley: University of California Press.

Siegel, Dorothy. 1974. Topics in English morphology. Massachusetts Institute of Technology, PhD dissertation.

Tranel, Bernard. 1987. French schwa and nonlinear phonology. *Linguistics* 25: 845–866.

Tranel, B. 1996. Exceptionality in Optimality Theory and final consonants in French. In *Grammatical theory and Romance languages*, ed. Karen Zagona, 275–291. Amsterdam: Benjamins.

Tryon, Darrell T., ed. 1995. *Comparative Austronesian dictionary: an introduction to Austronesian studies*. Berlin/New York: de Gruyter.

van Ogijk, Antonio. 1959. *Elementary grammar of the Bisayan language*. Cebu: Convento Opon.

van der Tuuk, H.N.. 1867/1971. *A grammar of Toba Batak. Translation, with incorporation of van der Tuuk's manuscript notes and corrections, of Tobasche spraakunst*. The Hague: Martinus Nijhoff.

Vanoverbergh, Morice. 1972. *Isneg-English vocabulary*. Honolulu: University Press of Hawaii.

Verheijen, Jilis A.J. 1986. *The Sama/Bajau language in the Lesser Sunda Islands. Pacific Linguistics Series D No. 7*; *Materials in Languages of Indonesia*, No. 32. Canberra: Department of Linguistics, Research School of Pacific Studies, Australian National University.

Walther, Markus, and Richard Wiese. 1999. Optimization versus lexical specification. Handout from the workshop Conflicting Rules in Phonology and Syntax, University of Potsdam.

Weinreich, Uriel, William Labov, and Marvin Herzog. 1968. Empirical foundations for a theory of language change. In *Directions for historical linguistics*, eds. Winfred Lehmann and Yakov Malkiel, 97–195. Austin: University of Texas Press.

West, Anne. 1995. Yami. In *Comparative Austronesian Dictionary: an introduction to Austronesian studies, Part 1: Fascicle 1*, ed. Darrell T. Tryon, 315–320. Berlin/New York: de Gruyter.

Wilbur, Ronnie Bring. 1973. *The phonology of reduplication*. Bloomington: Indiana University Linguistics Club.

Wolff, John. 1962. *A description of Cebuano Visayan, Division of Modern Languages*. Ithaca: Cornell University.

Woollams, Geoff. 1996. *A Grammar of Karo Batak, Sumatra*. Canberra: Department of Linguistics, Research School of Pacific Studies, Australian National University.

Yip, Moira. 2002. *Tone*. Cambridge: Cambridge University Press.

Yip, Moira. 2007. Tone. In *The Cambridge handbook of phonological theory*, ed. Paul de Lacy. 229–251. Cambridge: Cambridge University Press.

Zhang, Jie, Yuwen Lai, and Craig Sailor. 2010. Effects of phonetics and frequency on the production of Taiwanese tone sandhi. In *Proceedings of Chicago Linguistic Society* 43.

Zonneveld, Wim. 1978. *A formal theory of exceptions in generative phonology*. Dordrecht: Foris.

Zoll, Cheryl. 1996. Parsing below the segment in a constraint based framework. University of California, Berkeley, PhD dissertation.

Zorc, R. David Paul. 1979. *Core etymological dictionary of Filipino*. Batchelor: Darwin Community College.

Zsiga, Elizabeth, Maria Gouskova, and One Tlale. 2006. On the status of voiced stops in Tswana: against *ND. In *Proceedings of the North Eastern Linguistic Society 36*, eds. Christopher Davis, Amy Rose Deal, and Youri Zabbal. 721–734. Amherst: GLSA.

Zuraw, Kie. 2000. Patterned exceptions in phonology. University of California, Los Angeles, PhD dissertation.

Zuraw, Kie. 2006. Using the web as a phonological corpus: a case study from Tagalog. In *EACL-2006: Proceedings of the 11th conference of the European Chapter of the Association for Computational Linguistics/proceedings of the 2nd international workshop on web as corpus*, 59–66.