Regularities in the Derived Lexicon[*]

Kie Zuraw
University of California, Los Angeles

Abstract

This paper identifies regularities in the distributions of exceptions to Tagalog nasal substitution and proposes that although information about exceptionality must be listed in the lexical entries of the words involved, lower-ranked markedness constraints encode the regularities and are active in shaping the way new derived words are incorporated into the lexicon.

1. Introduction

Polymorphemic words that (i) are derived from roots by morphology that is not fully productive (e.g., *nation-al* but **country-al*), (ii) differ phonologically from their stems ways that are not fully predictable (*ártist*, *artíst-ic*, but *Árab*, *Árabic*), (iii) are semantically noncompositional (*dis-ease*), or (iv) have a bound stem (*con-cur*) require their own lexical entries to contain unpredictable information about them.[1] These lexical entries make up the *derived lexicon*.

Although derivational phonology is typically exceptionful (property *ii* above), there are regularities in the distribution of exceptions in the derived lexicon. This paper examines regularities in Tagalog nasal substitution, and proposes a model in which those regularities are reflected in the grammar, despite the necessity for listing the words involved.

2. Nasal Substitution

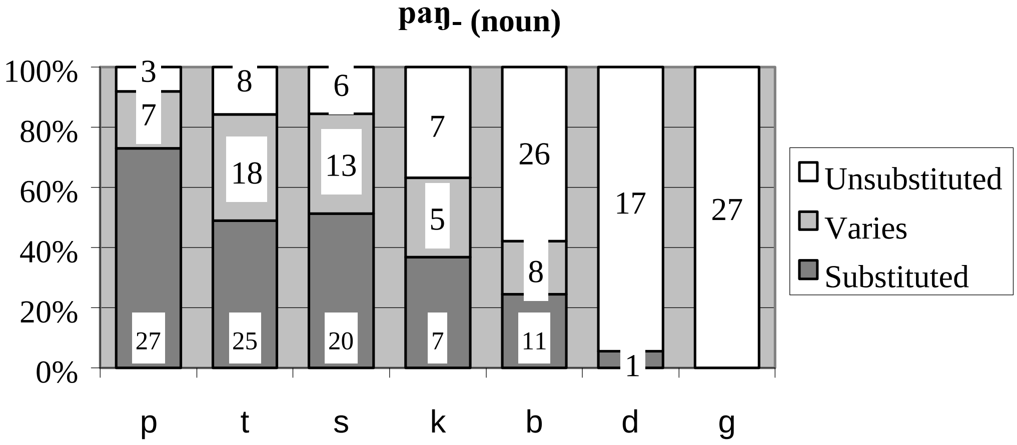
2.1 The data

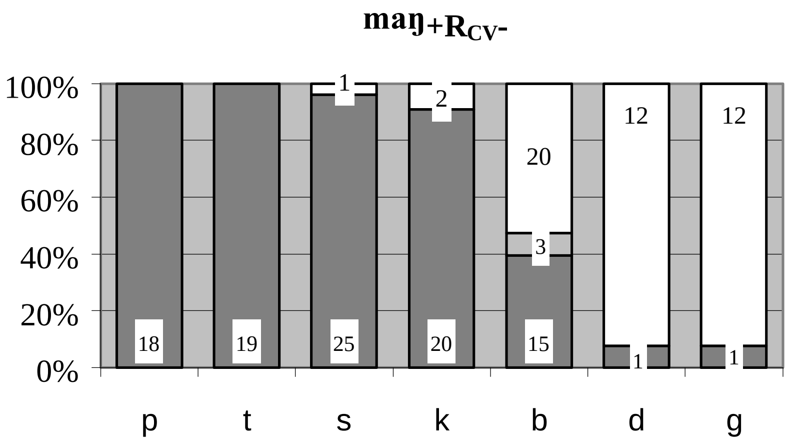
In Tagalog, nasal substitution coalesces a prefix-final nasal and a stem-initial obstruent (the coalescence analysis—as opposed to assimilation and deletion—is due to Lapoliwa 1981 and Pater 1996). The result is a nasal with the same place of articulation as the obstruent. Nasal substitution applies sporadically: (1) shows examples of substitution and of failure to substitute.

(1)	*p*	**p**ighatîʔ	‘grief’	pa-**m**i-**m**ighatîʔ	‘being in grief’
		poʔók	‘district’	pa**m**-**p**oʔók	‘local’
	t	pag-**t**úːloj	‘staying as guest’	kaː-pa-**n**ulúːj-an	‘fellow lodger’
		tabój	‘driving forward’	pa**n**-**t**abój	‘to goad’
	s	**s**úːlat	‘writing’	maː-**n**u-**n**ulát	‘writer’
				pa**n**-**s**úːlat	‘writing instrument’
	k	**k**amkám	‘usurpation’	ma-pa-**ŋ**amkám	‘rapacious’
		kaliskîs	‘scales’	pa**ŋ**-**k**aliskîs	‘tool for removing scales’
	b	mag-**b**igáj	‘to give’	ma-**m**igáj	‘to distribute’
		bigkás	‘pronouncing’	ma**m**-**b**i-**b**igkás	‘reciter’
	d	**d**aláːŋin	‘prayer’	ʔi-pa-**n**aláŋ-in	‘to pray’
		dinîg	‘audible’	pa**n**-**d**inîg	‘sense of hearing’
	g	**g**indáj	‘unsteadiness on feet’	pa-**ŋ**i-**ŋ**indáj	‘unsteadiness on feet’
		gáːwaj	‘witchcraft’	ma**ŋ**-**g**a-**g**áːwaj	‘witch’

Although the application of nasal substitution is sporadic and unpredictable, there are some regularities in its distribution.[2] Different morphological constructions have different overall rates of substitution, but within each construction, stems with a voiceless initial consonant are much more likely to substitute than stems with a voiced initial consonant, and stems with a fronter place of articulation are more likely to substitute than stems with a backer place of articulation. The chart in (2) illustrates the distribution of nasal substitution in just two common constructions, noun-forming *paŋ-* and professional-noun-forming *maŋ+R_{CV}-* (R_{CV} = CV reduplication). For example, of the 37 *p*-initial stems that take the noun-forming *paŋ-* construction, 73% substitute, but of the 45 *b*-initial stems, only 24% substitute (the voicing effect), and of the 19 *k*-initial stems, 37% substitute (the frontness effect).[3]

(2)





Nasal-substituted words must be listed, because (i) despite the lexical trends described above, exactly which words will undergo substitution is unpredictable; (ii) although the semantic connection between stem and derivative is always apparent, exact meanings are sometimes unpredictable; and (iii) there are sometimes unpredictable stress shifts.
(3) gives some examples of semantic unpredictability and of unpredictable stress shifts.

(3)	siʔîl	‘oppressed by a ruler’	ma-niʔîl	‘to strangle to death’
	balîk	‘return’	pa-malîk	‘hand rudder’
	tahîʔ	‘sewing’	maː-na-náːhiʔ	‘seamstress’
	túːbig	‘water’	ma-nubîg	‘to urinate’

2.2 An experiment

I conducted an experiment aimed at answering two questions: (i) is nasal substitution productive? and (ii) are speakers sensitive to the lexical patterns within nasal substitution? Nine native speakers of Tagalog living in Los Angeles were recruited. They ranged in age from 18 to 69, and had emigrated from the Philippines 3 to 12 years earlier.

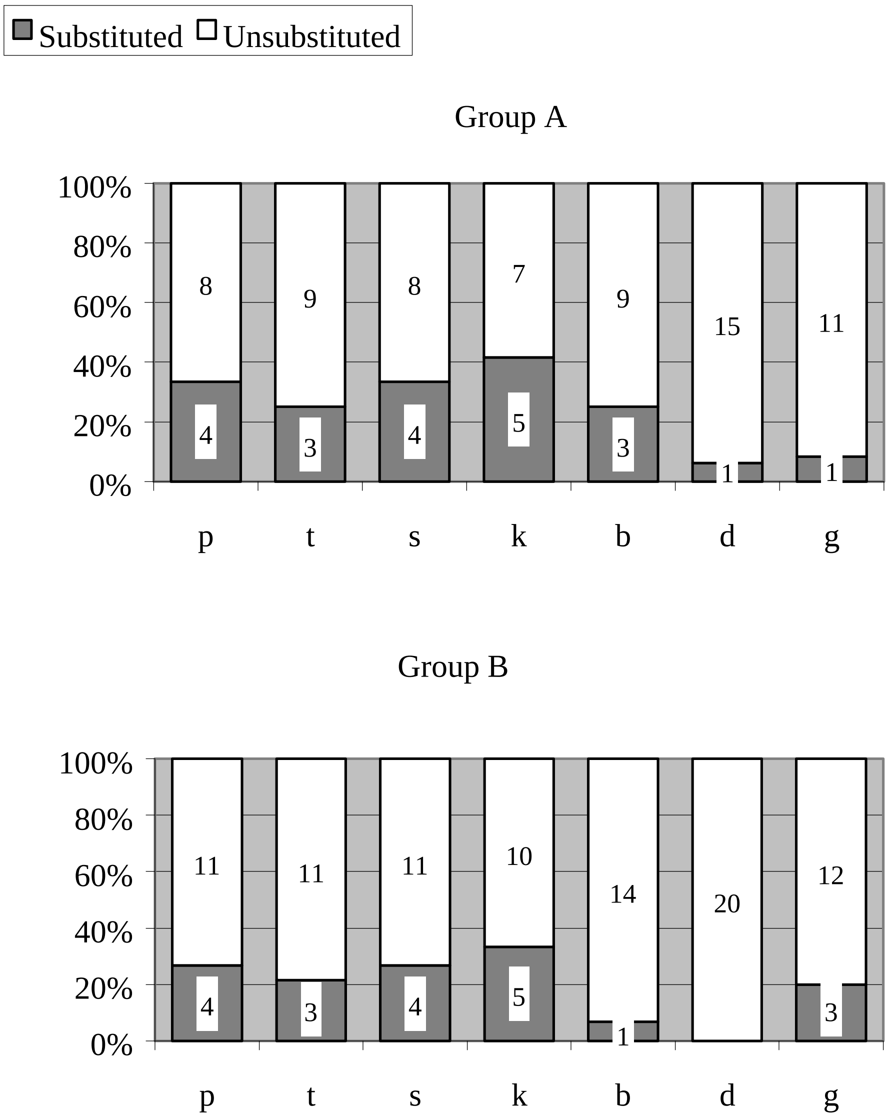
In the first task, participants were shown a series of cards, each of which had a crude illustration of person performing a farming or craft activity, with two sentences printed at the top. The sentences were designed as a “wug”-test (Berko 1958) for the *maŋ+*R_{CV}- construction, which forms professional and habitual nouns (similarly to English *–er*). Participants had to produce the *maŋ+*R_{CV}- form of a novel stem, deciding whether or not to perform nasal substitution. In the example shown in (4), the novel root is *bugnát*, presented in a construction (*pag+*R_{CV}-) that does not permit substitution. To fill in the blank, the participant would probably choose one of *maŋ-bu-bugnát* (no substitution, no assimilation), *mam-bu-bugnát* (assimilation only), or *ma-mu-mugnát* (substitution).

(4) Pagbubugnát ang trabaho niya. Siya ay ________________.
to-bugnat (topic) job his/her he/she (inversion)
His/her job is to *bugnat*. He/she is a ____________.

Participants in Group A (4 participants) were given some real roots mixed in with the novel roots, and were told that many of the words were rare and that if they didn’t know a word or its *maŋ+*R_{CV}- form, they should just guess. Participants in Group B (5 participants), were given only novel words after the training items, and were told that the words were made-up and there were no right or wrong answers.

Substitution rates, shown in (5), were much lower than the rates in the lexicon for *maŋ+*R_{CV}-, but were higher than zero. In other words, nasal substitution was neither very productive nor completely unproductive for this construction.

(5)

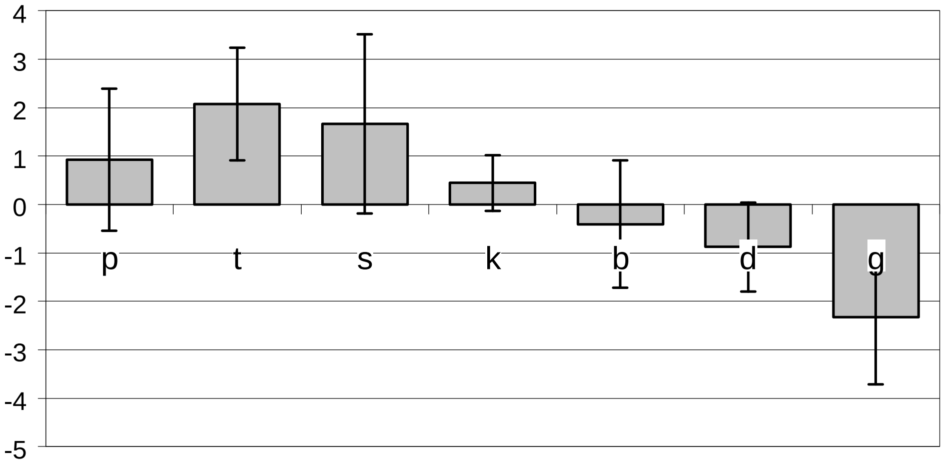


The second experimental task was designed to determine if participants were sensitive to the patterns of voicing and place of articulation seen in nasal substitution. Task II was administered immediately after Task I: starting with four novel-word practice items, each participant was given cards with the same illustrations and the same sentences as in Task I, but this time with the blanks filled in, as shown in (6). Each root was presented twice (but not consecutively; order was randomized), once substituted and once unsubstituted. The participant read the stimulus aloud, then rated it from 1 (bad) to 10 (good).

(6) Kung pagbubugnát ang trabaho niya, siya ay mamumugnát/mambubugnát.

(7) shows the average for each segment of the rating given to a substituted stimulus minus the rating given to the corresponding unsubstituted stimulus (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 unsubstituted stimulus higher.

(7)



The positive numbers for voiceless-initial roots and negative numbers for voiced-initial roots mean that over all, participants tended to prefer the substituted stimuli for voiceless-initial roots and tended to prefer the unsubstituted stimuli for the voiced-initial roots, reflecting the voicing effect. And, except for the unexpectedly low ratings for *p*, acceptability judgments also suggest sensitivity to the frontness effect.

2.3 The model

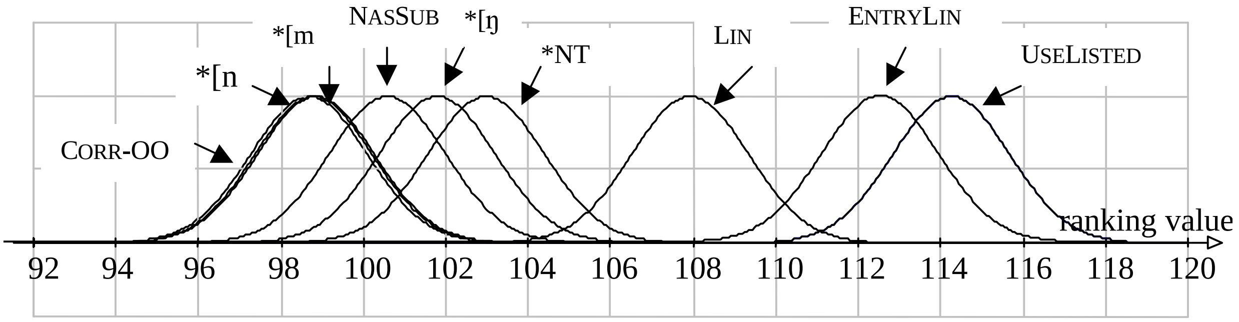
The experimental results suggest that nasal substitution and its patterns should be modeled in the grammar, despite the necessity for listing nasal-substituted words. The basic model that I will propose involves high-ranking input-output correspondence constraints that cause established words to be pronounced as listed, with lower-ranked markedness constraints that come into play when no listed form is available (as in a novel word). The relevant constraints are listed in (8)

(8) NASSUB A prefix-final nasal and a stem-initial obstruent must coalesce.[4]
*NT A sequence of a nasal and a voiceless obstruent is forbidden.[5]
This constraint requires nasal substitution in voiceless-initial stems.
*[ŋ, *[n, *[m Root-initial *ŋ* (*n*, *m*) is forbidden.[6]
One of these constraints is violated when nasal substitution occurs, since the coalesced nasal is root-initial (as well as prefix-final).
LINEARITY Coalescence or splitting[7] is forbidden.
This constraint is violated by nasal substitution.[8]
ENTRYLIN Coalescence or splitting is forbidden within a lexical entry.
CORR-OO Shorthand for IDENT-OO[SONORANT] and IDENT-OO[VOICE], where the correspondent output is the bare stem or any other unsubstituted form.
USELISTED Use as input a single lexical entry that has all the morphosyntactic and semantic properties of the utterance intent (see below).

I'm using a slightly different conception of *input* here than is standard: I assume that the real "input" to the tableau is the speaker's intent—that is, the morphosyntactic and semantic features that she wishes to express. Each candidate consists of an input-output pair. High-ranking constraints (not illustrated in the tableaux here) require that candidates' inputs have features that match the intent closely, but the inputs need not be the same for every candidate within a tableau. For example, if a speaker's intent is (roughly) 'to distribute (Actor Focus)' and she has a lexical entry */mamigáj/*, both */mamigáj/* and */maŋ+bigáj/* would be possible inputs. USELISTED favors candidates with */mamigáj/* as their input. But for a novel stem, such as *bugnát*, there is no single lexical entry expressing the intent 'to *bugnat*', and so all candidates will violate USELISTED. LINEARITY is violated whenever nasal substitution occurs. ENTRYLIN, however, can be violated only in candidates where the input is a single lexical entry (such as hypothetical */mampigáj/® [mampigáj]*).

When Boersma's (1997) Gradual Learning Algorithm is applied to a mini-lexicon of Tagalog which exhibits the voicing and frontness effects, the typical grammar learned is that in (9).[9] The grammar depicted in (9) is a stochastic grammar: each constraint has a ranking value (the center of each curve) along an arbitrary scale; in any given utterance, each constraint's ranking value is perturbed by a noise factor, generating a number that tends to be close to the ranking value (the numbers assigned to a constraint are normally distributed about the ranking value, as shown in (9)). The constraints are then ranked according to these numbers. The rankings that are generated vary, but within practical limits. For example, given the ranking values in (9), ENTRYLIN will outrank USELISTED 26% of the time, but *[m will outrank ENTRYLIN only 0. 000075% of the time.

(9)



So, in utterances where there is an appropriate listed word, the high rankings of USELISTED and ENTRYLIN ensure that it will almost certainly be faithfully parsed:[10]

	'to distribute'-AF	USELISTED	ENTRYLIN	Other Constraints
☞	/mamigáj/ → [mamigáj]			…
	/mamigáj/ → [mambigáj]		*!	…
	/maŋ+bigáj/ → [mambigáj]	*!		…
	/maŋ+bigáj/ → [mamigáj]	*!		…

In forming novel potentially nasal-substituted words, however, all candidates violate USELISTED and satisfy ENTRYLIN, so lower-ranked constraints decide. CORR-OO and LINEARITY tend to suppress nasal substitution, but given the grammar in (9), they can be outranked by other constraints that promote substitution. For example, the ranking shown below would produce nasal substitution for the novel stem *bugnát*, because NASSUB outranks LINEARITY, CORR-OO, and *[m. The probability of such a ranking is about 0.26%:

'to bugnat'-AF	UseListed	EntryLin	*[ŋ	*NT	*[n	NasSub	Linearity	Corr-OO	*[m
☞ /maŋ+bugnát/ → [mamugnát]	*						*	*	*
/maŋ+bugnát/ → [mambugnát]	*					*!			

2.3.1 The evolution of the lexicon

The probability of nasal substitution is not the same for all novel words. It is much higher on voiceless-initial stems, because a voiceless-initial stem substitutes if *either* NasSub *or* *NT outranks Linearity, Corr-OO, and the relevant *[nasal constraint. For example, the probability of a ranking that would produce substitution on a new *p*-initial stem is about 3.5% (compared to 0.26% for a *b*-initial stem). And the ranking of the *[nasal constraints in (9) means that fronter places of articulation tend to substitute more often than backer places: for *b* to substitute, NasSub must outrank *[m (as well as Linearity and Corr-OO), but for *g* to substitute, NasSub must outrank *[ŋ. since *[m tends to be ranked lower than *[ŋ does, it is more likely that NasSub will outrank *[m than that it will outrank *[ŋ. For example, compare the 3.5% probability of substituting a *p*-initial stem to the probability of substituting a *k*-initial stem, which is about 2.6%.[11]

Zuraw (in progress) gives a computational model of speaker-hearer interaction in the speech community in which these differences in rate of substitution are the seed for differences in the rates at which new words come to be listed in the lexicon as substituted. A key element of the model is the notion of *gradient listedness*: a lexical entry does not change instantaneously from being nonexistent to being fully available. Rather, its listedness is a function of how many times it has been heard—see the interesting effect found by Frisch (1999) in wordlikeness judgments of novel words: stimuli that had been heard twice were judged more wordlike than stimuli that had been heard just once. Listedness determines the probability that a lexical entry will be available in any given utterance. Thus, if */mamugnát/* is only partly listed, sometimes it will be available as an input, and UseListed will require that it will be used, but sometimes it will not be available, and lower-ranked constraints will decide whether the only available input, */maŋ+bugnát/*, undergoes substitution or not. This probabilistic availability of lexical entries means that a word's fate (to be substituted or not) is not sealed by the first speaker who coins it. Rather, the word's behavior starts out highly variable, and gradually becomes stable throughout the speech community.[12]

The model predicts that existing patterns in the lexicon should be perpetuated in new words. Although rates of nasal substitution in Spanish and English loanwords are still too low to determine whether the voicing and place effects are being perpetuated (and there are too few loan stems with established derived forms) the data on loanwords in another case—Tagalog vowel raising—suggest that there, low-ranking constraints are indeed shaping the incorporation of new derived words into the lexicon.

3. Conclusion

The general model presented here of exceptions in the derived lexicon is that lexical entries, when they exist, prevail; but, low-ranking constraints assert themselves when there is no lexical entry, or when an incompletely listed entry fails, by chance to be available. In cases like nasal substitution and vowel raising in Tagalog, bare stem forms are "primary" in the sense that they are more frequent, and loanwords usually occur in stem form for some time before derived forms develop. This means that low-ranked constraints will have the opportunity to determine outcomes more often in the derived lexicon than in bare roots or stems, and that the derived lexicon is the place to look for the effects of low-ranking constraints.

When a novel word is derived, CORR-OO and CORR-IO constraints discourage radical change from the stem form, which means that derived forms of newer loanword stems are less likely to alternate, and that low-frequency words are less likely to undergo derivational phonology. But, if a derived word is used often enough for a new lexical entry or allomorph to be created, the likelihood that that new entry or allomorph will differ from the stem reflects the influence of the low-ranking constraints.

Notes

[*] Because of space limits, this paper covers only the first half the talk it is based on. The second half of the talk applied the model to another case study, Tagalog vowel raising. For much more detail on nasal substitution, the experimental procedure and results, constraint definitions, the architecture and behavior of the model, and the vowel raising case, please see Zuraw (in progress). Comments welcome; please send them to ross@ucla.edu.

1. whether complete phoneme strings or merely a list of unpredictable properties

2. Data based on a complete count of all 1,736 obstruent-initial words with a potentially nasal-substituting prefix from English's (1986) dictionary.

3. Newman (1985) finds an implicational hierarchy reflecting similar effects in languages where nasal substitution is predictable if the stem-initial obstruent is known: If the language substitutes *g*, it also substitutes *d*, and if a language substitutes *d*, it substitutes *b*; similarly, substitution on *b* implies substitution on *k*, which implies substitution on *t*, *s*, and *p*. Thanks to Joe Pater for pointing out this interesting finding.

4. I have considered various phonetic and prosodic motivations proposed for nasal substitution in other languages (e.g., Archangeli, Moll, and Ohno 1998's *CC, Pater 1999's Alignment analysis of Indonesian), but they are not applicable to Tagalog. I think that Tagalog has simply inherited nasal substitution as an arbitrary alternation, and has analyzed it synchronically as driven by an arbitrary constraint.

5. See Hayes and Stivers (1996), Pater (1996).

6. These constraints are based on a pattern in the lexicon: root-initial nasals are relatively rare, with *n* and *ŋ* rarer than *m*. In general, more sonorant consonants are rarer root-initially, and fronter consonants are more common root-initially than backer consonants.

7. A splitting of segments $/m_1/ \rightarrow [m_1p_1]$ can be seen as violating LINEARITY, because in the input, 1 does not precede 1, but in the output, it does.

8. IDENT-IO[SONORANT] and IDENT-IO[VOICE] are also violated in nasal substitution.

9. The parameters of the algorithm must be set so that there is low initial plasticity. See Zuraw (in progress) for details.

10. The chance of generating a ranking in which USELISTED and ENTRYLIN outrank both *[m and CORR-OO (the constraints that would encourage an unfaithful parse in this case) is about 99.99986%.

11. The fact that these percentages are quite a bit lower than those seen in the experiment suggests that the ranking of LINEARITY is too high. This is a problem I am working on.

12. The reason the initial prevalence of nonsubstitution doesn't cause nearly all words to become listed as unsubstituted is that the hearer takes into account the probability that the speaker was using a prefix+stem input rather than a single, unsubstituted lexical entry.

References

Archangeli, D., L. Moll, and K. Ohno (1998). Why not *NC̥? To appear in the proceedings of the 34th annual meeting of the Chicago Linguistic Society.

Berko, J. (1958). The Child’s Learning of English Morphology. *Word* 14: 150-177.

Boersma, P. (1997). *How we learn variation, optionality, and probability*. Ms., University of Amsterdam.

English, L.J. (1986). *Tagalog-English Dictionary*. Manila: Congregation of The Most Holy Redeemer.

Frisch, S. (1999). Phonotactics in Phonology and Psycholinguistics. Paper presented at the Workshop on the Lexicon in Phonetics and Phonology, University of Alberta.

Hayes, B. and T. Stivers (1996). *The Phonetics of Post-Nasal Voicing.* Ms., University of California, Los Angeles.

Lapoliwa, H. (1981). *A generative approach to the phonology of Bahasa Indonesian*. Canberra: Australian National University.

Newman, J. (1985). Nasal replacement in Western Austronesian: An overview. *Philippine Journal of Linguistics* 15-16: 1-17.

Pater, J. (1996). *Austronesian Nasal Substitution and Other NC̥ Effects.* Ms., McGill University.

Pater, J. (1999). Generality and restrictiveness in constraint formulation: Austronesian nasal substitution and child consonant harmony. Handout from a talk given at the University of Massachusetts, Amherst.

Zuraw, K. (in progress). Exceptions and Regularities in Phonology. Dissertation, University of California, Los Angeles.