

# Surface Base Selection in Pengo<sup>1</sup>

Andrew Dowd

University of California, Santa Cruz

draft current 2/14/2009

## 1 Introduction

### *1.1. Background*

Albright (2002) provides a model for paradigm acquisition that begins by identifying a base form for the paradigm, and then develops a grammar of rules that derive all the other forms in the paradigm from this base. An important corollary of this model is that it restricts the UR to a surface form from among the output forms in the paradigm, and to the surface form from the same morphological context for all lexical items. This has been called the “single surface base hypothesis.”

This approach allows for an extremely computationally simple process of UR discovery. A learner attempting to construct a paradigm must simply learn the base, which is present in the data the learner is exposed to. The learner need not make any more complex inferences from the data to acquire the base of the paradigm. Albright (2002) shows just how easily this model may be algorithmically implemented.

Kenstowicz and Kisseberth (1977) argue that typological facts make the process of UR discovery more complex than a single surface base allows, and conclude that there is no general-purpose UR discovery procedure at all. Their particular arguments against the single surface base hypothesis as a basis for a discovery method involve systems in which it appears that the same surface form is not suitable as a base for all the lexical items of a particular class. They also raise a slightly larger issue, which is that there are systems in which there is a potential neutralization in all the forms in the paradigm, so the UR cannot be unambiguously reconstructed from any single surface form. Thus they conclude that abstract URs are a necessity.

This is a direct challenge to Albright’s model, which is incapable of constructing abstract URs. Therefore, in order to evaluate the overall feasibility of the model, it is instructive to consider whether or not these systems are accessible to an analysis in single surface base terms. Kenstowicz and Kisseberth’s examples of systems requiring information from more than one output form are from Russian and Pengo. Albright’s (2002, chapter 6) approach to the Russian data shows that while selecting a surface oblique form as the base does not reconstruct the correct paradigm in 100% of cases, the alternations it fails to predict are sufficiently rare to justify calling them exceptions.

---

<sup>1</sup> This is a greatly expanded version of a talk given at WCCFL XXIV in Vancouver in March of 2005. I would like to thank Adam Albright for his invaluable contributions to nearly every aspect of the development of this paper, Armin Mester for his gracious advising, Junko Ito and Pranav Anand for their participation, as well as Bruce Hayes and Afton Lewis for valuable comments, and James Isaacs and Ascander Dost for putting up with practice talks way back when.

This approach differs from an approach using abstract URs in that, while it allows “everywhere ambiguous” three-way alternations, they are always a result of memorized exceptions; that is non-phonological complications within the grammar.

### *1.2. Goals*

In this work I approach the Pengo data cited by Kenstowicz and Kisseberth and offer an essentially similar solution. I show that, even though no surface form in Pengo verbal paradigms is entirely non-neutralizing across all lexical items, so no single surface form can be used as a UR to unambiguously generate 100% of the attested forms, there is still a maximally informative form, one which appears to have escaped consideration by Kenstowicz and Kisseberth. Albright’s UR discovery procedure will select this form as the base, and derive the rest of the paradigm for each lexical item from it. This will fail to predict certain alternations attested in the data, and as in Russian, these forms will be claimed to be memorized exceptions. This reliance on memorized exceptions is argued to actually represent no net increase in memorization, however, as the base selection procedure actually significantly reduces idiosyncrasy elsewhere in the system. Additionally, the base selection procedure explains extant statistical patterns in the language by positing an analogical change from which the current pattern results.

In light of these treatments of the data from Russian and Pengo, it seems that Kenstowicz and Kisseberth’s objections to the single surface base hypothesis may not invalidate it at all. There are other examples in the literature of this type of multiple-neutralization system; for example Kenstowicz and Kisseberth mention Tonkawa, Yawelmani, and Turkish, but the Russian and Pengo results may encourage us that this pattern is actually illusory, and closer scrutiny may reveal that all of these systems are accessible to a single surface base analysis.

### *1.3. Pengo morphophonology*

Pengo is a tribal language spoken in the Koraput district in Orissa in India. It is in the Manda-Kui branch of South-Central Dravidian languages, and is closely related to Manda as well as Kui and Kuvi. Even in the 1950s when initial field research was carried out, the Pengo language was being edged out by Oriya, the majority language in the area. The Ethnologue reports 1,254 speakers in a 1961 census, but it is not clear whether the language is still spoken today. This makes it extremely difficult to find viable data on the language, as the only written source of vocabulary appears to be Burrow & Bhattacharya (1970).

What follows is an introduction to the main phonological and morphological characteristics of Pengo that will be relevant to the following discussion.

According to Burrow and Bhattacharya’s reckoning, Pengo has ten vowels and 22 consonants:

Table 1									
Pengo segmental inventory									
<i>a</i>	<i>a:</i>	<i>i</i>	<i>i:</i>	<i>u</i>	<i>u:</i>	<i>e</i>	<i>e:</i>	<i>o</i>	<i>o:</i>
<i>k</i>	<i>g</i>	<i>c</i>	<i>j</i>	<i>t</i>	<i>d</i>	<i>t</i>	<i>d</i>	<i>p</i>	<i>b</i>
	<i>ŋ</i>				<i>ŋ</i>		<i>n</i>		<i>m</i>
<i>h</i>						<i>s</i>	<i>z</i>		<i>v</i>
		<i>y</i>	<i>ɾ</i>			<i>r</i>	<i>l</i>		

In the notation I will be using, *j* represents the affricate *dz*, and *c* represents the corresponding voiceless affricate, which are palatalized before front vowels. orthographic *y* represents IPA [j]. While the above is the inventory given by Burrow and Bhattacharya, not all the contrasts within are entirely phonemic. *j* and *z*, for instance, are partly complementary in distribution, and mention will be made below of the *s/h* distinction.

Additionally, the nasal *ŋ* can occur either before a *g* or independently, but in the latter case a *g* has always been elided. Thus *ŋ* and *ŋg* are in complementary distribution, the *g* being elided finally and before consonantal suffixes and preserved before vowels. Similarly, *mb* is reduced to *m* finally and before consonants, but not all occurrences of *m* are due to elision of *b*.

One of the most extraordinary phonological processes in Pengo is the seemingly entirely productive metathesis of velar-labial clusters. The *kp* or *gb* clusters that are formed by suffixation of labial-initial suffixes to velar-final stems metathesize to *pk* or *bg*. This alternation has been profiled in some detail in Garrett and Blevins (in press).

Pengo has undergone a historically recent change of *s* to *h* in certain environments. The etymologically original *s* is preserved before voiceless stops and word-finally. Therefore *s* and *h* are in complementary distribution in the language<sup>2</sup>. As an example, this change produces alternations in *s*-final nouns, which become *h*-final before a vowel-initial suffix.

Table 2		
<i>s/h</i> alternations		
gloss	singular	plural
‘finger-ring’	<i>vatus</i>	<i>vatuhing</i>
‘sambhar’	<i>mangges</i>	<i>manggehing</i>

Pengo also has regressive voicing assimilation in obstruent-obstruent clusters. The voiced obstruents are devoiced before voiceless stop-initial suffixes, while the voiceless obstruents are voiced before voiced obstruent-initial suffixes.

Pengo verbs, in which I will be primarily interested, appear in a number of different morphological contexts, all consisting of a set of one or more suffixes. The

<sup>2</sup> Actually, Burrow and Bhattacharya report Oriya loanwords in which an original *s* is preserved in phonological environments where *h* would be expected, and in fact offer a number of *s*-initial words in their vocabulary list, often as variants of *h*-initial words. The distribution is thus probably somewhat freer than complementary.

forms I will be most interested in are the imperative, the past tense, the “special base,” the intensive-frequentative, the motion base, the gerund, and the infinitive.

The imperative is a suffix *-a*. A connecting glide *y* sometimes appears before the suffix after vowel final stems. This glide always occurs following front vowels, but is optional following back vowels:

Table 3		
imperatives		
gloss	past	imperative
‘spit’	<i>cuptan</i>	<i>cupa</i>
‘seize’	<i>astan</i>	<i>aha</i>
‘plough’	<i>ru:tan</i>	<i>rua</i>
‘sell’	<i>protan</i>	<i>proa/proya</i>
‘cut’	<i>koytan</i>	<i>koya</i>

Since the imperative suffix is vowel-initial, stems which participate in the *s-h* alternation described above show the final *h* in the imperative. Since this is a completely allophonic alternation, it does not result in any neutralization. There is for instance no pair like *\*ahtan/aha* to contrast with *astan/aha*. However, there are a few potential neutralizations in the imperative form. One arises from shortening of final long vowels in the imperative, leading to neutralization of long- and short-vowel-final verb stems. Another neutralization involves the glide insertion after vowel-final stems. Imperatives like *proya*, with root *pro-*, then fall together with imperatives of *y*-final roots like *koya* from the root *koy-*.

The past tense is formed by a suffix *-t*, which is followed by subject agreement morphology. Since it is only the past tense suffix itself which affects the stem phonology, I will give past tense forms with the 3sg masculine suffix *-an*. The past tense suffix *-t* causes the above-mentioned regressive voicing assimilation, resulting in a neutralization of all obstruent-final stems in this tense.

Table 4		
past tense forms		
gloss	past tense	imperative
‘give’	<i>hi:tan</i>	<i>hiya</i>
‘call’	<i>ku:ktan</i>	<i>ku:ka</i>
‘bore’	<i>pottan</i>	<i>pota</i>
‘dig’	<i>ka:rtan</i>	<i>ka:ra</i>
‘tread’	<i>toktan</i>	<i>toga</i>
‘gore’	<i>ustan</i>	<i>uza</i>
‘spread out’	<i>pa:stan</i>	<i>pa:ha</i>
‘grow’	<i>pantan</i>	<i>pana</i>
‘prod’	<i>kuntan</i>	<i>kunda</i>
‘sever’	<i>naʃtan/naʃtan</i>	<i>naʃa</i>
‘drink’	<i>uʃan</i>	<i>uʃa</i>

Other potential neutralizations affecting the past tense are the loss of the *d* in final *-nd* clusters, the change of *t* to *ɫ*, and the coalescence of the past tense suffix *-t-* and final *n* to *-ɫ*.

The special base, the intensive-frequentative, the motion base, the gerund, and the infinitive all have in common that they divide the verbal lexicon up into two essentially arbitrary lexical classes, which I will call A and B. Class A verbs take the A suffixes for these five forms, and class B verbs take the B suffixes.

Table 5		
verb classes		
	class A	class B
special base	<i>-t/-ta</i>	<i>-d/-da</i>
intensive/frequentative	<i>-pa</i>	<i>-ba</i>
motion base	<i>-ka</i>	<i>-ga</i>
gerund	<i>-ci/-hi/-si</i>	<i>-ji/-zi</i>
infinitive	<i>-teng</i>	<i>-deng</i>

As can be seen, the principal difference between the class A and class B suffixes is simply the voicing of the initial consonant. Since all these suffixes participate in the regressive voicing assimilation, this represents a potential two-way neutralization for all verbs ending in obstruents. Verb forms with class A suffixes would all appear with voiceless final obstruents in suffixed forms, and those with class B suffixes would all appear with voiced final obstruents. The potential typology is given below, comparing imperatives with gerunds for four hypothetical obstruent-final verbs; *ik-* and *ig-* in both classes:

Table 6		
factorial typology		
	class A	class B
voiceless-final	<i>ika~ikci</i>	<i>ika~igji</i>
voiced final	<i>iga~ikci</i>	<i>iga~igji</i>

These potential neutralizations rob these suffixed forms of informativity. It is impossible in every case to reconstruct the imperative form, *ik-* or *ig-*, from the suffixed form. As we will see, however, the attested typology in fact differs significantly from this hypothetical example.

These voice-changing suffixes do present the potential for other neutralizations as well. There is one additional neutralization in the so-called special base, which is the form of the verb used when the direct object is 1<sup>st</sup> or 2<sup>nd</sup> person. The class B special base suffix *-da* causes deletion of the *d* from final *nd* clusters, as in *kunda*, from the stem *kund-* ‘to prod,’ with imperative *kunda*. Additionally, there is a small class of exceptions to the regular *-ta/-da* special base. There are six common, mostly nasal-final, verbs that take an irregular suffix *-j* in the special base. These verbs elsewhere take the voiced suffixes, and typically cause a sound change in the infinitive suffix *-deng*, altering it to *-jeng*.

<i>-j</i> special bases		
imperative	special base	gloss
<i>ina</i>	<i>inj-</i>	‘say’
<i>tina</i>	<i>tinj-</i>	‘eat’
<i>puna</i>	<i>punj-</i>	‘know’
<i>mana</i>	<i>manj-</i>	‘be’
<i>vena</i>	<i>venj-</i>	‘hear’
<i>hala</i>	<i>hanj-</i>	‘go’

The intensive-frequentative base is, at first sight, much more irregular than most of these suffixes, as it is affected by the metathesis alternation whereby the *kp* or *gb* clusters that are formed by affixation of this suffix to velar-final verbs metathesize to *pk* or *bg*. Additionally, there is a class of verbs that do not take the standard *-pa/-ba* intensive suffixes at all, but instead take an invariant suffix *-ka*. This will be crucial later.

## 2 Theoretical preliminaries

### 2.1. The Single Surface Base Hypothesis

Many kinds of data appear to require reference to bases of morphological paradigms. This is certainly not an entirely uncontroversial statement, as some accounts of paradigms have involved no reference to bases within the paradigm at all, such as McCarthy’s (2002) view. But most other projects in phonology assume that there are underlying representations (URs) for phonological paradigms, and that these can be reconstructed from a single form in the paradigm, the base.

If we assume that morphological paradigms do contain privileged bases, we are faced with another issue. As linguists, we can identify the bases of paradigms by inferring them from base effects such as analogical changes. Language learners, however, must be able to identify bases for paradigms in the absence of such evidence, in order for these bases to create such effects. They must find the form that allows them to best predict the rest of the paradigm, in order to be able to produce novel forms of known words. Part of the project of phonology is uncovering how learners decide which form to use as the base. Essentially, we need a process of UR discovery for morphological paradigms.

Albright (2002) has proposed just such a discovery process. Essentially, in Albright’s model, learners select bases for paradigms based on a criterion of informativity. The most informative form in the paradigm is the one that reveals the most about the constant morphological and phonological properties of the stem. Effectively, it is the form that exhibits the fewest, or least serious, informativity-sapping neutralizations. An important corollary of this discovery procedure is that the base will always be a single, surface alternant, and it will be the same part of the paradigm for all lexical items.

The most informative form is what Kenstowicz & Kisseberth called “the position of maximal differentiation.” As an example, consider the imperative and past forms of various Pengo verbs.

Table 8		
neutralization in the past		
2SG imperative	3SGM past	gloss
<i>cupa</i>	<i>cuptan</i>	‘spit’
<i>tu:ba</i>	<i>tu:ptan</i>	‘blow’
<i>eca</i>	<i>eccan</i>	‘shoot’
<i>uja</i>	<i>uccan</i>	‘suck’
<i>ho:ka</i>	<i>ho:ktan</i>	‘wash clothes’
<i>maga</i>	<i>maktan</i>	‘sleep’

The imperative is formed by suffixing *-a* to the root, the 3sg masc. past is formed with the past tense suffix *-t* and the subject agreement suffix *-an*. The table illustrates that there is a voicing neutralization that takes place before the past tense suffix *-t*. obstruent-final roots may freely be voiced-final or voiceless-final in the imperative, but are restricted to voiceless-final in the past. This means it is possible to predict the past tense form from the imperative by supposing a devoicing rule, but it is not possible to predict the imperative form from the past tense form alone by any phonological rule. The imperative is thus the most informative form in this part of the paradigm. Kenstowicz and Kisseberth’s method for selecting the UR relied solely on this distinction between “neutralized positions” and the “position of maximal differentiation.”

Kenstowicz and Kisseberth, however, were looking for absolute categorizations of informative and uninformative forms. The methodology of the present project is somewhat different. Considering the above table, the various candidates for base (i.e., the imperative form and the past tense form) can be assigned percentile values of informativity. From the six imperative forms and the obstruent devoicing rule, all six past tense forms can be predicted. The imperative form would thus receive an informativity rating of 100%. From the six past tense forms, if no phonological rules are added, three of the imperative forms are correctly predicted, *cupa*, *eca*, and *ho:ka*. The other three forms, however, are predicted to be *\*tu:pa*, *\*uca*, and *\*maka*. The best that the past tense form can predict the imperative form is 50%, equal to chance. Since the imperative form scores higher on informativity for this simplified paradigm, our theory also predicts that the imperative should be selected as the base. Things change slightly when the data become more complicated and no form scores 100%.

Table 9		
two-way neutralizations		
2SG imperative	3SGM past	gloss
<i>proya</i>	<i>protan</i>	‘sell’
<i>koya</i>	<i>koytan</i>	‘cut’
<i>aha</i>	<i>astan</i>	‘seize’
<i>peza</i>	<i>pestan</i>	‘pick up’

From these four past tense forms and some of the phonological rules I gave above, such as glide insertion in imperatives and the s/h alternation, three of the corresponding

imperatives could be accurately predicted, *proya*, *koya*, and *aha*. *peza* would be incorrectly predicted as *\*peha*. Conversely, if a voicing rule were applied to the *s* before the past tense suffix *-t*, *peza* would be correctly predicted, but *aha* would not. It seems that the best the past form can predict the imperative for this sample is 75%.

From the four imperatives, and more phonology, such as the voicing assimilation rule, three past tense forms can be correctly predicted, *pestan*, *astan*, and *koytan*. *protan* is incorrectly predicted to be *\*proytan*. Likewise, if *y* glides are presumed to be epenthetic, *protan* is correctly predicted but *koytan* is not. Again, the best that the imperative form can do is 75%.

In this case, there is no single most informative form, and thus no clear winner in the base selection race. As we examine larger and larger samples of data, this type of complication becomes more and more common. It is rare to see an area in which a single form is 100% informative. Exceptions crop up, and no form can be used to predict 100% of the rest of the forms, but in samples of sufficient size, we can expect a single most informative form, the form with the highest percentile score for informativity. This is how the algorithm will select the base for a paradigm.

## 2.2. *Pengo as a counterexample*

For a long time, the general consensus in phonology seems to have been that the single surface base hypothesis is too restrictive to satisfactorily capture the full range of possible phonological systems. Kenstowicz and Kisseberth (1977, chapter 1), in particular, make the seminal claim that the UR must be significantly more abstract than a single surface form.

As part of an effort to illustrate this claim, Kenstowicz and Kisseberth entertain various hypotheses about the potential level of abstraction of URs. They refer to the single surface base hypothesis as *hypothesis B*’:

- (1) “The UR of a morpheme may include both variant and invariant phonetic properties. All of the variant properties selected to appear in the UR must occur in a single surface alternant of that morpheme, the basic alternant. The choice of the basic alternant is constrained by a principle of parallelism according to which the basic alternant for all morphemes of a given class (noun, verb, particle, etc.) must occur in the same morphological context.”

This hypothesis has the advantage of being a strong and concrete condition on URs. However, it is unable to admit URs with any abstract characteristics, such as underspecification or non-phonological information about alternations. This greatly restricts the predicted variability of phonological systems. If a phonological system were discovered that required access to more information than is present in a single output form across the paradigm, it would be strong evidence against this hypothesis.

Kenstowicz and Kisseberth believe that *Pengo* represents an example of just such a system. The presence of neutralizations in multiple forms in the *Pengo* verbal paradigm is crucial to this claim.

Table 10		
neutralization in the past, revisited		
2SG imperative	3SGM past	gloss
<i>cupa</i>	<i>cuptan</i>	‘spit’
<i>tu:ba</i>	<i>tu:ptan</i>	‘blow’
<i>eca</i>	<i>eccan</i>	‘shoot’
<i>uja</i>	<i>uccan</i>	‘suck’
<i>ho:ka</i>	<i>ho:ktan</i>	‘wash clothes’
<i>maga</i>	<i>maktan</i>	‘sleep’

As we saw above, this table shows that the past tense forms of obstruent-final verbs are predictable from the imperative forms, but not vice versa, due to the widespread neutralization of final obstruent voicing in the past tense. Kenstowicz and Kisseberth view this as an argument that the imperative form should be selected as underlying by any model consistent with hypothesis B”. This is a crucial step in the argument, one to which we will return in a later section.

Once the imperative is established as underlying in this sample, the parallelism constraint of B” requires that the imperative be the UR for all Pengo verbs, including the following:

Table 11		
-s final verbs		
2SG imperative	3SGM past	gloss
<i>aha</i>	<i>astan</i>	‘seize’
<i>gu:ha</i>	<i>gu:stan</i>	‘swallow’
<i>iha</i>	<i>istan</i>	‘strike’

With the assumption that the imperative is the base, hypothesis B” requires that the URs for these verbs be *ah-*, *gu:h-*, and *ih-*, respectively. This requires a rule converting *h* to *s* before voiceless segments. Such a rule does indeed produce all the attested alternations involving *s* and *h*, because the alternation is non-neutralizing.

It is important to note at this point that the predictive capacity of the imperative as the base is not diminished. In fact, Kenstowicz and Kisseberth’s argument is more subtle. There are, they claim, other considerations that suggest that *s* is a better candidate for the underlying form in these *s/h* pairs. These consist mainly in the presence of further *h/z* alternations in the forms under investigation that can only be described in plausible phonological terms by positing an underlying /s/. Kenstowicz and Kisseberth give two examples, which they feel constitute arguments against the supposition that /h/ could be the underlying form.

The first argument is from the transitive/intransitive alternations found in some Pengo verbs. These verbs occur in transitive/intransitive pairs, the intransitive ending in a voiced consonant, and the transitive in a voiceless consonant.

transitivity alternations		
intransitive	transitive	
imperative	imperative	gloss
<i>laba</i>	<i>lapa</i>	‘fit into’
<i>ruga</i>	<i>ruka</i>	‘hide’
<i>maga</i>	<i>maka</i>	‘lie/lay down’
<i>maza</i>	<i>maha</i>	‘turn’
<i>vi:za</i>	<i>vi:ha</i>	‘finish’

Notice that the transitive counterparts of *maza* and *vi:za* are not *masa* and *vi:sa*, as would be expected with a transparent application of devoicing, but instead *maha* and *vi:ha*. If the *h* is present in the underlying representation for the transitive forms, the voicing/devoicing rule will not be sufficient to predict these forms. An additional rule of *h/z* correspondence must be introduced. But if the underlying representation for the transitive forms contains *s* instead, the forms could be generated in the usual fashion, by devoicing *z*, and further phonology would then change the *s* into *h*.

The second instance of *h/z* alternations found by Kenstowicz and Kisseberth is in the gerund suffix. As has been mentioned before, certain suffixes in Pengo fall into arbitrary classes, class A beginning with voiceless consonants, and class B beginning with voiced consonants. The gerund is one of these voice-alternating suffixes. Its class B, voiced variant following vowel-final verb roots is *-zi*. Its class A, voiceless variant in the same environment is *-hi*. Again, positing the *h* in the underlying representation of *-hi* requires introduction of some new rule beyond the existing voicing alternation. If the underlying form contains *s*, however, the surface form will again fall out from general phonology.

This angle of argument that Kenstowicz and Kisseberth take is fairly strong, but it seems like it would not be particularly difficult to bite the bullet and maintain that *h* is present in all the underlying forms containing either *s* or *h*. For example, there is some evidence that the voicing alternation in transitive/intransitive pairs is not productive in the modern language, and we should not necessarily expect such an alternation to involve particularly transparent phonology. Additionally, the gerund has a number of other allomorphs that are not obviously phonologically related, namely *-ci* and *-ji*, so it would not be entirely surprising if the relation between *-zi* and *-hi* was more distant than Kenstowicz and Kisseberth imagine.

In short, the ability of Pengo to serve as a counterexample to hypothesis B” (the single surface base hypothesis) is somewhat limited by the unusual character of the voicing alternations common in the language’s verbal system. It is still possible to maintain that the imperative is the base form, and not lose any *predictive* capability over the more abstract alternatives advocated by Kenstowicz and Kisseberth. However, I will accept their reasoning and dismiss the imperative as a suitable candidate for the base, if only because I will shortly provide a candidate with a predictive ability demonstrably superior to the infinitive, as well as an explanatory power not shared by any abstract alternative.

### 3 Proposal

#### 3.1. Informativity

Informativity is a measure of how much of the structure necessary for predicting paradigmatic alternations is preserved in a particular surface form. A form can be depleted of its informativity by neutralizations, as we saw in the case of the past tense forms. However, informativity can also be depleted through non-phonological means. Unaffixed forms, for example, will lack any non-predictable information about particular suffixes simply because the suffix is not present. This has the net effect of making these forms less informative than the suffixed form.

Kenstowicz and Kisseberth's identification of the imperative as the base for Pengo verbal paradigms hinged on the superior informativity of the imperative when compared to the past tense form. The final obstruent voicing neutralization reduces the informativity of the past tense form.

However, this is the only pairwise comparison of different candidates for the base in their paper. No other possibilities are investigated. Pengo verbs are found in numerous other forms, most indicated by suffixes on the verb root. Among these suffixed forms, there is in fact a better candidate for the base in Pengo verbal paradigms than the imperative.

#### 3.2. Voice-changing suffixes

As seen in section 2, there are a number of Pengo verbal suffixes that alternate in voicing more or less arbitrarily. These are the special base, the intensive-frequentative, the motion base, the gerund, and the infinitive. Due to the strict regressive voicing assimilation in Pengo, these forms have dramatic phonological effects on the obstruent final verb stems that precede them. At first glance, this makes them rather unlikely candidates for basehood, as they could cause extremely broad neutralizations. Recall from Table 6 that the presence of these suffixes results in the following expected typology:

Table 6		
factorial typology		
	class A	class B
voiceless-final	ika~ikci	ika~igji
voiced final	iga~ikci	iga~igji

Obviously, this pattern does not exactly inspire confidence in the informativity of the suffixed form. All information about the voicing of the final obstruent in the unsuffixed forms is lost. But let us examine the distribution of these four possible forms in the actual language and see if there is any hope for improving the odds. All the numbers in the discussion that follows are based on my survey of the 203 verbs in Burrow & Bhattacharya's vocabulary section for which an infinitive or a gerund was given. Of the 130 such verbs ending in obstruents, the following distribution was observed:

Table 13		
observed distribution		
	Class A	Class B
	[-voi] suffix	[+voi] suffix
[-voi] obst root	60	27
[+voi] obst root	0	43

It is impossible to know what the independently expected distribution of such forms should be, there are simply too many factors to consider, such as overall phoneme frequency, frequency by position, frequency in verbs, etc. But if we rid ourselves of as many preconceived notions as possible and assume that a completely even distribution of root-final consonants is expected for each suffix class, we see something like this:

Table 14		
expected distribution		
	Class A	Class B
	[-voi] suffix	[+voi] suffix
[-voi] obst root	30	35
[+voi] obst root	30	35

Concentrating for the moment on the voiceless suffixes, we expect a distribution of 30/30, but observe 60/0. A quick binomial test will show that there is essentially zero probability that the observed distribution arose by chance ( $p = 0.0000000$ ).

If the suffixed forms are selected as bases, however, we expect the resultant neutralizations to collapse the four possible verb/suffix pairings into two. For example, obstruent-final verbs taking the voiceless suffixes would all end in voiceless consonants in the base form, and thus any difference in the final obstruent voicing in unsuffixed forms would be unrecoverable. This would result in an analogical change whereby all obstruent final verbs taking voiceless suffixes would become voiceless-final underlyingly. This hypothetical change would give us exactly the observed distribution for the verbs taking voiceless suffixes.

In the case of verbs taking voiced suffixes, the data are a bit more complicated. Assuming that the suffixed form is the base leads to the expectation that all verbs taking voiced suffixes should also have changed to underlyingly voiced-final. But the observed distribution is 43/27. Only 61% of these verbs are underlyingly voiced-final. This looks like an effect in the expected direction ( $p = 0.0154232$ ), but not nearly as strong an effect as might be expected, given the distribution of verbs taking voiceless suffixes.

### 3.2.2 *The intensive-frequentative*

In order to understand this effect, we must ask, which suffixed form is acting as the base? All the suffixed forms in Pengo are not created equal. They result in different neutralizations due to different phonological characteristics, and some may provide better candidates for basehood than others. The intensive form, in particular, deserves closer scrutiny. As has been discussed above, there is a class of verbs in Pengo that do not take

the usual alternating *-pa/-ba* intensive suffix. They take a non-alternating suffix *-ka* instead. This causes devoicing of stem-final obstruents. Interestingly, Burrow and Bhattacharya note that the *-ka* suffix is restricted to verbs that elsewhere take the voiced suffixes (*-ji* gerund, *-deng* infinitive, etc.)

If there is a correlation between having a voiceless stem-final consonant in the unsuffixed forms (like the imperative) and taking the invariant voiceless *-ka* intensive suffix, this could make the intensive a good candidate for the base. It would mean that the "underlying" voiceless consonant in the forms that alternate under suffixation was actually present in the base.

I have data on the intensive form for 11 of the 27 voiceless-final verb stems that take the voiced suffixes:

Table 15		
intensives of alternating forms		
imperative	intensive	gloss
<i>eca</i>	<i>ecka</i>	‘shoot w/ bow’
<i>kica</i>	<i>kicka</i>	‘pinch’
<i>paca</i>	<i>packa</i>	‘scratch’
<i>pi:ca</i>	<i>pi:cka</i>	‘squeeze’
<i>ʝoca</i>	<i>ʝocka</i>	‘weave’
<i>hi:pa</i>	<i>hi:pka</i>	‘sweep’
<i>kata</i>	<i>katka</i>	‘cut w/ axe’
<i>kaka</i>	<i>kabga</i>	‘vomit’
<i>ma:ka</i>	<i>ma:bga</i>	‘bake’
<i>ʔa:ka</i>	<i>ʔa:bga</i>	‘offer worship’
<i>ho:ka</i>	<i>ho:bga</i>	‘rub’

Of these 11 forms, the 4 that end in *k* undergo metathesis in the intensive, ending in *-bga*, the other 7 all take the *-ka* intensive suffix. The sample size is unfortunately small, but the important point to note is that *all* the voiceless-final verb roots for which I have intensives, excepting those that end in *k*, take the *-ka* intensive. There are no voiceless-final stems that take the *-ba* voiced intensive suffix on the surface, so (modulo metathesis) none of them have final consonant voicing in the intensive. If this trend is really general, then, out of the 27 alternating voiceless-final verb stems, the 20 which do not end in *k* could actually be regularly predicted from the intensive.

This does raise the issue of how to account for the forms with metathesis in the intensive. The Table 16 shows that among forms with metathesis, the intensive does not seem to be an adequate method of predicting the final obstruent voicing of the root in the unsuffixed forms, because metathesis is not restricted to a particular suffix class. All three attested combinations of root final and suffix initial obstruent voicing are possible with metathesis, voiceless-voiceless, voiced-voiced, and voiceless-voiced, leading to a three-way contrast:

Table 16		
metathesizing intensives		
imperative	intensive	gloss
<i>ku:ka</i>	<i>ku:pka</i>	‘call’
<i>ʈraka</i>	<i>ʈrapka</i>	‘hit’
<i>ɖe:ka</i>	<i>ɖe:pka</i>	‘seek’
<i>ɖrika</i>	<i>ɖripka</i>	‘break’ (tr.)
<i>tiga</i>	<i>tibga</i>	‘push’
<i>toga</i>	<i>tobga</i>	‘step on’
<i>paga</i>	<i>pabga</i>	‘split’ (intr.)
<i>pa:ga</i>	<i>pa:bga</i>	‘strike, kill’
<i>kaka</i>	<i>kabga</i>	‘vomit’
<i>ma:ka</i>	<i>ma:bga</i>	‘bake’
<i>ʎa:ka</i>	<i>ʎa:bga</i>	‘worship’
<i>ho:ka</i>	<i>ho:bga</i>	‘wash, rub’

Attempting to predict the unsuffixed forms from the intensive, postulating no voicing change, will incorrectly predict four forms, *kaka*, *ma:ka*, *ʎa:ka*, and *ho:ka*. The best ability of the model to predict unsuffixed forms from intensives with metathesis is thus 66.67%.

Since this is a fairly good figure, and the forms with metathesis represent such a small fraction of the total forms, I am going to call these four forms exceptions, and claim that these alternations are memorized by speakers and not derived regularly from a base. This is one of the provisions of the model, that the existence of exceptions is expected. Given the glosses of these verbs, it also seems safe to assume that they are all rather frequent, which is commonly correlated with irregular or exceptional phonology. Additionally, according to Burrow & Bhattacharya, the verb *kaka*, ‘to vomit,’ is principally found in (the reflexive of) the intensive form, *kabgiya*:

There are other exceptions to the rule that would derive underlying representations from intensives. These again appear to have an odd subregularity. The rule I have specified predicts that all forms with *-ka* intensives should either be voiceless obstruent final, or sonorant final. Out of 18 total verbs with *-ka* intensives in my sample, 7 are voiceless-final and 5 are sonorant final, but 6 end in voiced obstruents, so they alternate in the unsuffixed forms like the imperative. Under my proposal, all six of these forms must be regarded as exceptions, memorized by the learner. However, there might be an opportunity to reduce the net memorization by stipulating a single rule governing five of these exceptions. All five of these have unsuffixed forms which end in a nasal+obstruent cluster or *m<mb*. These clusters simplify under suffixation, so the nasal is missing in the intensive. So we have *kata~katka*, but *kunda~kutka*, and *hi:pa~hi:pka*, but *poma~popka*. In fact all the verbs ending in nasal+obstruent clusters for which I have intensive forms take the *-ka* intensive and participate in this alternation.

### 3.3 A base effect

The intensive-frequentative form, then, appears to possess sufficient informativity to predict the other forms in the Pengo verbal paradigm, due to the unusual distribution of suffix classes and final obstruents. In addition, this very distribution is an unexpected characteristic of the language that demands an explanation. However, considering the intensive-frequentative form to be a base for paradigmatic alternations in fact provides exactly the sort of explanatory power necessary to account for the broad conspiracy of statistical distributions that produces this informativity.

Specifically, this skewed distribution of final obstruents among verbs in suffix classes A and B in the modern language is evidence of an analogical change reorganizing the verb root final obstruents based on the intensive forms.

Presumably, since the attested typology has so little probability of having arisen by accident, there was a time in the history of the language when the voicing of verb root-final obstruents was entirely random with respect to suffix voicing, and verbs were free to vary in voicing under suffixation. The current state of the language, with final obstruent voicing dependent for the most part on the intensive-frequentative form, must have resulted from an analogical change that leveled the imperative form to the intensive.

In order for this change to occur, Pengo learners must have selected the intensive form as a base for the paradigm and extrapolated the unpredictable imperative voicing from it. Forms that had previously alternated would have been leveled out. This is why there are, for instance, no Pengo verbs that end in a voiced consonant but take the voiceless-initial suffixes.

I have no historical data on whether this change actually happened, or when, or to what members of the language family, but if my reconstructions are correct, then Pengo does in fact possess a single surface base, in the form of the intensive, for each verb in the language. Thus Pengo is consistent with the single surface base hypothesis.

This reasoning about the UR discovery process also answers a conceptual objection to the single surface base hypothesis. Kenstowicz and Kisseberth (1977) consider it obvious that “the underlying representation of a morpheme will appear unaltered only in some environments. There is no reason to expect that there will be a single environment in which all morphemes of a given class will be unaffected by a given morphophonemic rule.” While this seems true as a static typological generalization, from the perspective of a learner, it is irrelevant to the selection of a base for a paradigm. If there is a lesson to be learned from analogical change, it is that there is no reason to suspect that the base selected by learners should be free of all morphophonemic alternations.

Consider another, more textbook example of analogical change, Germanic paradigm leveling. Old English had a class of verbs with a rhotacism alternation in certain past tense forms. Modern English has lost the rhotacism due to analogical leveling, returning to the etymologically original *s* in all forms of these verbs.

Table 17		
Old English rhotacism alternation		
<i>ceosan</i> ‘choose’	OE	Mod.E
infinitive	<i>ceosan</i>	choose
past sg.	<i>ceosan</i>	chose
past pl.	<i>curon</i>	chose
past part.	<i>coren</i>	chosen

The same *s/r* variation existed in Old High German. While English leveled out *r* and kept *s*, German leveled out *s* and kept *r*.

Table 18					
English/German leveling					
English			German		
infinitive	past	past part.	infinitive	past	past part.
<i>choose</i>	<i>chose</i>	<i>chosen</i>	<i>küren</i>	<i>kor</i>	<i>gekoren</i>
<i>lose</i>	<i>lost</i>	<i>lost</i>	<i>verlieren</i>	<i>verlor</i>	<i>verloren</i>
<i>freeze</i>	<i>froze</i>	<i>frozen</i>	<i>frieren</i>	<i>fror</i>	<i>gefroren</i>

In German, the form selected as the base and generalized throughout the paradigm was one of the past forms with the alternant *r*. In Kenstowicz and Kisseberth’s terminology, however, the *r* in these forms is the result of a morphophonemic rule. Learners select surface forms as bases because they are accessible, not because they are morphophonologically simple. Base selection does not, in fact, respect the distinction between “original” forms and those derived by morphophonology.

### 3.4. Abstract URs

Accounts involving abstract URs to explain this sort of “everywhere ambiguous” alternation typically involve underspecification (Inkelas 1994, Inkelas, Orgun, and Zoll 1997). Non-alternating segments are taken to be fully specified in the UR, while alternating segments will be represented underlyingly by an incomplete featural specification, in this case, a segment underlyingly unspecified for voice. In this view, alternation is conditioned by information present in the UR that is not available in any surface representation.

Consider Table 19. Following the Prague School practice of including in underlying forms only those specifications common to all surface forms, the alternating final obstruents will be represented by an archiphoneme *K* underlyingly unspecified for voice:

Table 19		
underspecification		
	class A	class B
voiceless-final	ik-	iK-
voiced final	*	ig-

These empty voicing specifications will be filled in on the surface by rules, such as the regressive voicing assimilation found in Pengo. An underspecification view of three-way alternations suggests that the locus of variation rests in the stem-final obstruent.

However, the underspecification fails to predict the distribution of suffix-initial consonant voicing in this case. For example, nothing in principle prevents the unattested combination of a fully-specified voiced root-final obstruent with a voiceless suffix \**igci*. Nor can the lack of attested stems with underspecified iK- taking voiceless suffixes be explained.

Underspecification, archiphonemic or otherwise, is not a useful method for explaining these “everywhere ambiguous” alternations when they are *externally* conditioned. The essential problem is that the locus of variation is not found within the alternating root-final obstruent, but rests in the idiosyncratic voicing of the following suffix-initial consonant. And underspecification of voicing is not appropriate for explaining the voicing contrast between the class A and B suffixes, since their alternation is not conditioned phonologically at all.

Regardless of underspecification, the typology of exceptions noted in Table 16, for example, suggests the possibility that bases in Pengo may make reference to information in surface forms other than the intensive-frequentative. This would entail, minimally, a multiple base hypothesis, which Kenstowicz and Kisseberth schematize as *hypothesis C*:

- (2) “The UR of a morpheme includes those variant and invariant phonetic properties that are idiosyncratic. But all of the variant properties assigned to the UR must occur together in at least one phonetic manifestation of the morpheme. This manifestation can be referred to as the **basic alternant**.”

This would allow the exceptional forms whose final consonant voicing was not predictable from the intensive to select another basic alternant, say the imperative. However, the crucial aspect of the data is that metathesizing forms, since they may or may not alternate in voicing, are not predictable from *either* the unsuffixed or the intensive form. The intensive form will fail to predict the root-final consonant voicing in forms which alternate, and the imperative will fail to predict any of the suffix-initial consonant voicings.

Kenstowicz and Kisseberth ultimately propose that accounting for the idiosyncratic suffix-initial voicing is beyond the ability of any surface features, and represent the information diacritically, as (A root) and (B root) features. Only bases which actually contain the suffixes can predict suffix-initial voicing without recourse to such features.

However, diacritic features must presumably be memorized by the learner independently for each word, and require information from multiple surface forms. This makes the computational task considerably more difficult.

A more general deficiency in the hypothesis that Pengo URs are constructed without reference to a single privileged surface base is that it completely fails to explain the statistical correlation between verb-root-final obstruent voicing and intensive suffix-initial voicing. Allowing the information to be represented diacritically suggests that any possible combination of voicings should be freely attested, since predictability is

effectively irrelevant. Moreover, denying that the intensive is the single privileged base requires ignoring the evidence of the analogical change. Pengo learners must have at some point decided that the intensive form was the single surface form from which all the other forms should be derived, presumably switching to this form from another, and reorganized the phonology of the verbal paradigms accordingly.

In the following sections I will demonstrate that the intensive is in fact the most informative form in the Pengo verbal paradigm in the sense I have been using informativity, using results from computer simulations run on Albright's program.

## **4 Experimental methods**

### *4.1. The Minimal Generalization Learner*

#### *4.1.1. Phonology*

Running simulations on the Pengo data presents a few distinct challenges, because of the small sample size, complex phonology, and the idiosyncratic nature of the voicing alternation in the suffixes under consideration. For this reason, it was necessary to "cheat" in two particular ways when running these simulations.

Firstly, the learner is capable of discovering phonological rules in the language. It is designed to be provided with a list of phonotactically illicit sequences of segments, and to use the training data to discover phonological alternations just as it discovers morphological alternations. In attempting to accurately recreate the situation of children attempting to learn natural languages, this is important, because the evidence for phonological rules comes precisely from alternations in morphologically related words.

However, due to the fact that we were running simulations on such a small area of the language, there was not enough data for the learner to accurately determine the necessary phonology. Additionally, the learner often attempted to treat the idiosyncratic voicing of the voice-changing suffixes as a phonological alternation, and produced a number of distracting rules. This is in general a positive feature of the model, as it is intended to reflect the systems natural language speakers have for extending idiosyncratic or irregular alternations to novel forms, but it obscured the morphological alternations in question in this experiment.

So in this case the model was run using precompiled phonology. The learner did not learn any phonological generalizations from the training data, but instead used a list of rules which had been extracted ahead of time from linguistic examination of the language as a whole. Mainly these rules came out of Burrow and Bhattacharya.

Another way in which we were forced to "cheat" was in testing the model on pairwise mappings involving the special base. Since Burrow and Bhattacharya provide so few examples of suffixed forms of verbs, it was found that the number of forms for which an intensive form and a special base were both attested was too low to provide a trustworthy basis for running the algorithm. The four pairwise comparisons in question (deriving the intensive from the special base, deriving the special base from the intensive, deriving the imperative from the special base, and deriving the special base from the imperative) would be suspect due to the small sample size.

To increase the reliability of the data, then, the regular rules for special base formation provided by Burrow and Bhattacharya, as well as the listed exceptions, were used to synthesize a list of special bases for all the verbs in the data. It seems safe to assume that this list is accurate, as Burrow and Bhattacharya note no further exceptions to their generalizations, but it is still a move that ideally we would not want to make, because essentially the ability to predict unattested forms is what we're trying to prove. However, in the absence of any larger sample of data, it was deemed necessary in order to provide sufficiently large mappings for the learner.

#### 4.1.2. The learner

Before investigating the results of the computer modeling of Pengo verbal morphology, it will be instructive to explain the details of how the model works. The following will be a brief introduction to the model; for more detailed description, see Albright (2002), chapters 3 and 4, and Albright and Hayes (1999, 2002), among others.

The model identifies the most informative form in a paradigm by using each form as a candidate for the base, and attempting to derive each other form in the paradigm from it. The form which allows the most accurate reconstruction of the rest of the paradigm is the most informative. In order to do this, the model requires two things: a method for deriving one form in the paradigm from another, and a metric for assessing the accuracy of the derivations.

The implementation of the former takes the form of a “minimal generalization learner,” as described in Pinker and Prince (1988). It is trained on pairs of morphologically related forms, and attempts to learn the rules by which one form can be derived from another. It does this by induction, using a bottom-up process to make broader and broader generalizations based on the data available.

As an example, consider attempting to derive Pengo gerunds from imperatives. The model is presented with the following forms:

Table 20	
gerunds	
imperative	gerund
<i>uka</i>	<i>ukci</i>
<i>ura</i>	<i>urci</i>
<i>uca</i>	<i>uchi</i>
<i>proa</i>	<i>prohi</i>

Each row of the above table represents an ordered pair of forms in a particular morphological relation to one another. This relation can be characterized as a *structural change*, in this case suffixation, in a particular *context*. Formally, the change can be represented as a rule in the form  $A \rightarrow B$  and the context as  $/ C\_D$ . This yields a set of four word-specific rules for the above pairs, where # represents a word boundary:

- (3)  $a \rightarrow ci / uk\_ \#$   
 $a \rightarrow ci / ur\_ \#$

$a \rightarrow hi / uc\_ \#$   
 $a \rightarrow hi / pro\_ \#$

Word-specific rules essentially amount to memorization of existing forms, since they cannot be applied in any environment other than the word from which they were learned. But since the learner's task is to be able to apply its conclusions to novel forms, it seeks to generalize to more predictive rules by comparing pairs of forms with the same structural change. For instance, the first two rules above,  $a \rightarrow ci / uk\_ \#$  and  $a \rightarrow ci / ur\_ \#$  both have the same change,  $a \rightarrow ci$ , in different environments. The generalization algorithm compares the environments of these two rules and separates them into shared and non-shared portions. Specifically, it looks at the string of shared segments strictly adjacent to the change, and the phonological features shared by the first segment in which the two forms differ.

The generalization procedure attempts to retain as much shared information as possible across each generalization, giving the most specific rule that will cover all the input forms. This is why it is referred to as *minimal generalization*. Since the segments adjacent to the change  $a \rightarrow ci$  do not share any features other than being consonants, the algorithm will generate a new, generalized rule  $a \rightarrow ci / [-syllabic]\_ \#$ . If the strings in the environments for the two rules being compared are very similar, the generalized rule will be very specific, but iterating this generalization across the lexicon can produce more and more general rules.

When the algorithm encounters the change  $a \rightarrow hi / uc\_ \#$ , this change is not shared by either of the two word-specific rules above, so the algorithm sets up a new structural description for the new change. It cannot generalize to environments from the  $a \rightarrow ci / [-syllabic]\_ \#$  rule because the structural change is different. This word-specific rule will only be generalized when the algorithm encounters other rules with the change  $a \rightarrow hi$ , like  $a \rightarrow hi / pro\_ \#$ . From these two rules, a rule  $a \rightarrow hi / \_ \#$  is generalized. The environment for this rule overlaps significantly with the environment for the  $a \rightarrow ci / [-syllabic]\_ \#$  rule, so it can be seen that the algorithm sets up competing rules within the language.

#### 4.1.3. Evaluating rule systems

As can be seen above, the algorithm creates pairwise mappings, each an ordered pair of forms  $\langle X, Y \rangle$  where  $X$  is the input and  $Y$  is the output. The algorithm will learn a set of generalized rules for deriving each  $Y$  from each  $X$  for every such pair in the paradigm, and each set of rules is called a *subgrammar*. Once the learner has determined the generalized rules which can be used to derive the attested forms, it needs to have some way of determining which rules are better than others, since some rules will be in direct competition in particular environments. This is necessary in order to determine which of the possible rules should be used to generate novel forms.

The base selection model, then, contains an algorithm used to calculate the efficacy of the various putative rules. Rules are scored first on *reliability*. Reliability is defined as the ratio of the number of input forms the rule derives correctly (*hits*) over the number of forms it could potentially apply to (*scope*). For example, given the set of four pairs above, the  $a \rightarrow ci / [-syllabic]\_ \#$  rule could potentially apply to three; *uka*, *ura*, and

*uca*, because they meet the environment specified for the rule. Therefore the scope of this rule is 3. However, the rule only generates 2 hits, because it only describes the correct change for *uka* and *ura*. This rule's reliability given the forms above, then, is  $2/3 = 0.667$ .

In order to capture the intuition that high reliability based on a large sample is more trustworthy than high reliability over a small sample, reliability values are adjusted using confidence limit statistics as in Mikheev (1997). The reliability figure, which ranges from 0 to 1, is multiplied by a reliability parameter ranging from 0.5 to 1, depending on the size of the rule's scope. For the actual formulas used in this calculation, refer to Albright (2002) or Albright and Hayes (2002). The adjusted value is the rule's score for confidence.

The model uses confidence figures to generalize rules to novel forms. When a novel form is encountered, the algorithm compares it to all the existing rules to determine if it contains the proper environment for any of them. Each rule which applies to the environment in the input form is applied to produce a new output, in decreasing order of confidence, and each output is assigned a well-formedness score equal to the confidence score of the rule that produced it. Rules describing very productive processes will have very high confidence, and thus derive outputs with very high well-formedness scores, while very exceptional rules will have low confidence scores and thus derive outputs with low well-formedness scores.

If the model has competing rules that both apply in the environment in the input form, the model will derive competing output forms, each with a well-formedness score. The output with the highest score is assumed to be the output chosen in a forced choice task or judged best in an acceptability task. But since these well-formedness scores are gradient, they can also account for gradient acceptability or optionality.

#### 4.1.4. Base selection

In order to determine the most informative of the candidates for basehood, we need a metric for evaluating the accuracy of entire rule sets as derived by the learner. The way this is done is to test each learned subgrammar on its ability to reproduce the training set. Each subgrammar is used to derive outputs from all the inputs available in that subgrammar's particular mapping. For each input, the subgrammar selects the output that is derived with the highest confidence score. Then each of the derived outputs is compared to the real outputs to see if they match. The percentage of correct outputs for each subgrammar is that subgrammar's *accuracy*, not to be confused with the *reliability* scores which are assigned to each rule.

Subgrammars are also scored on three other categories; *average margin*, *average competitors*, and *average confidence*. To determine the *average margin*, each winning output is compared to the next best distinct output, and these margins are averaged over the whole subgrammar. This is useful in determining whether there are competing outputs that are generated with almost as much confidence as the winning outputs.

The *average competitors* score is the average number of competing outputs derived by the subgrammar for each winning output. The *average confidence* is the mean confidence of each winning output, which is equal to the confidence of the best rule in the grammar that derives that output.

In practice, these four metrics are all highly correlated with one another, but we will be most interested in what follows in the accuracy and average confidence.

Putting all this together, the algorithm is effectively considering each member of the paradigm as a candidate for base status, and constructing grammars that use each as an input to derive the remainder of the paradigm, using a series of pairwise mappings, and subgrammars for each mapping. A complete morphological grammar is a set of subgrammars for each possible pairwise mapping in the paradigm. The most informative base candidate is the form that permits the most accurate subgrammars.

## 4.2. Results

### 4.2.1. Relative Accuracy

The simulations were run on a total of three forms from the verbal paradigm: the imperative, the intensive, and the special base. This simplification to the full verbal paradigm was made in the interest of clarity, to avoid having an impenetrable maze of results. These three specific forms were chosen because they appear to represent the best candidates for basehood. The past tense, for instance, is obviously eliminated by the total neutralization of stem-final consonant voicing, and the motion base, infinitive, and gerund are expected to show near-identical numbers to the special base. The following table represents the results of the series of simulations run on the computer model:

↓input	output→	imperative	intensive	special base	mean
imperative	accuracy		0.738461	0.757281	0.747871
	avg. margin		0.201116	0.375807	0.288462
	avg. comp.		2	1.203883	1.601941
	avg. conf.		0.590428	0.766277	0.678352
intensive	accuracy	0.892307		0.923076	<b>0.907692</b>
	avg. margin	0.612923		0.678258	<b>0.645591</b>
	avg. comp.	0.446153		0.184615	<b>0.315384</b>
	avg. conf.	0.788959		0.774421	0.781690
special base	accuracy	0.932038	0.861538		0.896788
	avg. margin	0.737682	0.547709		0.642696
	avg. comp.	0.616504	0.430769		0.523637
	avg. conf.	0.872985	0.755894		<b>0.814439</b>

So what does all this mean? The three rows in this table represent three subgrammars for generating our simplified Pengo verbal paradigm. Examining the specifics of the data, the intensive scored the highest mean percentile scores for accuracy and average margin, and had the fewest average competitors, while the special base had the highest mean percentile score for confidence. The next subsections will compare the different subgrammars on their respective abilities to predict particular forms, as this seems to be a good way of judging their relative performances.

#### 4.2.2. Predicting the imperative

The four simulations involving the intensive had a sample size of only 66 verbs, limited by the number of verbs for which Burrow & Bhattacharya provide intensives, while the other two simulations had a sample size of 207, limited only by the number of verbs in the sample, because of the synthesized special bases. So comparing these results on level ground may be slightly misleading. In particular, the special base performed better than the intensive at predicting the imperative; 93% accuracy vs. 89%. This was unexpected, but is probably due to the differing sample sizes, and the small overall sample size. In absolute numbers, the special base failed to predict 14 of 207 imperatives, a much lower total than expected.

The tables in this section are taken straight from the raw results generated by the computer model, and so use the orthography developed for the computer simulations. Here capital *D*, *T*, and *N* represent retroflexes, capital *G* represents  $\eta$ , capital *A*, *E*, *I*, *O*, and *U* represent long vowels, the + symbol represents the final *-a* of the imperative form, and the # symbol represents the right word boundary in suffixed forms. “form1” is the input form, “form2” is the (incorrect) output, and “real” is the attested output. Column A represents the input string in the rule used to generate the output, column B the output string, P the left context of the rule, and Q the right context.

Table 22 lists the imperative forms that were incorrectly generated from intensive forms by the algorithm.

Table 22						
predicting the imperative from the intensive - errors						
form1	form2	real	A	B	P	Q
cubda#	-> cub+	cup+	by da#	-> + /		___
Enda#	-> En+	End+	by da#	-> + /		___
hlbda#	-> hlb+	hlp+	by da#	-> + /		___
hOgda#	-> hOg+	hOk+	by da#	-> + /		___
iDda#	-> iD+	iT+	by da#	-> + /		___
kaDda#	-> kaD+	kaT+	by da#	-> + /		___
kagda#	-> kag+	kak+	by da#	-> + /		___
koDda#	-> koD+	koT+	by da#	-> + /		___
krogda#	-> krog+	krok+	by da#	-> + /		___
kunda#	-> kun+	kund+	by da#	-> + /		___
nenda#	-> nen+	nend+	by da#	-> + /		___
progda#	-> prog+	prok+	by da#	-> + /		___
ujja#	-> uc+	uj+	by jja#	-> c+ /		___
Unda#	-> Un+	Und+	by da#	-> + /		___

Since the special base contains no information about the *-ka* intensives, it also should contain no information about the 27 verbs which are voiceless-final in the imperative, but take voiced suffixes in the special base *et al.* However, the special base was able to predict the imperative for 17 of these 27 verbs. The rules the computer model used to predict these forms exploited small islands of reliability, limited contexts in which a change shows consistency, possibly coincidentally.

For example, the rule *-jja* → *-ca* was successful for 10 out of 11 forms ending in *-jja* in the special base. Only one form ending in *-jja* had an intensive ending in *-ja*. Similarly, 5 out of 5 forms ending in *-dda* in the special base end in *-ta* in the intensive. Whether these generalizations would survive a larger sample size remains unclear, but given a sample of only 207 verbs, these alone account for a boost in overall reliability of approximately 5%, and probably a similar boost in average confidence. The effect of these islands of reliability on the overall score would be expected to be much smaller given a larger sample.

Note, however, that these generalizations may not be as coincidental as they appear. One of the lessons of this type of statistical analysis of phonological systems is that learners may be sensitive to small and seemingly coincidental regularities in the data, and form phonological generalizations based on them. The frequency of voicing alternations in *-dda* and *-jja* special bases in Pengo may in fact be an active phonological generalization, or the result of an analogical change. There is no way to tell without a considerably larger data set.

Deriving the imperative from the intensive resulted in a lower percentile score, but the results were still encouraging:

form1	form2	real	A	B	P	Q
hlpka#	-> hlk+	hlp+	by pka#	-> k+	/	___
hoTka#	-> hoT+	honD+	by ka#	-> +	/	___
kabga#	-> kag+	kak+	by bga#	-> g+	/	___
nEcka#	-> nEc+	nEnj+	by ka#	-> +	/ c	___
noTka#	-> noT+	nonD+	by ka#	-> +	/	___
pAbga#	-> pAk+	pAg+	by bga#	-> k+	/	___
popka#	-> pok+	pom+	by pka#	-> k+	/	___

The simulation failed where expected; on the *ne:nja~ne:cka* class of exceptions, on the four verbs exhibiting metathesis *and* voice-changing under suffixation, and on a verb *hi:pa~hi:pka* which ends in *p* and takes the *-ka* intensive, thus mimicking metathesis. This sums to 7 exceptions out of 66 verbs, but crucially the intensive suffered *no other failures* in predicting final consonant voicing. In every other case the final consonant voicing in the imperative was predictable from the intensive. For this reason I would expect that the reliability of the intensive would be greater given a larger sample size, although it is interesting to note that the most general *-ka* → ∅ rule was only successful at deriving the imperative in 10 out of the 18 cases in its scope.

#### 4.2.3. Predicting the intensive

In the first subgrammar, the imperative shows its crucial flaw as a base candidate and can predict only 74% of intensive forms. The principal problem is the lack of information about suffix voicing. Since suffix class membership is arbitrary, there is no way to predict whether a given verb will take the *-ba*, *-pa*, or *-ka* intensive. This problem is particularly evidenced by the average number of competitors, demonstrating that the

computer model was essentially always producing multiple potential outputs using competing rules, for any given intensive form.

predicting the intensive from the imperative - errors						
form1	form2	real	A	B	P	Q
ar+	-> arka#	arba#	by +	-> ka#	/	___
Do+	-> Doba#	Dopa#	by +	-> ba#	/	___
Du+	-> Duba#	DUba#	by +	-> ba#	/	___
hen+	-> henba#	henpa#	by +	-> ba#	/	___
hi+	-> hiba#	hIba#	by +	-> ba#	/	___
hlp+	-> hIppa#	hlpka#	by +	-> pa#	/	___
hon+	-> honba#	honpa#	by +	-> ba#	/	___
ka+	-> kaba#	kApa#	by +	-> ba#	/	___
kat+	-> katpa#	katka#	by +	-> pa#	/	___
nEnj+	-> nEnjba#	nEcka#	by +	-> ba#	/	___
pac+	-> pacpa#	packa#	by +	-> pa#	/ c	___
pom+	-> pomba#	popka#	by +	-> ba#	/	___
Roc+	-> Rocpa#	Rocka#	by +	-> pa#	/ c	___
tItr+	-> tItrka#	tItrpa#	by +	-> ka#	/	___
Trak+	-> Trabga#	Trapka#	by k+	-> bga#	/	___
TUN+	-> TUNba#	TUNpa#	by +	-> ba#	/	___
vec+	-> vecpa#	vecpa#	by +	-> ka#	/ c	___

The inability to predict *-ka* intensives is the critical failing of the special base as well. It failed to predict the correct intensive for 9 out of 66 forms, and every time the problem is the *-ka* intensive.

predicting the intensive from the special base - errors						
form1	form2	real	A	B	P	Q
arda#	-> arka#	arba#	by d	-> k	/	___ a#
hIbda#	-> hIbba#	hIpkka#	by d	-> b	/	___ a#
honDda#	-> honDba#	hoTka#	by d	-> b	/	___ a#
kadda#	-> kadba#	katka#	by d	-> b	/	___ a#
kunda#	-> kunba#	kutka#	by d	-> b	/	___ a#
nonDda#	-> nonDba#	noTka#	by d	-> b	/	___ a#
pomda#	-> pomba#	popka#	by d	-> b	/ m	___ a#
ujja#	-> ucka#	ujba#	by jj	-> ck	/	___ a#
Unda#	-> Unba#	Utka#	by d	-> b	/	___ a#

#### 4.2.4. Predicting the special base

Again the imperative is unable to predict suffix voicing. It is also unable to predict the irregular *mana ~ manj* type of special base, but the vast majority of exceptions involve the algorithm's widespread failure to predict whether a given form will take the *-t/-ta* or the *-d/-da* special base suffix. This leads to 50 incorrect predictions out of the 207 verbs in the sample:

Table 26

predicting the special base from the imperative - errors

form1	form2	real	A	B	P	Q
ar+	-> arta#	arda#	by +	-> ta#	/ r	___
cup+	-> cupta#	cubda#	by +	-> ta#	/	___
Di+	-> Dit#	DIt#	by +	-> t#	/	___
ec+	-> ecca#	ejja#	by +	-> ta#	/	___
En+	-> Enda#	Enta#	by +	-> da#	/	___
et+	-> etta#	edda#	by +	-> ta#	/	___
gil+	-> gilda#	gilta#	by +	-> da#	/	___
gUr+	-> gUrta#	gUrda#	by +	-> ta#	/ r	___
hAr+	-> hArta#	hArda#	by +	-> ta#	/ r	___
hen+	-> henda#	henta#	by +	-> da#	/	___
hlp+	-> hlpata#	hlpada#	by +	-> ta#	/	___
ho+	-> hot#	hOt#	by +	-> t#	/	___
hOc+	-> hOcca#	hOjja#	by +	-> ta#	/	___
hOk+	-> hOkta#	hOgda#	by +	-> ta#	/ k	___
hon+	-> honda#	honta#	by +	-> da#	/	___
in+	-> inda#	inj#	by +	-> da#	/	___
iT+	-> iTa#	iDda#	by +	-> ta#	/	___
je+	-> jet#	jEt#	by +	-> t#	/	___
jEc+	-> jEcca#	jEjja#	by +	-> ta#	/	___
jo+	-> jot#	jOt#	by +	-> t#	/ o	___
jOc+	-> jOcca#	jOjja#	by +	-> ta#	/	___
kak+	-> kakta#	kagda#	by +	-> ta#	/ k	___
kat+	-> katta#	kadda#	by +	-> ta#	/	___
kaT+	-> kaTta#	kaDda#	by +	-> ta#	/	___
kEr+	-> kErta#	kErda#	by +	-> ta#	/ r	___
ki+	-> kit#	kid#	by +	-> t#	/	___
kic+	-> kicca#	kijja#	by +	-> ta#	/	___
koT+	-> koTta#	koDda#	by +	-> ta#	/	___
kRo+	-> kRot#	kROt#	by +	-> t#	/ o	___
krok+	-> krokta#	krogda#	by +	-> ta#	/ k	___
kuR+	-> kuRda#	kuRta#	by +	-> da#	/	___
lAc+	-> lAcca#	lAjja#	by +	-> ta#	/	___
man+	-> manda#	manj#	by +	-> da#	/	___
mRi+	-> mRId#	mRit#	by i+	-> ld#	/	___
nil+	-> nilda#	nilta#	by +	-> da#	/	___
no+	-> not#	nOd#	by +	-> t#	/ o	___
o+	-> ot#	Od#	by +	-> t#	/	___
pac+	-> pacca#	pajja#	by +	-> ta#	/	___
plc+	-> plcca#	pljja#	by +	-> ca#	/ c	___
piR+	-> piRda#	piRta#	by +	-> da#	/	___
plt+	-> pltta#	pldda#	by +	-> ta#	/	___
prok+	-> prokta#	progda#	by +	-> ta#	/ k	___
pun+	-> punda#	punj#	by +	-> da#	/	___
Roc+	-> Rocca#	Rojja#	by +	-> ta#	/	___
ta+	-> tAt#	tat#	by a+	-> At#	/	___
TUN+	-> TUNda#	TUNta#	by +	-> da#	/	___
tUt+	-> tUtta#	tUdda#	by +	-> ta#	/	___
vac+	-> vacca#	vajja#	by +	-> ta#	/	___
ven+	-> venda#	venj#	by +	-> da#	/	___
vlt+	-> vltta#	vldda#	by +	-> ta#	/	___

The intensive is extremely accurate at predicting the special base. It fails to predict the idiosyncratic *-j* special bases, as did the imperative, but has no trouble recovering the voicing of forms with the *-ka* intensive.

Table 27  
 predicting the special base from the intensive - errors

form1	form2	real	A	B	P	Q
hlpka#	-> hlkta#	hlbda#	by pk	-> kt	/	___ a#
inba#	-> inda#	inj#	by b	-> d	/	___ a#
nEcka#	-> nEjja#	nEnja#	by ck	-> jj	/	___ a#
popka#	-> pokta#	pomda#	by pk	-> kt	/	___ a#
venba#	-> venda#	venj#	by b	-> d	/	___ a#

#### 4.2.5. Summary

Despite the fact that the special base turned out to be disproportionately good at predicting the imperative by exploiting islands of reliability, the intensive was overall the most reliable candidate for the base. It failed to predict some apparently systematic variations, but its accuracy across the paradigm is sufficiently high to regard these forms as memorized exceptions. Importantly, any other candidate for base would require more memorization in order to accurately derive all the other forms in the paradigm. This singles out the intensive as the most informative form in the paradigm.

The learning algorithm then selects this subgrammar as its grammar for the Pengo verbal lexicon, and will use it to generate novel forms for verbs outside the training set. The best way to test the theory that the intensive forms the base for the Pengo verbal paradigm would be to find native speakers of Pengo and test their ability to generalize to novel forms, using a *wug*-test. Unfortunately, in the absence of any available speakers, this research lies outside the scope of this paper.

## 5 Conclusion

My results for Pengo overall show that the intensive/frequentative form is the single surface base for the verbal paradigm. Firstly, because extant statistical patterns in the modern language suggest a process of analogical leveling around the intensive form at some point in the language's history, and secondly, because when the predictability relations among the forms in the paradigm are examined closely, the intensive is shown to be the form from which all the others may be most accurately derived. Thus any algorithm selecting surface bases while attempting to minimize overall memorization will select this form as the base.

Forms with voice-changing suffixes like the intensive appear superficially to be poor candidates for basehood because of the possibility for large-scale neutralizations in stem phonology, but close scrutiny of the data show that such neutralizations occur in sufficiently small numbers that they may be safely regarded as exceptions. This is due to a broad conspiracy of statistical generalizations within the Pengo verbal lexicon. These are the generalizations that provide evidence of the putative analogical change.

The misleading nature of the surface rule lists may be why Kenstowicz and Kisseberth overlooked the possibility that suffixed forms acted as the base. Since they were interested only in stem phonology, the erratic voicing of final consonants seemed like a red herring. Examining the learner's task, however, of producing novel forms of verbs while hearing only surface inflected forms, makes it clear how crucial this voicing alternation is.

The algorithm used above has no metric for measuring how exceptional exceptions are, but it seems plausible that memorization could be reduced by appealing to a few subregularities covering the forms whose paradigms are not predictable from the intensive form. For instance, the *-j* special bases, such as *venba ~ venj*, which also have irregular infinitives in *-jeN*, the nasal+obstruent cluster final verbs such as *kunda~kutka*, and the verbs that show both metathesis and voicing change, such as *kaka ~ kabga*, all form subclasses with several members each. Some memorization could be reduced if, instead of individually memorizing a rule for each form, one rule existed for each class.

Since I assume that at some point in the language's history prior to the analogical change I propose, the unsuffixed form was the base, it should not be surprising that most of these subregularities appear to be predictable from the unsuffixed form rather than the purported base, the intensive. This effect is a remnant of an older stage in the language's phonology.

Since the publication of Kenstowicz and Kisseberth (1977), it seems fair to say that most phonologists believe in the necessity of abstract underlying representations, and the single surface base hypothesis has fallen into disuse. Albright (2002) presents some results that challenge Kenstowicz and Kisseberth's conclusions regarding the single surface base hypothesis, and represents an attempt to revive the single surface base hypothesis to currency.

In light of this new evidence, it is important to continue looking for validation for the single surface base hypothesis. Since the Pengo language was proposed as the primary counterexample to this hypothesis, the above results should lend some serious credibility to this approach

## References

- Albright, Adam (2002). *The Identification of Bases in Morphological Paradigms*. UCLA dissertation.
- Albright and Hayes (1999). *An Automated Learner for Phonology and Morphology*. ms., UCLA. <http://www.linguistics.ucla.edu/people/hayes/learning/learner.pdf>
- Albright and Hayes (2002). "Modeling English Past Tense Intuitions with Minimal Generalization". In Maxwell, Michael (ed.) *Proceedings of the Sixth Meeting of the ACL Special Interest Group in Computational Phonology*. Philadelphia, July 2002. ACL.
- Burrow, T. and S. Bhattacharya (1970). *The Pengo Language: Grammar, Texts, and Vocabulary*. Oxford: Clarendon Press.
- Garrett and Blevins (in press). "Analogical Morphophonology" in *The nature of the word: Essays in honor of Paul Kiparsky*, ed. by Kristin Hanson and Sharon Inkelas Cambridge, Mass.: MIT Press.

- Inkelas, S. (1994). "The consequences of optimization for underspecification." NELS 25, 287–302. Available for download at: <http://roa.rutgers.edu/view.php4?roa=40>.
- Inkelas, S., O. Orgun, and C. Zoll (1997). "The implications of lexical exceptions for the nature of grammar." In I. Roca (Ed.), *Derivations and constraints in phonology*, pp. 393–418. Oxford: Clarendon.
- Kenstowicz and Kisseberth (1977). *Topics in Phonological Theory*. New York: Academic Press.
- McCarthy, John (2002). "Optimal Paradigms." Ms. ROA.
- Mikheev, Andrei (1997). "Automatic Rule Induction for Unknown-word Guessing". *Computational Linguistics* 23:405-423.
- Pinker, S. and A. Prince (1988). "On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition". *Cognition* 28:73-193.