Surface Base Selection in Pengo

Andrew Dowd

1. Introduction

My aim in this work is to defend what I’m calling the “Single Surface Base hypothesis.” This is the hypothesis that underlying representations (URs) are restricted 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.

In 1977, Kentowicz and Kisseberth published *Topics in Phonological Theory*, with a chapter entitled “The Problem of Underlying Representations.” In this chapter, they considered a number of hypotheses about the nature of URs, including a single surface base, which they called Hypothesis B’’. While this hypothesis has the advantage of being restrictive and concrete, Kenstowicz & Kisseberth argued against it on the grounds that phonological systems such as that of Pengo were inaccessible to such an analysis, because it appears that the same surface form in Pengo is not suitable as a base for all the lexical items of a particular class. Ultimately, they argued that the scope of variation in phonological systems made it necessary to postulate abstract URs for phonological paradigms.

I have a new analysis of the Pengo data that shows that, far from representing this sort of counterexample, Pengo is actually a prime example of how surface base selection works, and has an explanatory capacity beyond a model using abstract URs.

2. Pengo Phonology

Pengo phonology is fairly complex and unusual in some respects. 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*. For example, the verb *tig-* ‘to push,’ when followed by the intensive/frequentative suffix *–ba*, becomes *tibga.*

There are two other particular aspects of Pengo phonology that will be crucial. The first is that Pengo has strict regressive voicing assimilation in the obstruent clusters produced by suffixation. Stem-final obstruents are devoiced before voiceless-initial suffixes, and voiced before voiced obstruent-initial suffixes. For example, the verb *tog-* ‘to tread,’ when followed by the past tense suffix *–tan,* becomes *toktan*.

Also crucial is that suffix-initial voicing for a particular set of derivational verbal suffixes is not determined phonologically, but by membership in an arbitrary lexical class.

Table 1: *voice-changing suffixes*

	class A	class B
special base	*-t/-ta*	*-d/-da*
intensive/frequentative	*-pa*	*-ba*
motion base	*-ka*	*-ga*
gerund	*-ci/-hi/-si*	*-ji/-zi*
infinitive	*-teng*	*-deng*

Table 2: *interaction between voice-changing suffixes and assimilation*

imperative	gerund	intensive	special base
ara	arji	arba	arda
gaca	gachi	gacpa	gacca
jela	jelji	jelka	jelda
uja	ujji	ujba	ujja

© 2005 Andrew Dowd. Proceedings of the 24th West Coast Conference on Formal Linguistics, ed. John Alderete et al., 105-111. Somerville, MA: Cascadilla Proceedings Project.

If you take a look at table 1, you can see that if a verb is a member of "Class A," then the derivational suffixes it takes will begin with a voiceless obstruent, and if it is a member of "Class B," then it will take suffixes beginning in a voiced obstruent. This is as unusual as it sounds, but it's quite clear that it's actually a lexical idiosyncrasy.

This means that stem-final obstruent voicing is potentially neutralized in all cases under suffixation, being determined entirely by suffix class membership.

3. The distributional puzzle

In table 3 you can see the potential four way contrast. The imperative forms will show a contrast in the stem-final obstruent voicing, and it will be neutralized by the suffix voicing in the gerunds and other suffixed forms. In these forms, stem-final obstruent voicing is entirely determined by the voicing of the suffix-initial consonant.

Table 3: *potential four way contrast*

	voiced	voiceless
[+voi]obst root	anɖa ~ anɖji	
	paga ~ pagji	
	nu:za ~ nu:zzi	
[-voi]obst root	kata ~ kadji	uka ~ ukci
	ɽoca ~ ɽojji	uha ~ ussi
	ho:ka ~ ho:gji	preka ~ prekci

So for example look down the "voiced" column and you'll see *paga ~ pagji*, but *ho:ka ~ ho:gji*. Any rule that tried to predict stem-final suffix voicing from the gerund would fail to predict the alternation in the imperatives.

But the actual distribution in table 3 is different from the potential distribution. One of the boxes is empty. There are no verbs in the language that show a voiced stem-final consonant in the imperative and take the voiceless suffixes. Notice that these are forms that alternate under suffixation, and thus forms in which the imperative is not predictable from the suffixed forms. And the distribution of two known lexical idosyncracies seem to conspire to block this alternation from appearing.

This is a real puzzle under an account like Kenstowicz & Kisseberth's that involves abstract URs. Presumably there were once disagreeing forms in the lexicon, and speakers evidently failed to learn them. But why should they fail to learn these forms, if the necessary information was present in the UR? One possible answer is that speakers only have access to the information present in one surface form. In this case the suffixed form. Then this distribution would be the result of an analogical change that leveled the paradigm to the suffixed form. The imperative forms that disagreed in voicing with the suffixed forms disappeared and were replaced by agreeing forms, in this case, forms with a voiceless stem-final obstruent. For instance, you could imagine a verb *uga* with gerund *ukci*, whick would as a result of the change be leveled to the *uka/ukci* that we see in table 3. This is a fairly typical base effect.

But if this is true, why are there forms with voiceless stem-final obstruents that take voiced suffixes? The analogical change I proposed would seem to require that they also be leveled out, since they alternate under suffixation. To answer this question, we have to look at the suffixed forms a little more closely. Recall in table 1 that the intensive/frequentative suffix has a voiceless variant *–pa* and a voiced variant *–ba*. The distribution of voicing for these two suffixes follows the class lines; that is, if a verb takes the *–ba* suffix, it will also take the voiced gerund, infinitive, motion base, etc., and if it takes the *–pa* suffix, it will take all the other voiceless suffixes. However, not all verbs take one of these two intensive suffixes. There is a class of verbs that take an "irregular" invariant suffix *–ka*. This suffix has no voiced variant, but interestingly, is *restricted to verbs that elsewhere take the voiced suffixes*. This means that verbs taking the *–ka* intensive will always show a voiceless stem-final obstruent in the intensive form, and a voiced one in all the other suffixed forms. So if there is a correlation between the verbs with voiced suffixes that alternate under suffixation and the verbs that take the *–ka* intensive, we might have our solution. That is, none of the attested forms would disagree in voicing between the intensive/frequentative form and unsuffixed forms like the imperative. Table 4 shows the distribution when we examine the intensives.

Table 4: *distribution of suffixed forms*

	-ba	-pa	-ka
[+voi]obst root	yes	no	yes*
[-voi]obst root	yes*	yes	yes
son root	yes	yes	yes

Now, the single surface base model predicts that forms that alternate under suffixation will not be attested; that is, forms with voiceless stem-final obstruents that take the *–ba* intensive, forms with voiced stem-final obstruents that take the *–pa* intensive, and forms with voiced stem-final obstruents that take the *–ka* intensive. In this typology, all the other alternations allow suffix-final consonant voicing in the imperative to be predicted from the suffixed intensive form. Two of these boxes in table 4 are labeled "yes*." This means that forms with this combination of stem-final and suffix-initial voicing are attested, but only as what I am calling exceptions.

I now turn to these exceptions. First, we have the stems ending in voiced obstruents that take *–ka* in the intensive. I've identified six of these forms, which can be seen in table 5.

Table 5: *exceptions*

imperative	*intensive*
u:nda	u:tka
kunda	kutka
ne:nja	ne:cka
nonɖa	noʈka
raza	raska
honɖa	hoʈka

These forms, which are identified by Burrow & Bhattacharya as a class of exceptions, are roots that terminate in a nasal + obstruent cluster (or *m* from earlier *mb*, or *z*, possibly from an earlier cluster), and in addition to the regressive voicing assimilation under suffixation, the nasal is deleted. Outside of this class, there are no verbs in the language that alternate in final consonant voicing under suffixation of *–ka*.

The second class of exceptions is the verbs that take the voiced intensive suffix *–ba*, yet show a stem-final voiceless obstruent in the unsuffixed forms. I have identified four of these forms, which you can see in table 6.

Table 6: *exceptions*

imperative	intensive	gloss
kaka	*kabga*	'vomit'
ma:ka	*ma:bga*	'bake'
ɽa:ka	*ɽa:bga*	'offer worship'
ho:ka	*ho:bga*	'rub'

These four verbs end in a velar, and thus undergo metathesis when followed by a labial-initial suffix. While some forms that undergo metathesis still allow the unsuffixed forms to be predicted from the intensive, these four also undergo the voicing alternation. These metathesizing forms are known to be exceptional anyway; at least Burrow and Bhattacharya regard them as exceptional.

The upshot is, all of the forms this model predicts will be attested are, and of the three boxes we don't expect to be filled, one is empty, and two contain only exceptions with very specific subregularities. And if these forms are really exceptions, the I've shown that the base effect of leveling to the intensive form really exists.

4. The model

Identifying the intensive form as a base raises the question of what is a base, and what makes a base a base. As linguists, we can identify the bases of paradigms by inferring them from base effects such as analogical changes. That's what we just did in the previous section. 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.

Albright (2002) has provided a model for this process of UR discovery. 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, neutralizations.

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.

Let's look at an example of how this learner works. It involves slogging through some technical details briefly. Consider attempting to derive Pengo gerunds from imperatives. The model is presented with the forms in table 7.

Table 7: *imperatives/gerunds*

imperative	gerund
uka	*ukci*
ura	*urci*
uca	*uchi*
proa	*prohi*

Each row in table 7 represents an ordered pair <X,Y> of forms in a particular morphological relation to one another. This relation can be characterized as a *structural change*, in this case some sort of suffixation, in a particular *context*. Formally, the change can be represented as a rule in the form A _ 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:

1. *a _ ci / uk_#*
2. *a _ ci / ur_#*
3. *a _ hi / uc_#*
4. *a _ 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.

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 _ ci* in rules 1 and 2 do not share any features other than being consonants, the algorithm will generate a new, generalized rule *a _ 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 in 3, 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 _ 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 _ hi*, like *a _ hi* / *pro_#*. From these two rules, a rule *a _ hi* / _# is generalized.

This builds rules in a familiar manner; what's important is that the generalizations are always minimal, so the rules are always the most specific rules that will cover all the environments for a particular structural change. So from the above four rules, the algorithm produces two new generalized rules:

5. a _ ci / [-syllabic]_ #.
6. a _ hi / _#

The environments for these two rules overlap significantly, as well as overlapping with the environments for the single word rules, so it can be seen that the algorithm creates rules in competition in particular environments. This leads us to the next step, evaluating rules against each other to determine which rule wins each competition.

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 four pairs from table 7, rule 5, the *a _ 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.

The model doesn't only rate rules on reliability, however. In order to capture the intuition that high reliability based on a large sample is more trustworthy than high reliability over a small sample, the model also produces a value adjusted by confidence limit statistics as in Mikheev (1997). This means the reliability is multiplied by some value between .5 and 1, depending on the size of the rule's scope. The actual math for obtaining this value is explained in Albright (2002); it's too complicated to go into here. You can see how important this confidence adjustment is, however, when you consider that the individual word rules will all have 100% accuracy.

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. 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, paradigm gaps, or optionality.

To summarize, we start with a series of pairwise mappings, each an ordered pair of forms <X,Y> 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*. So, intuitively, the model is building up rules by generalization, and these rules are in competition, and the subgrammars built out of these rules will be in competition as well, because each subgrammar bears on each base candidate's fitness. Competition between subgrammars is essentially competition between base candidates.

So we need a metric for evaluating the accuracy of these subgrammars. The way this is done is to test each one 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*.

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.

5. Simulation results

I ran simulations of the Pengo data on Albright's learner. I compared three forms from the verbal paradigm: the imperative, the intensive, and the special base. This simplified paradigm is for the purpose of clarity, because it's a maze of data as it is.

Since there's a subgrammar for each input/output pair, and three different forms, and each has to be tested as a potential input and output, I tested six different subgrammars on their ability to replicate the training set. Each potential base candidate, then, has two subgrammars, one for each output form. These were averaged to find mean scores for each base candidate on the four different metrics of reliability.

The results are shown in table 8. looking down the mean column, the figures in bold are the highest mean figures for each particular metric of evaluation. Now, remember that the crucial number is accuracy.

Table 8: *simulation results*

_input	output_	imperative	intensive	special base	mean
imperative	accuracy		0.738461	0.757281	0.747871
	avg. margin		0.201116	0.375807	0.288462
	avg. competitors		2	1.203883	1.601941
	avg. confidence		0.590428	0.766277	0.678352
intensive	accuracy	0.892307		0.923076	**0.907692**
	avg. margin	0.612923		0.678258	**0.645591**
	avg. competitors	0.446153		0.184615	**0.315384**
	avg. confidence	0.788959		0.774421	0.781690
special base	accuracy	0.932038	0.861538		0.896788
	avg. margin	0.737682	0.547709		0.642696
	avg. competitors	0.616504	0.430769		0.523637
	avg. confidence	0.872985	0.755894		**0.814439**

The imperative had a mean accuracy around 75%, the special base had a mean accuracy of 89.7%, and the intensive had a mean accuracy of 90.8%.

The upshot is that overall, the candidate for base with the best mean accuracy was the intensive form. This means that the learner will select this form as the base for the paradigm, and use it to generalize to novel forms. This is consistent with the evidence from section 2 about the analogical change leveling to this form. Pengo learners are constructing the verbal paradigm using, in the general case, only the information present in the intensive/frequentative form.

6. Conclusion

Kenstowicz & Kisseberth (1977) argue that no single surface form can be the base fo the Pengo verbal paradigm. I've shown that one is; it's the intensive/frequentative form. I've also shown that this is the form Albright's model predicts will be selected by learners as the base.

Ever since 1977, most of phonology has accepted Kenstowicz and Kisseberth's conclusions regarding the necessity of abstract URs, based on examples like Pengo. I've shown that Pengo isn't just accessible to a single surface base analysis, it requires one. A model using abstract URs would be able to produce the Pengo phonological system, but it would fail to explain the analogical change. This model doesn't just account for the synchronic phonology, it *explains* the analogical change and the resultant statistical distribution.

References

Albright, Adam (2002). *The Identification of Bases in Morphological Paradigms*. UCLA dissertation.

Albright, Adam, and Bruce Hayes (1999). *An Automated Learner for Phonology and Morphology*. ms., UCLA. http://www.linguistics.ucla.edu/people/hayes/learning/learner.pdf

Albright, Adam, and Bruce 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, Andrew, and Juliette 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.

Kenstowicz, Michael, and Charles Kisseberth (1977). *Topics in Phonological Theory*. New York: Academic Press.

Mikheev, Andrei (1997). "Automatic Rule Induction for Unknown-word Guessing". *Computational Linguistics* 23:405-423.

Pinker, Steven, and Alan Prince (1988). "On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition". *Cognition* 28:73-193.

Proceedings of the 24th West Coast Conference on Formal Linguistics

edited by John Alderete, Chung-hye Han, and Alexei Kochetov

Cascadilla Proceedings Project   Somerville, MA   2005

Copyright information

ISBN 1-57473-407-5 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, e-mail: sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Dowd, Andrew. 2005. Surface Base Selection in Pengo. In *Proceedings of the 24th West Coast Conference on Formal Linguistics,* ed. John Alderete et al., 105-111. Somerville, MA: Cascadilla Proceedings Project.

or:

Dowd, Andrew. 2005. Surface Base Selection in Pengo. In *Proceedings of the 24th West Coast Conference on Formal Linguistics,* ed. John Alderete et al., 105-111. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #1212.