From Sequence Frequencies to Conditions in Bantu Vowel Harmony: Building a grammar from the ground up*

Diana Archangeli[a], Jeff Mielke[b] and Douglas Pulleyblank[c]
University of Arizona[a], University of Ottawa[b], University of British Columbia[c]

SUMMARY

We explore here the hypothesis that phonological grammars are emergent, formed by general principles that may involve little to no role for language-specific principles. Our basic proposal is that grammars develop from the identification of patterns of similarity, the calculation of frequencies of patterns of co-occurrence, and the development of generalized symbolic systems based on frequency data. We investigate six Bantu languages, all of which exhibit a canonical asymmetric height harmony pattern. Based on sizeable on-line databases, we examine the frequency of all possible vowel sequences in the six languages, using the frequency data to develop a nascent grammar for height harmony in each language. Our proposal is for a type of unsupervised learning, and we discuss various ways of establishing that the learning algorithm has converged on the correct grammar.

RÉSUMÉ

Nous explorons ici l'hypothèse que les grammaires phonologiques sont émergentes, formées par des principes généraux qui pourraient ne laisser que peu ou pas de rôle à des principes spécifiques à des langues données. Notre proposition de base est que les grammaires se développent à partir d'identification de schémas de similitude, de calcul des fréquences de co-occurrences, et du développement de systèmes symboliques généralisés basés sur des données de fréquence. Nous étudions six langues bantoues qui possèdent toutes un système d'harmonie vocalique de hauteur qui est canonique et asymétrique. En nous basant sur de grandes bases de données en ligne, nous examinons les fréquences de toutes les séquences de voyelles possibles dans les six langues, en utilisant des informations sur leurs fréquences afin de développer une grammaire initiale pour l'harmonie de hauteur dans chaque langue. Nous proposons un type d'apprentissage non supervisé, et nous discutons des différentes manières de vérifier que l'algorithme d'apprentissage en arrive à la grammaire correcte.

* This paper was supported in part by a Standard Research Grant from the Social Sciences and Humanities Research Council of Canada to Douglas Pulleyblank.

McGill Working Papers in Linguistics, Volume 22.1, Winter 2012

1 INTRODUCTION

Creating a grammar for a language involves creating an abstract, symbolic representation of the language itself. We explore here the hypothesis that creating the abstract symbolic representations to express language patterns is driven by deviations from the expected, in particular, deviations from the expected frequency of occurrence of patterns. The more robust the deviation, the more likely a pattern is to be symbolically represented, and so the more likely the pattern is to be encoded in the grammar.

Since the advent of generative grammar and works such as Chomsky and Halle (1968), analyses of phonological systems have assumed and relied on a Universal Grammar rich with formal components which constrain the set of possible individual grammars, and so constrain possible analyses. For example, under the Universal Grammar model, Optimality Theory (Prince and Smolensky 1993; McCarthy and Prince 1993, 1994; Archangeli and Langendoen 1997, the selections in McCarthy 2003, etc.) assumes a universal set of distinctive features, a universal set of constraints, and universal mechanisms for creating and evaluating a candidate set for each input form. Only the ranking of the universal constraints and the nature of input forms is determined by exposure to data; all else is provided by Universal Grammar.

To establish the need for Universal Grammar, we need to show that our best attempts to develop a model eschewing innate linguistic principles, an Emergent Grammar, is not viable. In other words, before assuming some property of a language is the result of an innate structure that is specifically linguistic, it is necessary to eliminate the possibility that that properties of language emerge from analysis based on (nonlinguistic) cognitive abilities. If it turns out that models of Emergent Grammar fail, then a structured Universal Grammar is indeed necessary. We examine this issue here.

Against this backdrop, we explore the question: How far can we get in deriving phonological patterns without assuming a rich Universal Grammar? In place of an innate linguistic component ("Universal Grammar", UG), we assume a minimal and plausible set of specific innate cognitive abilities, such as the ability to attend to and generalize from distributional frequencies (Pierrehumbert, 1993; Frisch et al., 2004). Linguistic analysis takes place in accord with these abilities allowing the abstract, symbolic grammatical structure to emerge from experience ("Emergent Grammar", EG): featural categories emerge, constraints emerge, constraint rankings emerge, and all are promoted into the abstract, symbolic grammar. (See Mohanan et al. 2010 for discussion of the general Emergence research program that we are exploring.)

With this very different set of starting assumptions, we are investigating the question of "How far can we get?" through a series of studies of vowel harmony; here, we report on a study of height harmony in Bantu, a naturally occurring pattern of reasonable complexity (see also Archangeli et al. 2011). Our hypothesis is that the grammatical representation of vowel harmony is derived from the way that generalizations are formed based on general learning principles interacting with language-specific data.

At the outset we make the simplifying assumption that the learner can identify objects (including sounds) as "the same" and so can identify the segments of the language, and has the segments largely in place prior to developing a lexicon (Werker and Tees 1983; Gómez and Gerken 1999; Maye et al. 2002; Richtsmeier et al. 2009, etc.). Additionally, we accept the arguments that language learners are sensitive to frequency distributions across the lexicon (Frisch 1996; Maye 2000; Gómez 2002; Pierrehumbert 2003; LaCross 2011; etc., but see also Onnis et al. 2003, 2004; Bonatti et al. 2005). Finally, we assume that humans create symbolic systems of representation (Deacon 1997).

(1) **Basic assumptions about human capabilities**

a. identify similarity
b. calculate frequencies
c. create symbolic systems

We do not assume that features or rules/constraints are a pre-established component of UG; rather, features emerge; (Mielke, 2005, 2008), and so do rules/constraints (Pulleyblank 2006; Mohanan et al. 2010, etc.).

We show that both features and conditions on the distribution of features can be determined by making use of information provided by co-occurrence frequencies for vowels with or without intervening consonants. Our conclusion is that the Bantu vowel harmony data are consistent with an Emergent Grammar model, hence fail to support a Universal Grammar model.

We begin with a sketch of the Bantu height harmony patterns and the nature of Emergent Grammar construction, followed by a description of the data used, section 2. This is followed in section 3 by a presentation of our methods, from raw observations to grammar construction, using Ciyao as an example. Sections 4-5 lay out the results from examining the languages in our sample. We consider implications of these results in the closing section.

2 DATA AND PATTERNS

In this section, we present the Bantu height harmony pattern and the analytic hypothesis that together form the basis of our study.

2.1 BANTU HEIGHT HARMONY

Bantu height harmony is a well-known and common vowel pattern in Bantu languages, though details vary and not all Bantu languages exhibit the pattern (Hyman 1999). Five-vowel systems frequently exhibit the widely attested “asymmetric" pattern, illustrated in (2) with data from Ciyao (Ngunga 2000). Height harmony is responsible for the alternation between high and mid vowels in the applicative and reversive suffixes.

(2) **Bantu height harmony in Ciyao**

a.	*‘-il’ applicative*			b.	*‘-ul’ reversive*		
	dim-	dim-il-	*cultivate*		siv-	siw-ul-	*close/open up*
	wut-	wut-il-	*pull*		uuv-	uuw-ul-	*hide/reveal*
	saam-	saam-il-	*move*		mat-	mat-ul-	*adhere/peel off*
	pet-	pet-el-	*ornament*		sweek-	sweek-ul-	*insert/pull out*
	soom-	soom-el-	*read/study*		som-	som-ol-	*pierce/extract*

As seen in (2a), the high front unrounded vowel /i/ is found after high and low vowels but disallowed in a syllable following a mid vowel, while the mid front vowel /e/ is only allowed in the mid-vowel context. The pattern is slightly different when the second vowel is back and round, as illustrated in (2b): the high back rounded vowel /u/ is found not only after high and low vowels but also after front mid vowels; /u/ is disallowed only in a syllable following the mid back rounded vowel /o/, with /o/ found instead.

2.2 CONSTRUCTING AN EMERGENT GRAMMAR

The hypothesis to be tested here is whether it is possible to construct a grammar of Bantu height harmony with minimal to no appeal to an innate linguistic endowment. The most robust relations are used to construct the grammar.[1] Since we are concerned with constructing a grammar for a vowel harmony pattern that is conditioned by stem vowels and observed in suffixes, the constructed grammar will be composed of left-to-right implicational statements. These implicational statements are derived from (non-directional) co-occurrence patterns: for example, if a given $V_1 \ldots V_2$ sequence is significantly underrepresented, then V_1, as the last vowel in a stem, implies that a suffix will not contain V_2.

[1] We take up the challenging issue of how to define “robust” in section 3.4.

(3) **The Emergent Grammar construction hypothesis**

1. calculate the frequency of occurrence of sounds S in some environment E
2. rank implicational relations of S in E based on robustness
3. project grammar out of the most robust implicational relations

Under this hypothesis, the learner has no pre-knowledge of the patterns to be identified, and so is tracking myriad sound-environment frequencies. Because we are looking at Bantu height harmony, we focus on what can be learned if the learner pays attention to $V_1 \ldots V_2$ frequencies. We propose that the learner expresses $V_1 \ldots V_2$ sequences as implicational relations between the two vowels, and that frequency determines the robustness of a particular relation in the language.

Thus, acquiring the grammar of Bantu height harmony involves calculating $V_1 \ldots V_2$ co-occurrence frequencies, represented as implicational relations between vowel pairs. The co-occurrence frequencies provide a ranking for the implicational relations; the learner constructs a grammar from the most robust of these implicational relations.

Before turning to our methods for testing this hypothesis, we justify the properties of our language sample.

2.3 THE LANGUAGE SAMPLE

To test the Emergent Grammar hypothesis, we began by identifying four critical properties each language must meet in order to be selected.

(4) **Criteria for test languages**

a. a relatively small vowel system: keep the number of calculations manageable, reduce confounds
b. the ‘same’ vowel system, again to reduce confounds
c. the ‘same’ phonological pattern, again to reduce confounds
d. accessible data, so the studies could be carried out

The choice of height harmony in Bantu as the test case allowed the satisfaction of all criteria. On the first and second points, the smallest vowel system consistently found in Bantu languages with height harmony is the standard five-vowel system, { i e a o u }. We limited test languages to those described as having only these vowels. On the third point, all test languages are from the set of languages described as having the asymmetric Bantu height harmony (Hyman 1999) and illustrated in (2). Finally, we identified an online data source with searchable word lists, the Comparative Bantu OnLine Dictionary (CBOLD, `http://www.cbold.ish-lyon.cnrs.fr/`).

(5) **Test languages**

- six Bantu test languages: Bukusu, Chichewa, Jita, Ikalanga, Kiga-Nkore, Ciyao (all from CBOLD)
- all languages have five vowels: { i e a o u }
- all languages described as having the height harmony system illustrated above (Hyman, 1999)

Details on these languages are given in (6) (see also Archangeli et al. 2011).

(6) **The test cases** [2]

- Bukusu: Niger-Congo, Bantu; Kenya; E31C in Guthrie (1967-71); CBOLD: KWL (1998); morphological alternations demonstrated in Mutonyi (2000); 889 verbs, 1412 nouns, 2658 lexical items in total (.fm)
- Chichewa: Niger-Congo, Bantu; Malawi; N31B in Guthrie (1967-71); CBOLD: Mtenje (2001); morphological alternations demonstrated in CBOLD; 2010 verbs, 2172 nouns, 4992 lexical items in total (.txt)
- Ciyao: Niger-Congo, Bantu; Malawi, Mozambique & Tanzania; P21 in Guthrie (1967-71); CBOLD: Ngunga (2001); morphological alternations demonstrated in Ngunga (2000); 2616 verbs, 3021 nouns, 6717 lexical items in total (.txt)
- Ikalanga: Niger-Congo, Bantu; Botswana & Zimbabwe; S16 in Guthrie (1967-71); CBOLD: Mathangwane (1994); morphological alternations demonstrated in Mathangwane (1999) (though asymmetric nature of harmony not discussed there); 1149 verbs, 1640 nouns, 2899 lexical items in total (.fm)
- Jita: Niger-Congo, Bantu; Tanzania; J25 in Guthrie (1967-71); CBOLD: Downing (1999); morphological alternations demonstrated in Downing (1999) (though asymmetric nature of harmony not discussed there); 870 verbs, 1010 nouns, 1925 lexical items in total (.txt)
- Nkore-Kiga: Niger-Congo, Bantu; Kenya; J13 in Guthrie (1967-71); CBOLD: Taylor (1959); morphological alternations demonstrated in Taylor (1985); 3007 verbs, 3847 nouns, 6913 lexical items in total (.txt)

An immediate question is whether these wordlists have sufficient data for the task at hand. In a review of the literature, Barrett (1995) suggests that by two and a half years of age, the lexicon contains around 500 words on average. Verbs are much less frequent than nouns, perhaps about a third of the lexicon if numbers are largely comparable to adult proportions. If a child achieves a lexicon of something like 14,000 words by age six (Clark, 1995), this might mean upwards of 4,500 verbs. The figure below shows the lexicon sizes and the proportion of verbs in each of the test languages. In all cases, the number of words in the list is equivalent to what could be expected in a child between 2 and 6 years old.

(7) **Are the CBOLD wordlists adequate?**

language	verbs	total	proportion of verbs
Bukusu	889	2658	.33
Chichewa	2010	4992	.40
Ciyao	2616	6717	.39
Ikalanga	1149	2899	.40
Jita	870	1925	.45
Nkore-Kiga	3007	6913	.43

The datasets for the test languages in CBOLD are comparable in size to the size of the vocabulary of a young child; it is plausible that generalizations made about these datasets might also be made about the developing child's lexicon.

[2] CBOLD includes references as indicated here, but omits complete bibliographical information; consequently our references are similarly incomplete. For each test language, there is a .txt data set version (.txt) and a FileMakerPro™ version (.fm) of the data set in CBOLD. We used the .txt version unless the data was more accessible through the .fm version.

3 METHODS

In this section, we lay out our strategy for projecting a grammar from data, using Ciyao to illustrate. Our starting point is to calculate the frequency with which different $V_1 \ldots V_2$ sequences occur in each language, and to rank the sequences based on those frequencies. Interpreting the sequences as implicational statements allows for the construction of further implications, by merging either antecedent or consequent. It also allows for eliminating implications which are subsumed by others. The end result for languages with height harmony is a set of robust implicational relations that express the harmonic pattern.

3.1 THE OBSERVATIONS

In selecting among the CBOLD languages, data accessibility was a primary concern. Many of the CBOLD wordlists included the category columns in (8) among others. This type of layout is advantageous for several reasons. First, the vowel sequences are readily accessible when the tone is not part of the typographic representation of each vowel, and the part of speech (POS) is given so that verbs (which show harmony) can be examined separately from nouns (which do not show as strong a harmony pattern; see Archangeli et al. 2011).

(8) **CBOLD word list (Ciyao verbs)**

Stem	**Tone**	**POS**	**Class**	**Gloss**
-n'weesula	HHLL	verb	15	abrase the skin
-loongana	HHLL	verb	15	accompany; go together
-soonjela	HHLL	verb	15	accuse
-pokolanya	HHLL	verb	15	arbitrate
-tumika	HHL	verb	15	act as a servant
-paambika	HHLL	verb	15	add to a load
-paambicila	HHLLL	verb	15	add to; increase (quantity)
-oonjecesya	LLHLL	verb	15	add to; increase (number)
-mta	HH	verb	15	adhere (as a swarm of bees)
-maambatila	HHLLL	verb	15	adhere; stick
-nyaambatila	HHLLL	verb	15	adhere, stick to
etc.				

From such lists of verbs, we identified relevant $V_1 \ldots V_2$ sequences in the following way. First, we limited our domain of study to (verb) stems because that is the strongest domain for height harmony in Bantu: Height harmony does not cross the prefix-stem boundary.[3] The prefix-stem boundary is a well-established boundary in Bantu phonology, and is easily identifiable due to the pervasive system of class prefixes. Furthermore, verbs have syntactic, semantic, morphological, and phonological properties that distinguish them from nouns, so there are many clues to support the hypothesis of two categories of word types.

Second, in making observations about $V_1 \ldots V_2$ sequences, we treated long vowels as single vowels. Thus, in a stem like *-n'weesula* 'abrase the skin', there are three vowels participating in $V_1 \ldots V_2$ sequences, *e...u...a*, not four (i.e., not *e...e...u...a*).

Finally, in our observations, we included all vowels of the stem, including the Final Vowel (the vowel [a] that typically occurs at the end of Bantu verbs). Note that "V_1" and "V_2" do not refer to absolute positions; for any two vowels in sequence, "V_1" refers to the first vowel in the sequence and "V_2" to the second vowel in the sequence.

[3] In Archangeli et al. (2011), we demonstrate that the harmony pattern does extend somewhat to nouns in some of the test languages. An informal characterization is that, within a given language, the height harmony pattern is pervasive among verbs and is a tendency of some degree in nouns.

Given these restrictions, items in the word lists (8) result in the observations in (9), paired with the source stem for convenience.

(9) **Sample observations**

Source stem	**Observations**
-n'weesula	e ... u, u ... a
-loongana	o ... a, a ... a
-soonjela	o ... e, e ... a
-pokolanya	o ... o, o ... a, a ... a
-tumika	u ... i, i ... a
-paambika	a ... i, i ... a
-paambicila	a ... i, i ... i, i ... a
-oonjecesya	o ... e, e ... e, e ... a
-mata	a ... a
-maambatila	a ... a, a ... i, i ... a
-nyaambatila	a ... a, a ... i, i ... a

Frequencies of $V_1 \ldots V_2$, then, are based on the number of times a particular sequence is found. For instance, in (9), the sequence *a...a* occurs 5 times while the sequence *a...e* never occurs. As an illustration, the counts for the tiny sub-lexicon in (9) are given in (10). "T" refers to the total in each row/column. Note that the grand total (26) refers to the total number of $V_1 \ldots V_2$ sequences, not to the total number of words considered.

(10) **$V_1 \ldots V_2$ frequencies for data in (9)**

		V2					
		i	**e**	**a**	**o**	**u**	**T**
V1	**i**	1	0	5	0	0	6
	e	0	1	2	0	1	4
	a	4	0	5	0	0	9
	o	0	2	2	1	0	5
	u	1	0	1	0	0	2
	T	6	3	15	1	1	26

The figures in (10) are comparable to an early stage in the linguist's data collection; they might also be comparable to a stage in acquisition when the lexical items learned add up to 26 $V_1 \ldots V_2$ sequences. As additional lexical items are added (either by linguist or by learner), the numbers in cells change to reflect the newly acquired data.[4]

Using the method described above for identifying relevant $V_1 \ldots V_2$ sequences, we count the number of occurrences of the 25 possible $V_1 \ldots V_2$ sequences in each language. Values for Ciyao verbs are shown in (11). Because verb stems typically contain two or more syllables, the total number of $V_1 \ldots V_2$ sequences, 5,758, is greater than the total number of verb stems, 2,583.

[4] Under our model, a grammar is projected at any point once (nascent) lexical items are acquired. As numbers change due to the acquisition of new items, the projected grammar may also change. In this way, the grammar is a grammar for a particular lexicon; the more items found in the lexicon, the more stable the grammar is because it takes so many more lexical items of a particular pattern to result in a sufficiently robust change in the distributions for the grammar to be revised.

(11) **Raw frequencies, Ciyao verbs**

		V2					
		i	**e**	**a**	**o**	**u**	**T**
	i	361	0	815	8	93	1277
	e	3	269	395	0	111	778
V1	**a**	271	6	933	4	241	1455
	o	10	117	360	355	0	842
	u	183	3	793	4	423	1406
	T	828	395	3296	371	868	5758

3.2 INTERPRETING THE NUMBERS

The next step is determining how to understand these numbers, because a given value changes its meaning depending on the context. For example, in (11) there are 360 occurrences of *o...a* and almost the same number of occurrences of *o...o*, 355. Yet these two numbers must be interpreted very differently. The 355 *o...o* examples constitute a vast *over*-representation of this sequence since there are only 371 $V_1 \ldots V_2$ sequences where the second vowel is [o] – thus, only 16 items of this set show a vowel other than [o] as V_1.

By contrast, *o...a* is *under*-represented with 360 examples. There are over 3,000 items with [a] as V_2. If the five possibilities for V_1 were evenly distributed, we would expect 600 items with the sequence *o...a*, instead of the observed 360.

In short, whether a sequence is under- or overrepresented depends on not only how many times the sequence appears but also how many times each vowel is observed in the appropriate position. What is needed is to determine how many observations are expected for each sequence, in order to determine whether the observed value is surprising or not. The Observed/Expected ratio has traditionally been used for quantifying under- and overrepresentation of co-occurrence patterns in the lexicon (Pierrehumbert 1993; Frisch et al. 2004).

Expected frequencies are based on the number of occurrences of V_1 and V_2, divided by the total number of occurrences:

$$V_1 \ldots V_2\text{: } \frac{\text{V1 freq} \times \text{V2 freq}}{\text{total}}$$

The expected frequencies for the two sequences just considered are calculated below:

$$\text{o...o: } \frac{842 \times 371}{5758} = 54$$

$$\text{o...a: } \frac{842 \times 3296}{5758} = 482$$

The expected values for each of the 25 $V_1 \ldots V_2$ sequences are given in (12).

(12) **Expected frequencies, Ciyao verbs**

		V2					
		i	**e**	**a**	**o**	**u**	**T**
	i	184	88	731	82	193	1277
	e	112	53	445	50	117	778
V1	**a**	209	100	833	94	219	1455
	o	121	58	482	54	127	842
	u	202	96	805	91	212	1406
	T	828	395	3296	371	868	5758

Finally, the observed and expected values are represented as a ratio, observed/expected. When the observed value is equivalent to the expected value, this ratio is 1.0. Values greater than one indicate that there are more observed sequences than were expected while values less than one show that there were fewer observations than were expected. The Ciyao observed/expected ratios are shown in (13a), and the interpretation of the magnitude is given in (13b).[5]

(13) **a. Observed/Expected ratios, Ciyao verbs**

		V2					
		i	**e**	**a**	**o**	**u**	**T**
	i	1.97	0.00	1.11	0.10	**0.48**	1277
	e	0.03	5.04	0.89	0.00	0.95	778
V1	**a**	1.30	0.06	1.12	0.04	1.10	1455
	o	0.08	2.03	0.75	6.54	0.00	842
	u	0.91	0.03	0.99	0.04	**2.00**	1406
	T	828	395	3296	371	212	5758

b. **Magnitude of O/E values** (examples are bolded in (13a))
0.5: half as many as expected, e.g. *i...u*
2.0: twice as many as expected, e.g. *u....u*

Again we have an interpretation issue, because the numerical values for underrepresentation are all found between 0 and 1 while the numerical values for overrepresentation are found between 1 and infinity. To put the values on an intuitively comparable scale, we convert the observed/expected ratios to $\log_2$ values.

$$\text{Observed/Expected log ratio: } \log_2 \frac{O}{E}$$

Underrepresented values are less than 0 while overrepresented values are greater than 0; magnitudes are comparable: 1 means that the sequence is overrepresented by a factor of 2 while –1 means it is underrepresented by a factor of 2, illustrated again by the bolded $\log_2$ values for *i...u* and *u...u*.

(14) **a. Observed/Expected $\log_2$ ratios**

		V2					
		i	**e**	**a**	**o**	**u**	**T**
	i	0.98	$-\infty$	0.16	–3.36	**–1.05**	1277
	e	–5.22	2.33	–0.17	$-\infty$	–0.08	778
V1	**a**	0.37	–4.06	0.16	–4.55	0.14	1455
	o	–3.60	1.02	–0.42	2.71	$-\infty$	842
	u	–0.14	–5.01	–0.02	–4.50	**1.00**	1406
	T	828	395	3296	371	212	5758

b. **Magnitude of $\log_2$ O/E values (examples in bold above)**
–1: half as many as expected, e.g. *i...u*
+1: twice as many as expected, e.g. *u....u*

[5] Totals are left in at the ends of rows and bottoms of columns to help keep perspective on the number of items on which these figures are based. They are total counts of sequence types, not totals created by adding rows or columns. Thus, there are 1277 items with [i] as V_1 and 828 items with [i] as V_2, etc.

We are now in position to determine which of the $V_1 \ldots V_2$ sequences are the most robust in each language: we can achieve this by ranking the sequences by the absolute value of their $\log_2$ values. To illustrate, the top twelve most robust Ciyao observations are given in (15).

(15) **Ciyao top twelve by magnitude**

	sequence	a. $\log_2$ values	b. $\lvert\log_2\rvert$ values
1.	*i…e	$-\infty$	∞
2.	*e…o	$-\infty$	∞
3.	*o…u	$-\infty$	∞
4.	*e…i	-5.2207987282	5.2207987282
5.	*u…e	-5.00677514785	5.00677514785
6.	*a…o	-4.55072674085	4.55072674085
7.	*u…o	-4.50130418213	4.50130418213
8.	*a…e	-4.05619770658	4.05619770658
9.	*o…i	-3.59788321211	3.59788321211
10.	*i…o	-3.36246611276	3.36246611276
11.	o…o	2.71007548827	2.71007548827
12.	e…e	2.33347924807	2.33347924807

As seen in (15), over- and underrepresentation can be shown in two ways. Over- and underrepresentation can be seen by considering the $\log_2$ O/E value, positive for overrepresentation, negative for underrepresentation. Alternatively, overrepresentation can be indicated by pairing the relative sequence (e.g. o...o) with the absolute $\lvert\log_2\rvert$ value, and underrepresentation indicated by pairing a starred sequence (e.g. *i...e) with the absolute $\lvert\log_2\rvert$ value.

To summarize, at this point, we have shown how counting the number of occurrences of vowel sequences can establish a hierarchy of the robustness of different patterns, and we have verified that there is sufficient data for the expected values to be significant. Of particular interest in the domain of language is that the gradient deviations from expectation can end up grammatically encoded as nongradient, categorical patterns. For example, in Ciyao while the sequence *e...i* is dispreferred, it does occur in some words. However, the suffix *-il* 'applicative' categorically surfaces as [el] after mid vowels and as [il] elsewhere (e.g., [el] after [e] (e.g., [pet-el-] 'ornament' but [saam-il-] 'move'): Importantly, the suffix does not occasionally show up as *-il* after mid vowels (*[pet-il-], etc.). Thus, the gradient tendencies found in the lexicon nonetheless translate into categorical grammatical effects. We explore how this can happen, again with no appeal to Universal Grammar, in the next section.

3.3 BUILDING A GRAMMAR

Here we address the question of how a grammar might be constructed based on information available from the distributional tendencies within a language. We propose that grammar construction is based on observations of the various over-and underrepresented properties within a language, expressed as $\log_2$ O/E as laid out in section 3.2. Since there is no *a priori* list identifying which properties are relevant in a language – or even which ones *might* be relevant – observations are collected about anything and everything that the learner takes note of. Many of these observations will not have significant skewing towards either under- or overrepresentation, so will not participate directly in grammar construction. In our example here, we focus on the height harmony pattern in our Bantu test languages; this is not to imply that there are no other grammatical properties to be learned in these languages.

We represent observations in terms of implicational relations. For example an observation about a $\{V_1 \ldots V_2\}$-sequence $x...y$ is expressed as follows:

$$\text{if } V_1 = x \text{ then } V_2 = y \qquad \textit{or} \qquad x \rightarrow y$$

An underrepresented $V_1 \ldots V_2$ sequence **x...y* is expressed negatively:

$$\text{if } V_1 = x \text{ then not } V_2 = y \qquad \textit{or} \qquad x \rightarrow \neg y$$

Each implication is assigned a $\log_2$ O/E value; the absolute values of the logs allows for a complete ranking of implications from largest to smallest effect, based on largest to smallest $|\log_2 \text{O/E}|$. Manipulating these implications to eliminate redundancy and to generalize the implications results in a nascent grammar.

The discussion up to this point has elaborated on the first three principles of grammar building (see (16)), observing, tracking, and ranking implicational relations. We now turn to the question of relationships among the set of implicational relations, the last three points in (16); our approach is to first expand this set, and then contract it through eliminating redundancy and increasing generality.[6] We continue with the Ciyao example to make the discussion concrete.

(16) **Principles for building a grammar**

a. Observe and track frequency of occurrence of observations.
b. Express observations as implicational relations.
c. Rank members of the family of related implicational relations from greatest to least effect.
d. Within a given implication family, expand on alternative possible implicational relations.
e. Eliminate redundancies among members of the family of implicational relations.
f. Increase coverage expressed by the family of implicational relations (generalize).

3.3.1 EXPANDING THE FAMILY OF IMPLICATIONAL RELATIONS

We use the term *family* for a set of implicational statements that fit a single frame. Here, we are interested in the family expressed as *If* V_1 *= x then (not)* V_2 *= y*. Other families include *If* V_2 *= x then (not)* V_1 *= y*, *If V = [F] then (not) V = [G]*, etc.

Expanding on the alternatives (16d) is done in two ways: (i) Each vowel in the language plays the role of V_1, and each vowel plays the role of V_2. This has already been assumed in the above discussion (section 3.2), where data is presented in a 5×5 grid and every cell is filled, even where the value is 0. In this way, the complete absence of some sequence is noted: for instance, in Ciyao verbs there are no sequences [i...e], [e..o], or [o...u]. (ii) Every possible *set* of vowels can take the role of V_1 and at the same time, every possible *set* of vowels can take the role of V_2. That is, relations are not expressed solely as relations between a specific pair of vowels, but they are also expressed between sets of vowels. We are interested in not only the frequency ranking of [i...e], [u...e], and the like, but also in the frequency of [{ i, u }...e] and of [e...{ i, u }], etc. The upshot for a language with 5 vowels is a set of 961 implicational relations. A few of these statements for Ciyao are given in (17). (Each of the implications is given a positive statement, e.g. (17-1) *If* V_1 *= { i } then* V_2 *= { i }* and (17.2) *If* V_1 *= { i } then* V_2 *= { e }*. The log O/E values indicate whether the implication is under- or overrepresented. An alternative would be to use |log O/E| and state the implications either positively or negatively. Had we used this alternative, (17.2) *If* V_1 *= { i } then not* V_2 *= { e }* with $|\log \text{O/E}| = \infty$. The two means of representation are notational equivalents.)

[6] The linearity of a list may imply that these six points are sequentially ordered in time. This is not necessary – nor, we would argue, desirable. The linguist – and learner – works with available data to construct a grammar of the forms available, and adjusts that grammar as further information comes to light. Thus, these principles are in constant use for grammar construction.

(17) **Expanding the family of implicational relations: Ciyao**

	implication	log O/E
1.	If V_1 = { i } then V_2 = { i }	0.975
2.	If V_1 = { i } then V_2 = { e }	$-\infty$
3.	If V_1 = { i } then V_2 = { a }	0.016
4.	If V_1 = { i } then V_2 = { o }	-3.362
5.	If V_1 = { i } then V_2 = { u }	-1.050
6.	If V_1 = { i, e } then V_2 = { i }	0.301
7.	If V_1 = { i, e } then V_2 = { e }	0.932
8.	If V_1 = { i, e } then V_2 = { a }	0.041
...		
961.	If V_1 = { i, e, a, o, u } then V_2 = { i, e, a, o, u}	0

Frequencies are determined for each of these 961 implicational relations, so that all implications can be ranked for size of effect with respect to all other implications. It is the combination of the rankings and the expanded family of implicational relations that allows the nascent grammar to emerge through reducing redundancy and increasing generalization.

Importantly, the expanded family membership is a result of manipulations of the basic implications, not the result of an independent set of observations. That is, this is a manipulation of one type of symbolic representation to create a new set of symbolic representations. The representations are becoming increasingly abstract. Redundancy and generalization continue the symbolic representation, as demonstrated below.

3.3.2 PROJECTION AND PROMOTION FROM A FAMILY OF IMPLICATIONAL RELATIONS TO A NASCENT GRAMMAR

Because we are exploring the idea that the symbolic representation of a grammar is driven by deviations from the expected, and that the more robust the deviation, the more likely that property is to be in the grammar, we propose beginning with the most robust deviations from the expected transition frequencies, and project those implicational relations to the grammar. In Ciyao, for example, the most robust implications are the three corresponding to the non-occurring sequences, *e...o, *o...u, and *i...e. Consequently, the three implications { e } $\rightarrow \neg$ { o }, { o } $\rightarrow \neg$ { u }, and { i } $\rightarrow \neg$ { e } project into the grammar.[7]

(18) **Ciyao grammar projection & promotion, iteration 1, the three most robust implications**

a. { e } $\rightarrow \neg$ { o } (absolute (∞))

b. { o } $\rightarrow \neg$ { u } (absolute (∞))

c. { i } $\rightarrow \neg$ { e } (absolute (∞))

Simply being projected into the grammar is not sufficient for an implication to be a statement in the final grammar because of the drive to become increasingly symbolic, accomplished through reducing redundancies and increasing generalization. We conceptualize this process as consisting of layers of generalizations, each layer constituting a level of increased generalization. We refer to the inclusion of an implication at a higher level as *promotion within the grammar*. The promoted statements together form the grammar, though less general layers, on which the promoted layer is based, are not eliminated. The core promotion principles are given in (19b,c).

[7] In (18) and subsequent iterations, the numbers in parentheses for each point give the $|\log_2 O/E|$ values; underrepresentation is indicated by negative implications while overrepresentation is indicated by positive implications.

(19) **Emergent Grammar Principles governing the projection and promotion of implications**

a. **Projection.** Beginning from the most robust deviations from the expected and proceeding until an implication is at the level of chance, assign an implication to the grammar.

b. **Redundancy.** Promote nonredundant implications.

i. *When assigning a subsumed implication.* Where the effect of an implication is a subset of a previously promoted implication or set of implications, the new implicational statement is not promoted.

ii. *When assigning an inclusive implication.* Where a previously promoted implication is a proper subset of the new implication, promote the new implication to the grammar and remove the previously promoted statement from the set of promoted implications.

c. **Generalization.** A single general implication is promoted over two compatible specific ones.

i. *Compatibility.* The two implications are both positive or they are both negative.

ii. *Antecedent generalization.* If two (compatible) implications share an antecedent, combine the consequents to express as a single implication.

iii. *Consequent generalization.* If two (compatible) implications share a consequent, combine the antecedents to express as a single implication.

Redundancy prevents repetition of information that is already promoted in the grammar, by failure to promote a new implication or by removal from the promoted set of an old implication that is rendered redundant by the projection of a new implication. For example, the fourth ranked of the robust implications in Ciyao is { i, u } $\rightarrow \neg$ { e }. When compared to the Ciyao grammar iteration in (18), we see that { i } $\rightarrow \neg$ { e } is subsumed within { i, u } $\rightarrow \neg$ { e }. Redundancy (19b) results in the promotion of the more general of these two statements, leaving only { i, u } $\rightarrow \neg$ { e } promoted for the next iteration of Ciyao grammar. We indicate implications that are not members of the promoted set with a single strike through the text.[8]

(20) **Ciyao grammar projection & promotion, iteration 2, the four most robust implications**

a. { e } $\rightarrow \neg$ { o } (absolute (∞))

b. { o } $\rightarrow \neg$ { u } (absolute (∞))

~~{ i } $\rightarrow \neg$ { e }~~ (absolute (∞)) due to Redundancy (19b-ii)

c. { i, u } $\rightarrow \neg$ { e } (5.94)

This same principle comes into play when the next implication is considered, based on *e...{o, i}. The corresponding implication, { e } $\rightarrow$ $\neg$ { o, i }, renders (20a) redundant, so the new implication is promoted over the original.

(21) **Ciyao grammar projection & promotion, iteration 3, the five most robust implications**

~~{ e } $\rightarrow \neg$ { o }~~ (absolute (∞)) due to Redundancy (19b-ii)

a. { o } $\rightarrow \neg$ { u } (absolute (∞))

b. { i, u } $\rightarrow \neg$ { e } (5.94)

c. { e } $\rightarrow$ $\neg$ { o, i } (5.75)

[8] See footnote 7 for an explanation of "5.94" in item (20c).

The sixth implication illustrates the other aspect of Redundancy: an implication is not promoted into the grammar if it is already contained within an existing implication. In this case, the observation is *e...i, already contained within *e...{ o, i }. The new implication fails to be promoted because it would be redundant.

(22) **Ciyao grammar projection & promotion, iteration 4, the six most robust implications**

a. { o } $\rightarrow \neg$ { u } (absolute (∞))

b. { i, u } $\rightarrow \neg$ { e } (5.94)

c. { e } $\rightarrow \neg$ { o, i } (5.75)

~~{ e } $\rightarrow \neg$ { i }~~ (5.22) due to Redundancy (19b-i)

The seventh most robust implication, { e, a } $\rightarrow \neg$ { o }, provides new information: it does not introduce redundancy nor does it share sufficient information with another implication for the two to generalize. The result is that this implication is promoted to the grammar with no further ado.

(23) **Ciyao grammar projection & promotion, iteration 5, the seven most robust implications**

a. { o } $\rightarrow \neg$ { u } (absolute (∞))

b. { i, u } $\rightarrow \neg$ { e } (5.94)

c. { e } $\rightarrow \neg$ { o, i } (5.75)

d. { e, a } $\rightarrow \neg$ { o } (5.17)

The eighth implication is { e, u } $\rightarrow \neg$ { o }, which bears the "consequent generalization" relation to (23d): both have { o } as the consequent (see (19c-ii)). As a result, by Generalization, the more complex version, formed by the merger of these two implications, is promoted, while the two originals are not promoted.

(24) **Ciyao grammar projection & promotion, iteration 6, the eight most robust implications**

a. { o } $\rightarrow \neg$ { u } (absolute (∞))

b. { i, u } $\rightarrow \neg$ { e } (5.94)

c. { e } $\rightarrow \neg$ { o, i } (5.75)

~~{ e, a } $\rightarrow \neg$ { o }~~ (5.17) due to Generalization (19c-iii)

~~{ e, u } $\rightarrow \neg$ { o }~~ (5.14) due to Generalization (19c-iii)

d. { e, a, u } $\rightarrow \neg$ { o } due to Generalization (19c-iii)

In summary, we have illustrated how the eight most robust implications would be projected for Ciyao; the resulting nascent grammar has four promoted implications.

(25) **Ciyao grammar promoted implications (projecting only the eight most robust implications)**

a. { o } $\rightarrow \neg$ { u }

b. { i, u } $\rightarrow \neg$ { e }

c. { e } $\rightarrow \neg$ { o, i }

d. { e, a, u } $\rightarrow \neg$ { o }

3.4 STOPPING THE LEARNER

In principle, any implication involving a pattern that is more robust than chance distribution can be projected to the grammar. In practice, it appears that the learner must stop well before reaching the level of chance, that is, a robustness of $\log_2 = 0$. We can see this in the case of Bantu height harmony. Grammaticization must stop before adding an implication to the grammar that would predict the wrong results under morpheme concatenation. That is, all implications that result in productive morpheme alternation must be included in the grammar, but implications that do not result in alternations should not be added – or must at least be of a different status. In the Bantu height harmony case, an implication like $\{ \text{e} \} \rightarrow \neg\{ \text{i} \}$ is important to include because of alternations in the applicative: [dim-]/[dim-il-] 'cultivate/applicative', [wut-]/[wut-il-] 'pull/applicative', [saam-]/[saam-il-] 'move/applicative' but [pet-]/[pet-el-] 'ornament/applicative', *[pet-il-]. By contrast, an implication like $\{ \text{e} \} \rightarrow \neg\{ \text{u} \}$ should not be added to the grammar because suffixes with /u/ surface as [u], not [o], following /e/ verbs: the reversive of [sweek-] 'insert' is [sweek-ul-] 'pull out', not *[sweek-ol-]. Since the absolute $\log_2$ value for a condition like $\{ \text{e, o} \} \rightarrow \neg\{ \text{i, u} \}$ is 1.94 in Ciyao, and this constraint would predict alternations of the unattested type, it is crucial that the learner be stopped before including such a condition.

The problem, as with all unsupervised learning, is what causes the learner to stop? One family of hypotheses about what causes grammaticization to stop is that there is some threshold below which implications are not projected into the grammar, a proposal which immediately raises the question of "what is the threshold?". We explore three hypotheses here. The first is that grammaticization stops when the magnitude of $\log_2$ O/E is less than some value (we use the value 2.5). All implications with $\log_2$ O/E values more robust than 2.5 are considered for projection into the grammar; those with a lower value are not. The second hypothesis is that grammaticization is driven by the subimplications, not the implications themselves, and that grammaticization stops when the magnitude of $\log_2$ O/E for at least one relevant subimplication is less than some value (we use 1.25). The final hypothesis considered here is that grammaticization halts when any positive implication is encountered. These three hypotheses are given in (26).

(26) **Hypotheses for putting a halt to grammaticization**

1. **Hypothesis 1: implication magnitude** Learner stops when the magnitude of $|\log_2 \text{O/E}|$ is less than some value; value investigated here = 2.5
2. **Hypothesis 2: subimplication magnitude** Learner stops when the magnitude of $|\log_2 \text{O/E}|$ is less than some value for any component implication; value investigated here = 1.25
3. **Hypothesis 3: no positive implications** Learner stops under some non-threshold implication; here, any positive implication

As we will see, hypothesis 2 fares the best against the data examined here.

4 RESULTS

We turn now to the results of our tests. For each language examined, there are three sub-tests to be considered, based on the three different hypotheses about how grammaticization ceases.

4.1 CRITERIA FOR SUCCESS

Recall that the pattern under consideration is Bantu Height Harmony, illustrated in (2), repeated in (27) for convenience.

(27) **Bantu height harmony in Ciyao, repeated from (2)**

a. *'-il' applicative*			b. *'-ul' reversive*		
dim-	dim-il-	*cultivate*	siv-	siw-ul-	*close/open up*
wut-	wut-il-	*pull*	uuv-	uuw-ul-	*hide/reveal*
saam-	saam-il-	*move*	mat-	mat-ul-	*adhere/peel off*
pet-	pet-el-	*ornament*	sweek-	sweek-ul-	*insert/pull out*
soom-	soom-el-	*read/study*	som-	som-ol-	*pierce/extract*

This is the harmonic pattern for all languages tested.

The implications that must be extracted from the frequencies in order to account for the height harmony pattern are the critical test conditions – the ones that place restrictions on high vowels after mid vowels (28). While there are undoubtedly other patterns to be identified in the distribution of vowels in each language, our focus remains on these three test implications.

(28) **Critical test implications for Bantu Height Harmony I**

a. A high front vowel may not follow a mid front vowel: *e...i.
b. A high front vowel may not follow a mid back vowel: *o...i.
c. A high back vowel may not follow a mid back vowel: *o...u.

A language with Bantu Height Harmony may express these three restrictions in different ways, for instance grouping the two "high front vowel" restrictions together to give "A high front vowel may not follow a mid vowel" or grouping the two "mid back vowel" restrictions together to give "A high vowel may not follow a mid back vowel". Additionally, a language might combine one of these with a restriction of no direct relevance to the Bantu Height Harmony case (italics show the irrelevant portion): "*[o] or* a high front vowel may not follow a mid front vowel."

In addition to encoding constraints prohibiting the sequences of (28), in order to encode the "asymmetric" pattern, it is crucial that derived grammars *not* include a condition prohibiting the sequence e...u. As seen in (27), forms like *sweek-ul-* are grammatical and must be allowed by the grammar.

(29) **Critical test implications for Bantu Height Harmony II**

a. A high back vowel may follow a mid front vowel: no *e...u condition in the grammar

The Emergent Grammar hypothesis claims that if a language has the Height Harmony pattern, the restrictions in (28), but not (29), will be expressed by the implications promoted in the grammar; presence in the grammar is based on frequency robustness and the Promotion Principles (19), as limited by one of the three subhypotheses for when to put a halt to grammaticization (26).

As we discuss each test language, we show the grammars that result given the Emergent Grammar Principles (19) and each of the three stopping strategies. The first three languages, Ciyao, Chichewa, and Ikalanga, all promote an appropriate grammar within the iteration space allowed by the three cut-off hypotheses.

4.2 CIYAO, CHICHEWA, AND IKALANGA: SUCCESS

Figure (30) shows the promoted grammar set for Ciyao. The implications that are critical in accounting for Bantu Height Harmony are indicated under the heading "Results required for success": all three of (28a,b,c) should appear in this column. The implications that are promoted in the grammar but that are not critical are in italics. The horizontal lines indicate where grammaticization ceases, based on each of the three hypotheses. The hypothesis and whether it succeeds or fails is indicated in the rightmost column.

(30) **Test 1 results: Ciyao verbs**

		Results required for success		
a.	V1 {e} → V2 not {*o*, i}	*e...i	(28a)	
b.	V1 {o} → V2 not {i, u}	*o...{ i, u }	(28b,c)	
c.	*V1 {i, a, u} → V2 not {e, o}*	*n/a*		
d.	*V1 {i, e, a, u} → V2 not {o}*	*n/a*		
e.	V1 {e, o} → V2 not {i}	*{ e, o }...i	(28a,b)	H3: Succeeds
f.	*V1 {o} → V2 {o}*	*n/a*		H1: Succeeds
g.	*V1 {e} → V2 {e}*	*n/a*		H2: Succeeds

Key: H1: Log magnitude of all promoted implications is >2.5
H2: Log magnitude of all components of promoted implications is >1.25
H3: All promoted implications are negative

For Ciyao, Bantu Height Harmony is covered by three promoted implications, given in (31). (Note that the implication in (31a) includes extraneous information shown by italics: { o } appears in the consequent, unnecessary for the Bantu Height Harmony grammar.)

(31) **Promoted Bantu Height Harmony implications for Ciyao**

a.	{ e } → ¬ { i }	*from*	V1 {e} → V2 not {*o*, i}
b.	{ o } → ¬ { i, u }	*from*	V1 {o} → V2 not {i, u}
c.	{ e, o } → ¬ { i }	*from*	V1 {e, o} → V2 not {i}

The Bantu Height Harmony pattern is expressed by the three implications in (31), all of which are promoted in iterations before grammaticization is cut off regardless of which of the three hypotheses is selected. Correctly, the allowed sequence e...u is not prohibited (29). It is appropriate to note as well that we do not assume that the grammar that results directly from the Emergence principles of (19) is the final state of the grammar. The results we are examining here constitute a stage in the development of a full phonological grammar.

We now consider Chichewa. The promoted grammar set for Chichewa is given in (32), using the presentation conventions laid out above.

(32) **Test 2 results: Chichewa verbs**

		Results required for success		
a.	*V1 {i, e, a, u} → V2 not {o}*	*n/a*		
b.	*V1 {i, a, u} → V2 not {e, o}*	*n/a*		
c.	V1 {o} → V2 not {i, u}	*o...{ i, u }	(28b,c)	
d.	V1 {e} → V2 not {*o*,i}	*e...i	(28a)	
e.	V1 {e, o} → V2 not {i}	*{ e, o }...i	(28a,b)	H3: Succeeds
f.	*V1 {o} → V2 {o}*	*n/a*		H1: Succeeds
g.	*V1 {e} → V2 {e}*	*n/a*		
h.	*V1 {i} → V2 {e, o, u}*	*n/a*		
i.	V1 {o,*i*} → V2 not {u}	*o...u	(28c)	H2: Succeeds

Key: H1: Log magnitude of all promoted implications is >2.5
H2: Log magnitude of all components of promoted implications is >1.25
H3: All promoted implications are negative

As in Ciyao, in Chichewa the Emergent Grammar principles of (19) serve to promote implications that account for Bantu Height Harmony, in a fashion consistent with the three cut-off hypotheses of (26). Again, the well-formed sequence e...u is not prohibited. The promoted grammars of Ciyao and Chichewa include the same three implications.[9]

The same three implications also emerge in Ikalanga, The promoted grammar set for Ikalanga is given in (33), again following the same presentation conventions.

(33) **Test 3: Ikalanga verbs**

		Results required for success		
i.	V1 {o} → V2 not {i, u}	*o...{ i, u }	(28b,c)	
ii.	*V1 {e, u} → V2 not {o}*	*n/a*		
iii.	V1 {e, o} → V2 not {i}	*{ e, o }...i	(28a,b)	
iv.	*V1 {u} → V2 not {e, o}*	*n/a*		H3: Succeeds
v.	*V1 {o} → V2 {o}*	*n/a*		
vi.	*V1 {e,i, u} → V2 not {o}*	*n/a*		
vii.	V1 {e} → V2 not { *o*,i}	*e...i	(28a)	
viii.	*V1 {i, u} → V2 not {e, o}*	*n/a*		H1: Succeeds
ix.	*V1 {e} → V2 {e}*	*n/a*		
x.	*V1 {a,i, u} → V2 not {e}*	*n/a*		H2: Succeeds

Key: H1: Log magnitude of all promoted implications is >2.5
H2: Log magnitude of all components of promoted implications is >1.25
H3: All promoted implications are negative

Again, the same three restrictions emerge, successfully characterizing Bantu Height Harmony pattern, regardless of which cut-off point is used. The same set of implications is promoted for each of these three languages, given in (34). In none of the languages is e...u prohibited.

(34) **Promoted Bantu Height Harmony implications for Ciyao, Chichewa, and Ikalanga**

a. { e } → ¬ { i } *from* V1 {e} → V2 not {*o*,i}
b. { o } → ¬ { i, u } *from* V1 {o} → V2 not {i, u}
c. { e, o } → ¬ { i } *from* V1 {e, o} → V2 not {i}

4.3 JITA, AND BUKUSU: FAILURE OF TWO CUT-OFF HYPOTHESES

Jita presents the first case in which one of the cut-off hypotheses fails. Based on patterns of suffix alternation, it appears that [u] is allowed after [e] in Jita; see (29). According to Hypothesis 1 (whereby the $\log_2$ magnitude of all promoted implications is >2.5), however, [u] should be prohibited after [e], an incorrect prediction. The promoted grammar for Jita is given in (35).

[9] The three implications enter the grammar in different iterations. Whether this difference results in a different final grammar for the two languages goes beyond the scope of this paper.

(35) **Test 4 results: Jita verbs**

		Results required for success		
a.	V1 {e} → V2 not {o,i}	*e...i	(28a)	
b.	V1 {e, o} → V2 not {i}	*{ e, o }...i	(28a,b)	
c.	V1 {o} → V2 not {i, u}	*o...{ i, u }	(28b,c)	
d.	*V1 {a, u} → V2 {e, o}*	*n/a*		
e.	*V1 {e, a, u} → V2 {o}*	*n/a*		H3: Succeeds
f.	*V1 {o} → V2 {o}*	*n/a*		H2: Succeeds
g.	**V1 {e, o} → V2 not {i, u}**	***{ e, o }...{ i, u }**		H1: Fails

Key: H1: Log magnitude of all promoted implications is >2.5
H2: Log magnitude of all components of promoted implications is >1.25
H3: All promoted implications are negative

Hypothesis 1 fails in Jita because the implication *V1 {e, o} → V2 not {i, u}* is overly restrictive: it prevents not only the rare to non-occurring *o...u, o...i*, and *e...i* sequences, but it also rules out the acceptable *e...u* sequence. Note that both Hypotheses 2 and 3 correctly stop the grammaticization of implications before inappropriate implications are promoted.

A grammar of Bantu Height Harmony also emerges in Bukusu if we assume Hypothesis 2, but both Hypotheses 1 and 3 are seen to fail. The promoted grammar set for Bukusu is given in (36), using the familiar presentation conventions.

(36) **Test 5 results: Bukusu verbs**

		Results required for success		
a.	*V1 {i, u} → V2 not {e, o}*	*n/a*		
b.	*V1 {i, a, u} → V2 not {e}*	*n/a*		
	V1 {o} → V2 not {u}	*o...u	(28c)	H1: Fails
c.	*V1 {i, e, u} → V2 not {o}*	*n/a*		H3: Fails
d.	*V1 {e} → V2 {e}*	*n/a*		
e.	*V1 {o} → V2 {o}*	*n/a*		
f.	V1 {o} → v2 not {i, u}	*o...{ i, u }	(28b,c)	
g.	V1 {e} → V2 not {o,i}	*e...i	(28a)	H2: Succeeds

Key: H1: Log magnitude of all promoted implications is >2.5
H2: Log magnitude of all components of promoted implications is >1.25
H3: All promoted implications are negative

While all three hypotheses correctly allow the sequence e...u, both cut-off Hypothesis 1 and 3 nevertheless fail. The Hypothesis 1 requirement that the $|\log_2|$ for all promoted implications be >2.5 is too stringent: Only one of the necessary prohibitions passes this high bar. Similarly, the Hypothesis 3 requirement that positive implications mark the cut off point also sets too high a bar. Only Hypothesis 2 admits the implications necessary to fully account for the Bantu Height Harmony pattern.

4.4 NKORE-KIGA: FAILURE FOR DIFFERENT REASONS

The Emergent Grammar Principles of (19) fail with respect to Nkore-Kiga, but the failure is because all three hypotheses cut off grammaticization too soon, before the necessary implications have been projected into the grammar. This is similar to the Bukusu case, where two of the hypothesized cut-offs failed to include critical implications.

(37) **Test 6 results: Nkore-Kiga verbs**

		Results required for success		
a.	V1 {o} → V2 not {i, u}	*o...{ i, u }	(28b,c)	
b.	*V1 {i, u} → V2 not {e, o}*	*n/a*		
c.	*V1 {i, a, u} → V2 not {e}*	*n/a*		H3: Fails
d.	*V1 {o} → V2 {o}*	*n/a*		H1: Fails
				H2: Fails

Key:
H1: Log magnitude of all promoted implications is >2.5
H2: Log magnitude of all components of promoted implications is >1.25
H3: All promoted implications are negative

As inspection of (37) reveals that all of Hypotheses 1, 2, and 3 fail to account for absence of [i] after a mid front vowel [e]. Promotion is cut off too early.

4.5 INTERIM SUMMARY

At this point, we have seen that the model succeeds in promoting appropriate grammars for 5 of the 6 languages if we adopt cut-off Hypothesis 2. In Jita, Hypothesis 1 failed because it admitted an implication which is counter to the Bantu Height Harmony pattern; see (29). In Bukusu, by contrast, Hypotheses 1 and 3 failed to admit implications necessary to the Bantu Height Harmony pattern; see (28). That is, failure is possible both by commission and by omission.

(38) **Interim summary**

	Hypothesis 1	Hypothesis 2	Hypothesis 3
Ciyao	Success	Success	Success
Chichewa	Success	Success	Success
Ikalanga	Success	Success	Success
Jita	Failure	Success	Success
Bukusu	Failure	Success	Failure
Nkore-Kiga	Failure	Failure	Failure

Hypothesis 1: Log magnitude >2.5
Hypothesis 2: All components >1.25
Hypothesis 3: No positive implications

All three hypotheses failed, however, when faced with the Nkore-Kiga data. In the next section we briefly address the interaction of incompatible effects which, we propose, offers an explanation – and invites a reconsideration – of the Nkore-Kiga analysis. This interaction shows that with a different understanding of the data, Nkore-Kiga also fits within the Emergent Grammar hypothesis.

5 INTERACTIONS IN NKORE-KIGA

In this section, we address the question of why the system failed with Nkore-Kiga. Nothing unusual about Bantu Height Harmony is recorded in the sources on the language (Taylor 1985; Taylor 1959 via CBOLD). Examples given in (39) show the familiar asymmetric pattern.

(39) **Height harmony in Nkore-Kiga**

stem		'-ir' benefactive		'-urur/uur' reversive	
hígik-a	'support'	hígik-ir-a	'fix in place'		
zinga	'roll up'			zing-urur-a	'unwind'
bík-a	'bring news of the death of'			bík-urur-a	'contradict a report of the death of'
kúb-a	'fold'	kúb-ir-a	'surround'	kúb-urur-a	'unfold'
gamb-a	'speak'	gamb-ir-a	'tell'	gamb-uur-a	'speak behind someone's back'
rém-a	'become lame'	rém-er-a	'become lame'		
tég-a	'trap'			tég-urur-a	'unset a trap'
kóm-a	'tie up'	kóm-er-an-a	'tie together'	kóm-oror-a	'unfasten'

The problem in Nkore-Kiga is not in the nature of the harmony pattern itself. However, inspection of the data reveal a different asymmetry between Nkore-Kiga and the other languages. In Nkore-Kiga, there is an unexpectedly large number of forms which begin with an *e...i* sequence – a sequence we do not expect to see in large numbers. Examples are given in (40).

(40) **Nkore-Kiga problem**

stem	*word*	
[ésiga]	kwésiga	'trust'
[étinda]	kwétinda	'hide'
[égiza]	kwégiza	'pretend'
etc.		

Larry Hyman (pc) notes that many of the relevant stems involve a reflexive prefix 'e-' as shown here (examples include the infinitival prefix /ku-/):

(41) **Reflexive prefix** *e-*

/ku-búza/	[kubúza]	'to cause loss, be lost to'
/ku-é-búza/	[kwébuza]	'to hide oneself'

We make the assumption that initial vowels of vowel-initial stems are not part of the phonological stem, hence outside of the relevant domain for height harmony. The assumption is plausible both because of their morphological status (41) and because of their syllabification with the preceding prefix (40). This gives the structure below:

kwé[siga] 'trust', kwé[tinda] 'hide', kwé[giza] 'pretend', etc.

Reevaluating the Nkore-Kiga data making this assumption yields quite different results, summarized in (42).

(42) **Test 6 results, revisited: Nkore-Kiga verbs; initial vowels of vowel-initial stems excluded**

		Results required for success		
a.	V1 {o} → V2 not {i, u}	*o...{ i, u }	(28b,c)	
b.	*V1 {i, e, u} → V2 not {o}*	*n/a*		
c.	V1 {e} → V2 not {o,i}	*e...i	(28a)	
d.	*V1 {i, u} → V2 not {e, o}*	*n/a*		
e.	V1 {e, o} → V2 not {i}	*{ e, o }...i	(28a,b)	
f.	*V1 {i, a, u} → V2 not {e}*	*n/a*		H3: Succeeds
vii.	*V1 {o} → V2 {o}*	*n/a*		H1: Succeeds
viii.	*V1 {e} → V2 {e}*	*n/a*		H2: Succeeds

Key: H1: Log magnitude of all promoted implications is >2.5
H2: Log magnitude of all components of promoted implications is >1.25
H3: All promoted implications are negative

The anomalous results for Nkore-Kiga are resolved through a deeper understandingof the morphological and phonological properties of the language. With the assumption that initial [e] in Nkore-Kiga is syllabified as part of the prefix, not as part of the stem, a grammar is promoted that satisfies all cut-off Hypotheses and that accounts for the Bantu Height Harmony pattern. It is of course important to solve the problem of how such interacting patterns are learned, a topic that is not within the scope of this paper to deal with.

6 CONCLUSION & REFLECTIONS

We have argued here that the frequency distribution of vowels in $V_1 \ldots V_2$ sequences provides sufficient evidence to identify the implications that are crucial for characterizing Bantu Height Harmony. In doing this, we have laid out the groundwork for developing the first steps towards an Emergent Grammar. In this concluding section, we explore this model further, then consider implications for the grammars of specific languages, in particular the Bantu Height Harmony languages focused on in this paper.

6.1 EMERGENT GRAMMAR

We made a small set of assumptions about the capabilities of language learners:

(43) **Assumptions**

a. Learners identify vowel sequences
b. Learners calculate frequency of distributions
c. Learners rank vowel pairs by frequency
d. Learners create symbolic systems

With these assumptions, one result is that every time a lexical item is learned, the frequencies change slightly. If the grammar of a language is based directly on these frequencies as Emergence proposes, then the grammar is never “complete”: The addition of each lexical item causes a readjustment in frequencies – and this may result in a different grammar.

However, the degree to which a single datum will affect the grammar depends on how many lexical items are already present. As the lexicon grows larger, the impact of individual items diminishes. As the data set grows, the chances for new frequency distributions with perceptible impacts lessens.

Under the Emergent view, the information that emerges as a result of these assumptions gives rise to implicational statements, a symbolic representation of the $V_1 \ldots V_2$ sequences. These implications are projected into a more complex symbolic system, a nascent grammar, in accord with the Emergent Grammar Principles, (19). These Principles also promote some of the implications within the grammar; with each promotion, the grammar becomes less closely tied to the frequencies that gave rise to the implications, and so becomes increasingly symbolic (cf. Deacon 1997). Consider, for instance, the Ciyao implication { i, u } → ¬ { e }. This implication comes to play a role in the nascent grammar because its two components are promoted and the generalized constraint emerges: { i } → ¬ { e } is an absolute prohibition; the implication that is subsumed in { i, u } → ¬ { e }, namely { u } → ¬ { e }, is ranked tenth in terms of robustness, but acquires significance in part because of the strength of { i } → ¬ { e }. In effect, by pooling the strength of compatible constraints, the grammar generalizes, becoming more symbolic, less concrete.

This again raises the question of "when does it all stop?". We asked this question above in the context of which implications project into the grammar; here the issue is perhaps related but different: When is the system sufficiently symbolic? The Emergent Grammar response is that the grammar continues to become more symbolic as long as there are principles allowing for more symbolic representations of the grammar. The challenge to the linguist is to identify those principles.

With the analysis of Bantu Height Harmony presented here, the analysis does not extend beyond projecting implications into the grammar and promoting certain of those projections. Even with this rudimentary version of the Bantu grammars, there are interesting results to consider.

6.2 VARIATION AMONG THE "GRAMMARS" OF BANTU HEIGHT HARMONY

Despite differences in frequencies, all nascent grammars in the test languages contained virtually the same implications. For instance, the Ciyao implications from (31), repeated in (44), are identical to those which emerge for Chichewa, Ikalanga, and Jita.

(44) **Promoted Bantu Height Harmony implications for Ciyao**

a.	{ o } → ¬ { i, u }	*from*	V1 {o} → V2 not {i, u}
b.	{ e, o } → ¬ { i }	*from*	V1 {e, o} → V2 not {i}

The emergent Bukusu grammar differs; it is given in (45).

(45) **Promoted Bantu Height Harmony projections for Bukusu**

a.	{ o } → ¬ { i, u }	*from*	V1 {o} → V2 not {i, u}
b.	{ e } → ¬ { i }	*from*	V1 {e} → V2 not {*o*,i}

The difference is that Ciyao (and Ikalanga, Jita, and Chichewa) characterize the pattern with some redundancy, including { o } in the antecedent of { e, o } → ¬ { i }, even though this is also expressed by { o } → ¬ { i, u }. The Bukusu grammar contains no such repetition.

This comparison shows that different sets of implicational statements can result in the same surface patterns. One important area for research in Emergent Grammar is to understand the next steps in developing the symbolic system; once that is worked out, we will be in position to determine whether the Bukusu grammar is or is not identical (in this regard) to the Ciyao-Ikalanga-Jita-Chichewa grammar.

6.3 OTHER PROMOTED IMPLICATIONS

Our focus here has been on the implications that are critical for expressing Bantu Height Harmony. However, other implications are promoted in these nascent grammars; these additional implications are an interesting set. Figure (46) lists the implicational conditions that are promoted, there is a "yes" in the language column if the language promotes that implication under any of the three hypotheses. In the case of Nkore-Kiga, a "yes" indicates that the implication is included even if vowel-initial status is not taken into consideration; cases marked "(yes)" are promoted only when the vowel-initial cases are excluded.

(46) **Additional implications in nascent grammars**

		Bukusu	Ciyao	Chichewa	Ikalanga	Jita	Nkore-Kiga
a.	{e} → ¬ {o}		yes			yes	
b.	{i, a, u} → ¬ {e, o}		yes	yes			
c.	{i, e, a, u} → ¬ {e, o}		yes				
d.	{i, e, a, u} → ¬ {o}			yes			
e.	{o} → {o}	yes		yes	yes		yes
f.	{e} → {e}	yes		yes	yes		(yes)
g.	{i} → {e, o, u}			yes			
h.	{e, u} → ¬ {o}				yes		
i.	{u} → ¬ {e, o}				yes		
j.	{i, e, u} → ¬ {o}	yes			yes		(yes)
k.	{i, u} → ¬ {e, o}	yes			yes		yes
l.	{i, a, u} → ¬ {e}	yes			yes		yes
m.	{a, u} → ¬ {e, o}					yes	
n.	{e, a, u} → ¬ {o}					yes	

A quick scan down the "consequent" side of these implications reveals that every one of them limits the distribution of {e, o} in V_2 position. There is a very strong preference for a sequence of mid vowels to be identical in these languages, in accord with the positive (46e,f); the negative implications express the same pattern by prohibiting {e, o} after some combination of the three other vowels. Each language, however, expresses this preference with a different set of conditions. For instance, in Ciyao, the implications allow mid vowels only after { o } while the other four all include the positive implications requiring identity between mid vowels in $V_1 \ldots V_2$ sequences. This is reassuring: languages differ even when they are highly similar, and here we see one way that Ciyao differs from the others.

6.4 CONCLUSION: HOW MUCH DO WE GET AWAY WITH(OUT)?

Emergent Grammar is the hypothesis that there is little to no genetic endowment with specific principles governing the phonology of language. The Emergent Grammar exploration of languages with Bantu Height Harmony presented here demonstrates that frequencies can lead to implicational statements that characterize the harmonic pattern. Key to this analysis is the assumption that humans are capable of identifying similarity, calculating frequencies, and creating symbolic systems.

REFERENCES

Archangeli, D. and Langendoen, T. (1997). *Optimality Theory*. Blackwell.

Archangeli, D., Mielke, J., and Pulleyblank, D. (2011). Greater than noise: Frequency effects in Bantu height harmony. In (to appear) B. Botma and Noske, R., editors, *Phonological Architecture: Empirical, Theoretical and Conceptual Issues – Papers in Honour of Norval S.H. Smith*. Mouton de Gruyter, Berlin.

Barrett, M. (1995). Early lexical development. In Fletcher, P. and MacWhinney, B., editors, *The Handbook of Child Language*, pages 362–392. Basil Blackwell, Oxford.

Bonatti, L., na, M. P., Nespor, M., and Mehler, J. (2005). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychological Science*, 16(6):451–459.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. Harper & Row, New York.

Clark, E. V. (1995). Later lexical development and word formation. In Fletcher, P. and MacWhinney, B., editors, *The Handbook of Child Language*, pages 393–412. Basil Blackwell, Oxford.

Deacon, T. (1997). *The Symbolic Species: The Co-evolution of Language and the Brain*. W.W. Norton & Company, New York.

Downing, L. (1999). Jita electronic dictionary provided to CBOLD. http://www.cbold.ish-lyon.cnrs.fr/; includes data from the Tanzania Language Survey.

Frisch, S. (1996). *Similarity and Frequency in Phonology*. PhD thesis, Northwestern University.

Frisch, S., Pierrehumbert, J., and Broe, M. (2004). Similarity avoidance and the OCP. *Natural Language and Linguistic Theory*, 22(179–228).

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, 13(5):431.

Gómez, R. and Gerken, L. (1999). Artificial grammar learning by one-year-olds leads to specific and abstract knowledge. *Cognition*, 70(2):109–135.

Guthrie, M. (1967-71). *Comparative Bantu: An Introduction to the Comparative Linguistics and Prehistory of the Bantu Languages*. Vols. I-IV. Greggs, London.

Hyman, L. M. (1999). The historical interpretation of vowel harmony in Bantu. In Hombert, J.-M. and Hyman, L. M., editors, *Bantu Historical Linguistics: Theoretical and empirical perspectives*, pages 235–295. CSLI, Stanford.

LaCross, A. (2011). *The Role of Language-specific Phonology: Tracking Linguistics Variables in Khalkha Mongolian*. PhD thesis, University of Arizona.

Mathangwane, J. (1994). Kalanga wordlist. CBOLD.

Mathangwane, J. (1999). *Ikalanga Phonetics and Phonology: A Synchronic and Diachronic Study*. Stanford Monographs in African Languages, Stanford, California.

Maye, J. (2000). *Learning Speech Sound Categories on the Basis of Distributional Information*. PhD thesis, University of Arizona, Tucson.

Maye, J., Werker, J., and Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3):101–111.

McCarthy, J., editor (2003). *Optimality Theory in Phonology: A Reader*. Blackwell, Malden, MA and Oxford.

McCarthy, J. and Prince, A. (1993). Generalized alignment. In Booij, G. and van Marle, J., editors, *Yearbook of Morphology*, pages 79–153. Kluwer, Dordrecht.

McCarthy, J. and Prince, A. (1994). The emergence of the unmarked: Optimality in prosodic morphology. In Gonzàlez, M., editor, *Proceedings of the North East Linguistic Society 24*, pages 333–379. GLSA Publications, Amherst, MA.

Mielke, J. (2005). Ambivalence and ambiguity in laterals and nasals. *Phonology*, 22.2:169–203.

Mielke, J. (2008). *The Emergence of Distinctive Features*. Oxford University Press, Oxford.

Mohanan, K. P., Archangeli, D., and Pulleyblank, D. (2010). The emergence of Optimality Theory. In Uyechi, L. and Wee, L.-H., editors, *Reality Exploration and Discovery: Pattern Interaction in Language and Life*, pages 143–158. Center for the Study of Language and Information, Stanford University.

Mtenje, A. (2001). Chewa wordlist. CBOLD.

Mutonyi, N. (2000). *Aspects of Bukusu morphology and phonology*. PhD thesis, The Ohio State University.

Ngunga, A. (2000). *Phonology and Morphology of the Ciyao Verb*. CSLI, Stanford.

Onnis, L., Christiansen, M., Chater, N., and Gómez, R. (2003). Reduction of uncertainty in human sequential learning: Evidence from artificial grammar learning. In *Proceedings of the 25th annual conference of the cognitive science society*, pages 886–891.

Onnis, L., Monaghan, P., Christiansen, M., and Chater, N. (2004). Variability is the spice of learning, and a crucial ingredient for detecting and generalizing in nonadjacent dependencies. In *Proceedings of the 26th annual conference of the Cognitive Science Society*, pages 1047–1052.

Pierrehumbert, J. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech*, 46:115–154.

Pierrehumbert, J. B. (1993). Dissimilarity in the arabic verbal roots. In *Proceedings of the North East Linguistics Society 23*, pages 367–381.

Prince, A. and Smolensky, P. (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*. Technical Report RuCCS-TR-2. Rutgers University Center for Cognitive Science, New Brunswick, NJ.

Pulleyblank, D. (2006). Minimizing UG: Constraints upon constraints. In Baumer, D., Montero, D., and Scanlon, M., editors, *Proceedings of the 25th West Coast Conference on Formal Linguistics*, pages 15–39. Cascadilla Proceedings Project. http://lingref.com/, document 1430, Somerville, MA.

Richtsmeier, P., Gerken, L., Goffman, L., and Hogan, T. (2009). Statistical frequency in perception affects children's lexical production. *Cognition*, 111:372–377.

Taylor, C. V. (1959). *A simplified Runyankore-Rukiga-English and English-Runyankore-Rukiga dictionary*. Eagle Press, Kampala.

Taylor, C. V. (1985). *Nkore-Kiga*. Croom Helm Ltd., Beckenham, England.

Werker, J. and Tees, R. (1983). Developmental changes across childhood in the perception of non-native speech sounds. *Canadian Journal of Psychology*, 37:278–286.