Chaha Phonotactics

Dougal Sutherland

April 2011

Chaha is an Ethio-Semitic language with a complex morphophonological structure. Working from data given in Rose (2007), this paper considers the phonotactic structure of its verb roots, finding particular importance in place and voicing features.

The full data used for this analysis is presented in Appendix A. It is a sample of 48 perfective forms of verbs from Rose (2007), some of which were given with explicit roots, and some of which were guessed based on the rules laid out therein.

It should be noted at the outset that the dataset presented here is on the small side, particularly for computational approaches—Hayes and Wilson (2008) and Adriaans and Kager (2010) used orders of magnitude more distinct types in most of their experiments—and consists only of citation forms of verbs rather than transcriptions of more natural speech. Nevertheless, the data given is sufficient to provide a first look into the phonotactics of Chaha, though it may prove difficult to distinguish some accidental gaps from fundamental restrictions.

Having only dictionary entries, though, is perhaps not as detrimental as it may at first seem. Hayes and Wilson (2008) cite several studies as pointing towards the conclusion that phonological intuitions are better modeled without taking token frequencies into account.

After a brief overview of the morphophonology of Chaha verbs (Section 1), we give an account of the roles of place (Section 2) and voicing (Section 3) features in the distribution of consonants in Chaha root verbs. We then employ phonotactic learning software to see what insight can be gained there (Section 4), and give an overview of what we have learned (Section 5).

1 Verbal Morphology

Chaha's verbs exhibit the templatic morphology typical of Semitic languages. Roots, typically consisting of three consonants, carry the basic semantics, while morphological effects are realized by fitting the consonants of the root into the stem.

For example, the root /k'ms/ 'taste,' when applied to the perfective pattern CəCəCə, becomes /k'əməsə/ 'he tasted.' We are concerned here only with the root and perfective forms, though of course others exist as well.

In addition to the typical triconsonantal Semitic roots, however, Chaha also has lexical classes of verbs which pattern somewhat differently, quadriliteral roots, several types of reduplicating roots, and "weak" roots which affect vowel quality rather than causing a consonant to be present. For details, consult the description by Rose (2007).

We are concerned here primarily with the distribution of root consonants. The vowels of the perfective form are not particularly interesting from a phonotactic standpoint, and there appears to be only one widespread and systematic phonological change between roots and perfective forms, described in Section 1.1 below.

1.1 Consonant Strength

The most apparent phonological change occurring in the perfective form is the mutation of the penultimate consonant, as shown in Table 1.

Root	Perfective	Root	Perfective	Root	Perfective
dβr	dəpərə	mgr	məkərə	zgr	zəkərə
grz	gənəzə	βr (122)	βənərə	zr (1212)	zirəsərə

Table 1: Phonological changes visible in the perfective form. The roots /βr/ and /zr/ are known to be reduplicating, with 122 and 1212 patterns, respectively. Note that in both cases, the change applies to the penultimate constant only.

The first row of Table 1 shows the second consonant of a triliteral root being devoiced in the perfective form. (/β/ is additionally hardened into a stop.) In the second line, we see /r/ being hardened to /n/ in the same contexts. Note the interaction with reduplication here: although we know from other evidence that the roots are reduplicated, the mutation clearly takes place after the reduplication is complete.

Rose (2007) describes the sets of vowels involved in this process as "strong" and "weak," and breaks them up as in Table 2. She claims that the process formerly was realized as gemination, but has taken this form since Chaha has lost overt geminate consonants.

Weak	Strong
β/b w/b^w	p p^w
d ʄ z ʒ	t ĉ s ʃ
g g^w g^j	k k^w k^j
x x^w x^j	k k^w k^j
r	n

Table 2: Chaha consonant series from Rose (2007).

2 Place

The Chaha root forms of this dataset have a clear phonotactic connection to place of articulation. If one considers the place features to consist of `[labial]`, `[coronal]`, and `[dorsal]`, then we have a constraint against two consecutive segments of a root occurring at the same place of articulation. (This applies to the root of verbs, so that, for example, 1233 reduplicated verbs will have their last two consonants share a place of articulation.)

There are two situations where we see verbs in the data violate this constraint: where one of the segments is /r/ or /n/, or in a handful of words where the violating consonant is a glide /w/ or /j/.

The rule is broken frequently by /r/ and /n/, for example in the word /bdr/ → /bətərə/ 'be first.' It appears that this place rule simply ignores /r/ and /n/—two phonemes which are intimately related to one another in Chaha, and have even been argued to be allophones (Rose, 2007). That exception is at least easy to represent featurally: if we make `[+coronal,+sonorant]` a "place of articulation" along with the other three, our constraint stands.

The only other violations in the data occur in words where one of the segments is either /w/ or /j/. We see this with /w/ in the weak roots /fwx/ → /f^wəxə/ and /t'fw/ → /t'əf^{w}ə/. In the former case, /f/ and

	Second Place				
	labial	coronal	r/n	dorsal	p-value
labial	-	10	9	3	.696
coronal	8	-	9	3	.829
r/n	5	10	-	2	.584
dorsal	5	1	10	-	.084

Table 3: The number of transitions from the place on the left to the place on top in the roots of the lexicon, ignoring the few words with multiply-placed roots. The p-value is the probability that transitions out of the first state are skewed with respect to the overall frequency of the second state's occurrence.

/w/ share the `[labial]` feature and /w/ and /x/ share the `[velar]` place; in the latter, /f/ and /w/ again share `[labial]` articulation. The situation with /tjg/, /k'jt/ and /bkj/ is essentially identical.

Weak verbs, however, behave somewhat unusually in Semitic languages, and one could also very well analyze the /w/ or /j/ segment as having been present historically but since deleted. If one completely ignores the glide and e.g. analyzes the root of /fʷəxə/ as /fx/, then, there are no violations.

All of these objections to the co-occurrence constraint have one feature in common: all are sonorants, and indeed they comprise most of Chaha's sonorant inventory. There is one sonorant lacking, however: /m/ never borders a labial sound in its eleven appearances in our lexicon. It has five borders with dorsals, two with /r n/, and seven with other coronals. The probability of a result at least that skewed assuming random distribution between labials, coronals, and dorsals is around .064; if one counts /r n/ as a separate class, it is .039. We thus have reasonably strong evidence to say that the fact that /m/ did not border a labial in our lexicon is due to more than chance, and hence that it is *not* an exception to our place restriction.

A natural follow-up question is: given that we have an essentially inviolable constraint against place repetition, are there preferences for which place follows one another? For example, is a dorsal sound more likely to be followed by /r n/ than one would expect?

We consider the transition probabilities between our four "place" specifications (labial, coronal, /r n/, and dorsal), which are shown in Table 3. The probability that transitions are skewed is shown in the p-value column. More specifically, for each row, the probability of the targets of its transitions (i.e. the entries along the row) being distributed differently from the overall distribution of transitions to that state is calculated using a multinomial likelihood test. The null hypothesis probability of an output state is the proportion of the time that state is chosen as the target of a transition, calculated relative to the other states which are not the source state.[1]

Only the p-value for dorsals is even close to being considered significant. Because of the multiple comparisons problem, however, it is not especially surprising that we see one result with a p-value of 0.84 out of four results. We still cannot reasonably discount the null hypothesis that (beside the restriction on identical places) there is no connection between pairs of place features.

[1]Here and elsewhere, the p test is conducted through Markov Chain Monte Carlo sampling, as the counts observed are too low for the standard χ^2 approximation to apply. Specifically, this is the `chisq.test` function of the R software package (R Development Core Team, 2010), using the `simulate.p.test` argument.

Voicing Pattern	roots given		all	
	root	surface	root	surface
U U U	2	2	2	2
U U V	-	-	1	-
U V U	3	3	11	6
U V V	-	-	2	1
V U U	-	-	2	2
V U V	-	5	5	8
V V U	1	1	3	3
V V V	6	1	8	2
p-value	.027	.041	.004	.021

Table 4: The distribution of the [voice] feature across the triconsonantal roots in the dataset.

3 Voice

The distribution of voicing features among root consonants is less clear in Chaha than that of place features, but also certainly plays a role. Canonical triliteral roots in the sample data have their voicing features distributed as shown in Table 4, where V represents a voiced consonant and U an unvoiced. The closely-related strength features are shown in Table 5, where + indicates [+strong], - means [-strong], and x is for consonants which are neither strong nor weak (such as /m/).

The p-value shown in Tables 4 and 5 is the probability that we would see results at least as skewed towards the patterns shown if there were no phonotactics operating at all, but voiced/unvoiced phonemes just showed up in the same proportion as they are actually observed. (The same applies to strong/weak/neither consonants, of course.) This null hypothesis claims that if $\Pr(\texttt{V}) = \frac{2}{3}$, then $\Pr(\texttt{VUV}) = \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{4}{27}$, and moreover that VVU and UVV hold identical probabilities. More technically, the p-value is the output of a multinomial likelihood test, where the null hypothesis probability vector is determined by assuming independence among the observed class probabilities.

Note that in seven of the eight tests, the odds are extremely poor that the observed patterns are due to chance. This tells us either that there are phonotactic rules related to voicing or strength, or that there are correlated rules defined on another property that happen to cause these patterns in voicing.

There does not appear to be an elegant way to express these rules, however, or at least not that is apparent thus far. In this type of situation, it may be helpful to turn to computational modeling, which can draw out the necessary data to make the elegant patterns (if they exist) more apparent.

4 Computational Simulations

Given the number of possible feature combinations, it is difficult to derive other phonotactic constraints merely by inspection. We turn now to a computational approach to seek other, perhaps less obvious constraints that may be present.

4.1 Models

We tried two freely available phonotactic learning systems on this data: those of Hayes and Wilson (2008) and of Adriaans and Kager (2010). Other widely models widely discussed in the literature, particularly that

Strength Pattern	roots given		all	
	root	surface	root	surface
x x x	-	-	-	-
x x -	-	-	-	-
x x +	-	-	-	-
x - x	-	-	-	-
x - -	1	-	3	-
x - +	-	-	1	-
x + x	-	-	-	-
x + -	-	1	2	4
x + +	-	-	-	-
- x x	-	-	-	-
- x -	-	-	-	-
- x +	-	-	-	-
- - x	-	-	-	-
- - -	5	-	7	-
- - +	-	-	1	-
- + x	-	-	1	-
- + -	-	4	1	6
- + +	-	-	2	3
+ x x	-	-	-	-
+ x -	-	-	3	1
+ x +	2	2	4	2
+ - x	2	1	4	1
+ - -	-	-	-	-
+ - +	1	1	1	1
+ + x	1	2	1	4
+ + -	-	1	-	1
+ + +	-	-	2	1
p-value	.032	.298	.015	.007

Table 5: The distribution of the strength features (see Section 1.1) across the triconsonantal forms of the dataset.

of Albright (2009), seem not to be publicly available.

Adriaans and Kager (2010) develop an Optimality Theory oriented system, which aims to simulate early phonotactic learning by infants, in particular its application to the segmentation of continuous speech into distinct words. These goals proved to be difficult to reconcile with our goals in this investigation. As makes sense for its research goals, this model focuses almost exclusively on finding signs that the speech stream has transitioned from one word to the next (Adriaans, 2011). This is not our goal, and in fact it does not seem particularly possible with the lexicon at hand: initial tests with Adriaans's StaGe model discovered only contiguity constraints and no markedness constraints at all. This not being particularly useful, we turn to the model of Hayes and Wilson (2008).

The model developed by Hayes and Wilson is not designed to be cognitively plausible, but rather simply an "inductive baseline" for the kinds of knowledge that are at all available to a phonotactic learner. In some respects, then, it is more amenable to our aim of learning about Chaha phonotactics without particular regard to how an infant would do so.

This model derives a maximum-entropy phonotactic grammar, which consists of a set of (violable) constraints and associated weights on those constraints. The lower the weight of the constraint, the more expected violations in the corpus, so that rare sequences are penalized as well as those which do not occur at all. The model takes as input a list of phonemes and associated features for those phonemes and constructs *a priori* natural classes of phonemes. It then examines n-gram sequences of natural classes and seeks those that occur less frequently than would be expected if there were no phonotactic constraints, in a manner somewhat similar to that employed in Section 3, but in an automated way and in a far larger search space.

The model also supports searching for constraints on higher levels of the autosegmental tier of Goldsmith (1979). Hayes and Wilson (2008) use the vowel projection to handle a vowel harmony system and a metrical projection to deal with a stress system.

For our simulations, we explored a variety of data sets and with several different sets of parameters. We performed experiments learning both directly on the roots, and also models learning on the perfective forms with a consonantal projection. Some got only triliteral input data, while others got it all. Some models were restricted to biphone probabilities; others could use triphones only where one of the phones contained just the features `[±consonantal, ±word_boundary]`; others could use unlimited triphones. Some could use so-called complement classes to learn logical implications, while others could not. Some models used strength features, some used voicing and nasality.

Accordingly, some simulations finished in minutes, while others ran for several days straight before being abandoned due to time constraints. Most of the output was also quite difficult to interpret and is almost certainly overfitting the data. Given more time, different priors and search settings should be evaluated in order to maximize the applicability of Hayes and Wilson's (2008) learner to particularly small datasets.

The results below are taken from a medley of output grammars and then independently evaluated. Although each grammar taken as a whole does provide a not-unreasonable look at the verbal phonotactics of Chaha, time constraints prevented poring through hundreds of rules for each model to filter out what was overfitting and combine the bigrams and trigrams of natural feature classes into something more interpretable.

4.2 Results

Many of the models discovered a large number of fairly "obvious" constraints, such as noting the lack of a labial approximant.

The models with access to the perfective forms tended to initially learn the constraint

*[+consonantal][+consonantal],

but then weight it at 0 and replace it with e.g. the group of constraints

*[+consonantal][-alveolar]

*[-alveolar][+consonantal]

*[-sonorant][+consonantal]

*[+consonantal][-ejective]

to account for /k'ənt'ə/. (In this case, the second constraint of the group is ranked more highly than the other three combined, which is desirable because it most closely resembles the constraint set out by Rose, 2007—that the second vowel is dropped when a root has /n/ as its second root and a coronal stop as its third.)

The models also come up with many examples that may shed some light in combination with the voicing constraints of Section 3, or might just be accidental:

*[#][p]

*[b,d][#]

*[k, kʷ, p][#]

The complexity of the expression of these constraints in feature terms seems to be an argument for their being accidental characteristics of the small dataset. The last one above, for example, is expressed in the output grammar as

*[-continuant,-ejective,-alveolar,-voiced][+word_boundary]

and each of those features is needed for the constraint to hold, suggesting that this is not really just one constraint working.

More time is needed to sift through the output of these models, to gain a better understanding of how these rules interact with the voicing/strength constraints, and so on. More data—even just the rest of the data from Rose (2007)—might also be quite helpful. Some of this work may be completed in the near future.

5 Conclusions

We have definitively shown one general feature of the phonotactics of Chaha verbs, the place of articulation constraint of Section 2. We did not, however, find any gradient effects there.

We have also shown gradient preferences for certain voicing / consonant strength patterns in Section 3, though an elegant characterization eludes us as of yet.

The computational models of Section 4 have shown some promise, but more work is needed to thoroughly understand their output, particularly in connection with the voicing constraints.

This treatment has focused exclusively on the root and perfective forms of verbs. It is likely that other verb forms, let alone other parts of speech, will have somewhat different phonotactics; Smith (2001) demonstrates that nouns often have much more lenient phonological restrictions than do verbs. Nevertheless, the same basic concepts (particularly of place and of voicing) will likely be relevant to all of Chaha phonotactics, and essentially the same approach should prove profitable in tackling the rest of the language once this aspect has been more fully understood.

References

Adriaans, F. (2011). *The Induction of Phonotactics for Speech Segmentation: Converging evidence from computational and human learners.* Phd dissertation, Universiteit Utrecht, Utrecht, The Netherlands.

Adriaans, F., & Kager, R. (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language*, *62*(3), 311–331.

Albright, A. (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology*, *26*(1), 9–41.

Goldsmith, J. (1979). *Autosegmental Phonology.* New York: Garland.

Hayes, B., & Wilson, C. (2008). A Maximum Entropy Model of Phonotactics and Phonotactic Learning. *Linguistic Inquiry*, *39*(3), 379–440.

R Development Core Team. (2010). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. Available from `http://www.r-project.org`

Rose, S. (2007). Chaha (Gurage) Morphology. In A. Kaye (Ed.), *Morphologies of Asia and Africa* (pp. 399–424). Eisenbraums.

Smith, J. (2001). Lexical category and phonological contrast. In R. Kirchner, J. Pater, & W. Wikely (Eds.), *Papers in experimental and theoretical linguistics: Workshop on the lexicon in phonetics and phonology* (Vol. 6, pp. 61–72). Edmonton: University of Alberta.

A Data

The data used in this analysis, from Rose (2007). Note that this data corrects a few errors in the assignment statement, as well as eliminating two duplicate verbs and hypothesizing roots for the majority of verb forms.

Perfective	Type	Special?	Root	Perfective	Type	Special?	Root
k'əməsə	A		**k'ms**	kɨrət'əmə	A	4	**k'rt'm**
gənəzə	A		**grz**	t'əβət'ə	A		**t'βt'**
nəpərə	A		**rβr**	məkərə	A		**mgr**
sənəfə	A		**srf**	kəfətə	A		**kft**
dəpərə	A		**dβr**	manəxə	C		mrx
banərə	C	122	br	zapətə	C		zpt
ĵəpərə	B		ĵpr	gʲənəzə	B		gʲrz
məkʲərə	B		mkr	met'ərə	B		mtr
k'ʷəmərə	D		k'mr	bʷənəsə	D		brs
mɨsəkərə	A	4	mskr	gɨrətəmə	A	4	grdm
kətəfə	A		**ktf**	bətərə	A		**bdr**
nədəfə	A		**ndf**	zəkərə	A		**zgr**
tənəfə	A		trf	ĉ'ənəfə	B		ĉ'rf
anəβə		a	arβ	dak'ə		a	dak'
səma		a	sma	wət'ək'ə		w	wt'k'
fʷəxə		w	fwx	t'əfʷə		w	t'fw
tegə		j	tjg	bəkʲə		j	bkj
fet'ət'ə		122	**ft'**	bazəzə		122	**βz**
k'ʷənərə		122	**k'ʷr**	k'ɨmət'ət'ə		1233	kmt'
fɨk'əfək'ə		1212	fk'	dɨrəzəzə		1233	drz
nɨsənəsə		1212	nsn	sɨrətətə		1233	srt
zɨrəsərə		1212	zr	fɨrək'ək'ə		1233	frk'
sɨrəsərə		1212	sr	bənərə		122	**βr**
k'ənt'ə	A		k'nt'	kʲ'ətə		j	k'jt

Table 6: The data used in this paper. Bold roots were given explicitly; others were guessed according to the rules of Chaha morphology given by Rose.