Modeling OCP-Place in Amharic with the Maximum Entropy phonotactic learner

Rebecca S. Colavin, Roger Levy and Sharon Rose
University of California, San Diego

1 Introduction

One essential part of a native speaker's knowledge is the characterization of what logically possible sound sequences constitute legitimate possible words in the speaker's language, or phonotactics. Although formal phonotactic models were originally categorical—classifying every sound sequence dichotomously as either being a well-formed possible word or not (Chomsky & Halle, 1968) the recognition that well-formedness judgments are gradient has driven the more recent development of *probabilistic* models of phonotactics (Albright, in prep; Hayes & Wilson, 2008). These models involve two substantive commitments: first, a speaker's phonotactic knowledge is encoded as a probability distribution over sound sequences, with higher probability implying greater well-formedness; second, this probability distribution is hypothesized to be in substantial alignment with the lexicon, so that learning a language's phonotactics can be understood as inferring a probability distribution that assigns high likelihood to the lexicon. Since any speaker's experience with a lexicon is finite, however, a probabilistic phonotactic model must also include a propensity toward *generalizing* to novel sequences, assigning higher probability to some nonce words than to others. A model of English phonotactics must thus assign higher probability to [bwad] than to [bnad], for example, for the former sequence is preferred to the latter by native speakers, although both are nonce words (Albright, in prep). In this case, the distinction can be understood as a preference at the representational level of phonological features: namely, that [-continuant][+nasal] sequences do not appear in English, while [-continuant][+approximant] features do. Hence understanding phonotactics requires investigation into the representational basis for phonological generalization.

This paper reports an investigation into phonological generalization and its representational basis for the case of the Obligatory Contour Principle as applied to place of articulation (OCP-Place) in Amharic, a Semitic language of Ethiopia. OCP-Place in Amharic applies to the co-occurrence of consonants within the verb root. This constraint poses modeling challenges at three levels. First, relative to other case studies in phonotactics such as English onsets, Amharic verb roots are long and complex, consisting minimally of three consonants, and sometimes four or five. Second, OCP-Place is gradient, with violations rare but attested. Third, there are a large number of verbs in Amharic with identical consonants. Identical consonants could be analyzed as surface-true, and therefore included in the assessment of phonotactic restrictions. On the other hand, identical consonants could be analyzed as lexically reduplicated, in which case the copy would be absent from the lexical representation of the root and not included in the assessment of OCP-Place phonotactics. Here, we use the Hayes and Wilson (2008) Maximum Entropy (Maxent) Phonotactic Learner to model the acquisition of phonotactic knowledge in Amharic, training the model on a lexicon of Amharic verb roots, and testing it both on its ability to accurately encode the contents of the lexicon and on its ability to predict native speaker well-formedness judgments of nonce verbs. We compare automatically learned phonotactic grammars with hand-coded grammars specifically designed to encode OCP-Place. We find that an automatic learner with access to information about lexical reduplication is able to discover constraints more effectively than a learner with access to only surface lexical representations, underscoring the importance of data representation in the acquisition of phonotactic knowledge.

2 Phonotactics and experimental testing

Experimental work has shown that there is a connection between the lexical probability of the sound sequences that compose a nonce word (a manifestation of phonotactic restrictions) and gradient speaker judgments of word acceptability.

2.1 Experimental studies and probabilistic models

Experimental studies that evaluate gradient speaker judgments for nonce words that controlled for phonotactic probability have primarily focused on specific sub-parts of English words. For example, Scholes (1966), Coleman, (1996), Albright and Hayes (2003), Hay et al. (2003), Treiman et al. (2000) and Albright (in prep) have collected English speaker judgments for nonce words that differed in the probability of their onsets, rimes or consonant clusters. Judgments for probabilistically controlled whole nonce words are somewhat rarer (Vitevitch et al. 1997, Frisch et al. 2000). Research focusing on whole words in other languages include Frisch & Zawaydeh (2001) for Arabic, Myers & Tsay (2005) for Mandarin, and Kirby & Yu (2007) for Cantonese.

Probabilistic phonotactic models use lexical probabilities to predict speaker judgments. These models learn by assigning a high likelihood to the lexicon, and define probability distributions over sounds and sound sequences. Those probability distributions can then be used to predict the acceptability of nonce forms. Currently, the model that performs the best in predicting speaker judgments is the Hayes and Wilson (2008) Maximum Entropy phonotactic learner. For English onsets, the correlation between the predictions of the Maxent learner and speaker judgments (Scholes 1968) is remarkable: r = .946 (Pearson's correlation). The judgments were based on a binary legal-illegal task rather than a gradient judgment task, which may account for the high correlation. It is important to test the model on more challenging data.

2.2 The Maximum Entropy phonotactic learner

The power of the Maxent learner compared to earlier models, is that it uses a phonologically motivated representation of distinctive features and natural classes to compose a Maxent weighted grammar of constraints. A Maxent grammar containing constraints C_i each with weight w_i assigns a probability distribution over the set Ω of possible sound sequences in the language, according to the following formula:

$$(1)\quad \mathbf{P}(x) = \frac{\exp[-\mathbf{Maxent\,value}(x)]}{\mathbf{Z}},$$

where the Maxent score for a logically possible sound sequence *x* is defined as

$$(2)\quad \mathbf{Maxent\,value} = \sum_i w_i C_i(x)$$

the value $C_i(x)$ is the number of violations of C_i in *x*, and the normalizing constant *Z* is defined as follows:

$$(3)\quad Z = \sum_{x' \in \Omega} \exp[-\mathbf{Maxent\,value}(x')]$$

This representation enables the model to make generalizations that would be impossible with a less fine-grained representation. Returning to our example of [bwad] versus [bnad]-a segmental bigram model would assign equal (zero) probability to the onsets [bw] and [bn], since neither bigram appears in an English onset. However, the Maxent learner has access to the natural class representations and thus can learn that other stop-[+approximant] sequences do occur in the lexicon but stop-[+nasal] sequences do not. It would therefore assign a higher penalty to [bn] than [bw].

Given a segment inventory defined in terms of distinctive features, the Maxent learner creates the set of all possible natural class sequences defined for the language (where the maximal length of the sequences is a researcher-defined parameter of the model). This set of natural class sequences defines the set of potential constraints. The Maxent Learner then iterates over the following processes:

1. Selecting a constraint from the set of potential constraints and adding it to the grammar.
2. Reweighting the new constraint set according to the principle of Maxent.

We elaborate on each of these in turn.

2.2.1 Weighting the grammar

For weighting the constraints in the grammar, the Maxent learner proceeds similarly to other Maxent models. The goal is to assign a penalty weight to each constraint in the grammar such that the probability of the learning data *D* is maximized. There is no single step method for determining the set of constraint weights that maximize P(*D*), so an iterative hill climbing algorithm is used.

2.2.2 Constraint selection

The model selects a candidate constraint from the set of potential constraints that represents the most under-represented sequence given the current grammar. The model estimates O(C_i)/E(C_i), the ratio between the number of Observed and Expected violations, for each of the candidate constraints C_i via Monte Carlo sampling. O/E values between 0 and 1 indicate that a sequence is under-represented. The model first selects the most *accurate* constraints. This is the set of constraints that have O/E values below a certain threshold with an adjustment (Mikheev, 1997) such that for the same O/E values, those with high Expected values are more salient than those with lower Expected values. Then, within that set, a heuristic is used to select the most *general* constraint. Shorter constraints are considered more general than longer ones (so *[+voice] would be considered more general than *[+voice][+voice]) and constraints described with few natural classes are considered more general than those that have many ([+voice] is therefore more general than [+voice, -continuant]).

3 Simulations

Although the Maxent learner performs extremely well in predicting the acceptability of unattested onsets in English, it has not been evaluated against speaker judgments across a word or for languages other than English[1]. The goal of our simulations is therefore to evaluate the performance of the Maxent learner for words in a language with different representational issues and phonotactic constraints that those of English onsets.

3.1 OCP-Place in Amharic

We evaluated the performance of the Maxent learner on the consonant phonotactics of verb roots in Amharic, a Semitic language of Ethiopia. Like other Semitic languages, Amharic verb roots are subject to OCP-Place (Greenberg 1950, Bender and Fulass 1978, McCarthy 1986, 1988, Pierrehumbert 1993, Buckley 1997, Frisch et al. 2004, Rose and King 2007): consonants with the same place of articulation co-occur less frequently within roots than would be expected, all else being equal. Although a hallmark of Semitic languages, OCP-Place has also been identified as a restriction in Russian (Padgett 1995), Javanese (Mester 1986) English (Berkley 1994) and Muna (Coetzee and Pater 2008).

OCP-Place is considered a gradient restriction because it may be violated. For Amharic, as for other Semitic languages, gradiency is expressed over two dimensions: place of articulation and location of violation. Regarding place of articulation, homorganic consonants are under-represented in Amharic verb roots

[1] Hayes and Wilson (2008) modeled the phonotactics of Wargamy and Shona, but did not compare the predictions of the models to speaker judgments. Pater & Kager (2010) have applied the MaxEnt learner to Dutch rimes and tested speaker judgments.

but the co-occurrence of dorsal consonants is rarest, followed by labials and then coronals (for coronals, the largest group, the restriction operates over coronals having the same manner of articulation). This holds for longer verb roots, too. For location of violation, verb roots with OCP-Place violations at the left edge of the root (C_1C_2X) are rarest, followed by right-edge violations (XC_2C_3), with non-adjacent violations being the least rare (C_1XC_3).

This expression of OCP-Place over verb roots poses a particular challenge to the Maxent learner. To encode the gradiency over location of violation requires constraints of length 3 (to distinguish left-edge, right-edge and non-adjacent violations) but the Maxent learner is designed to favor shorter constraints. For example, the model might select a constraint over adjacent segments of length 2 *[Dorsal][Dorsal] rather than the longer *[word boundary][Dorsal][Dorsal] and this may well affect the performance of the model in predicting gradiency according to location of violation. In addition, Amharic has roots with four, five and even six consonants.

3.2 Identical consonants

Amharic has numerous roots containing identical consonants (36% of the roots in our database). There are several common patterns in Ethio-Semitic languages such as 122, 1212, 1233 or 12323 (where numbers correspond to consonants), as well as other less common types. These patterns are usually analyzed as a single consonant that is spread to another position (McCarthy 1979, 1983) or as reduplication (Buckley 1990, Hudson 1995, Rose 1997, Gafos 1998) or both. However, these are cases of lexical or phonological reduplication, rather than a productive morphological derivational pattern. Two approaches could be taken to the representation of identical consonants in this study. One, they could be analyzed as is, treating the distribution as lexical and static. This method has the advantage of being surface-true, but by so doing, it is necessary for OCP-Place to be stated as a restriction over homorganic, but non-identical consonants, as identical consonants would be overrepresented in particular positions, thereby undermining OCP-Place. Berent & Shimron (2003) have shown that Hebrew speakers judge identical and homorganic consonants differently. Two, they could be analyzed according to the standard phonological analysis (McCarthy 1979, 1983), treating them as containing a single underlying consonant and assuming a process of reduplication or spreading.

As there are arguments in favor of both of these representations, we performed two parallel series of simulations: one, the Identical approach, assumes that identical consonants are qualitatively the same as other consonants, and two, the Reduplication approach, assumes that only one consonant is included in the verb root and subject to evaluation of OCP-Place.

3.3 Methodology

The following section describes the training data, test data, procedures and evaluation for each of the two simulations.

3.3.1 Training data

In both simulations, models were trained on 4243 verbs (in 3ms perfective citation form) drawn from Kane (1990) containing an inventory of 25 consonants shown in table 1. Because of software limitations, labialized consonants were confounded with their non-labialized counterparts. This does not affect the calculation of OCP-Place, since it operates over primary place of articulation.

Place	Segments
Labial	p’, b, f, m, w
Coronal	stops, affricates: t, t’, d, ʧ, ʧ’, ʤ fricatives: s, s’, ʃ, z, ʒ sonorants: n, ɲ, r, l, j
Dorsal	k, k’, g
Glottal	h

Table 1: Segment inventory used in training data

The training data included not only the standard Semitic tri-consonantal roots, but also weak roots (those appearing to lack a surface root consonant), roots with 4, 5 or 6 consonants and roots with identical consonants. Roots with 3 or more consonants and no identical consonants were encoded directly as they appear in the dictionary, but the encoding of weak roots and roots with identical consonants required some analytical decisions.

Weak roots. Weak roots in Semitic languages are those roots that fail to display the canonical number of surface consonants, and are typically analyzed as containing a glide (see e.g. Kaye 2007). In Ethio-Semitic languages, the historical presence of a glide is marked in the surface form by a front vowel (ex. hedə ‘to go’ <hjd) or the palatalization of the preceding consonant (ex. rəʧʧ’ə ‘to sprinkle <rt’j) (Hudson, 1979). In addition, in languages like Amharic that have lost guttural consonants, the historical presence of a guttural is marked by [a] (Unseth, 2002), (ex. gəbba ‘to enter’ and lakə ‘to send’ are historically derived from *gbʔ, and *lʔk respectively). Speakers may determine the location of the missing surface consonant due to these clues and the location of morphological gemination on the penultimate root consonant. We included weak roots in the training data, as OCP-Place still restricts the other consonants. However, weak roots were encoded with the place holder 'X' in the position of the consonant

missing from the expected root position. For example, the weak root gəbba is encoded as [gbX] and the weak root hedə as [hXd]. For the MaxEnt learner we gave 'X' a single distinctive feature, 'x', and left it underspecified for all other features; all other segments are underspecified for 'x'.

Identical consonant roots. In the Identical simulation, roots such as zərətt'ət'ə 'to trip up by entangling the feet' which follows the 1233 pattern, are encoded with identical consonants: /zrt't'/. In the Reduplication simulation, the same root would be encoded as /zrt'X/, using a featureless placeholder X for the copied consonant in the same manner as for weak roots. The placeholder method allows reference to the templatic pattern and word edge employed by the root, in this case a quadriconsonantal pattern, and keeps it distinct from an unrelated triconsonantal root, in this case, /zrt'/ for the verb zərrət'ə 'to insult; to be stunted'.

Table 2 summarizes the encoding choices and the contrast between the representation of identical consonant roots in each of the two simulations. The shaded areas indicate examples where the encoding for simulation 1 is different from simulation 2.

Root type	Example	root	sim 1	sim 2
surface true	bəggənə "get furious"	bgn	bgn	bgn
weak	awwədə "perfume"	wd	Xwd	Xwd
reduplicative	bədəbbədə "beat"	bd	bdbd	bdXX

Table 2: Root types and representational encoding

3.3.2 Test data

Computational phonotactic models require data against which the predictions of the model can be compared. We evaluated all model predictions against a set of speaker judgments of nonce verb forms previously collected in Ethiopia by the third author[2]. A set of triconsonantal nonce verb stimuli were created based on consonant distributions in a dictionary study (Rose and King 2007). The nonce verbs, which used only the 14 most frequent and evenly distributed consonants, contained 90 forms with OCP-Place violations representing a range of predicted acceptability according to location of violation (left-edge, right-edge, non-adjacent) and place of articulation (dorsal, coronal, labial). Twenty native speakers of Amharic were asked to rate the nonce forms (all conjugated

[2] This study was developed with Lisa King and data were collected in 2001 in Addis Ababa, Ethiopia.

identically as CəC:əCə) on a 1-6 scale with 1 = very Amharic-like and 6 = not like Amharic at all. Speakers significantly dispreferred nonce forms with OCP violations over controls (f-test: p<.0001) – see figure 1.

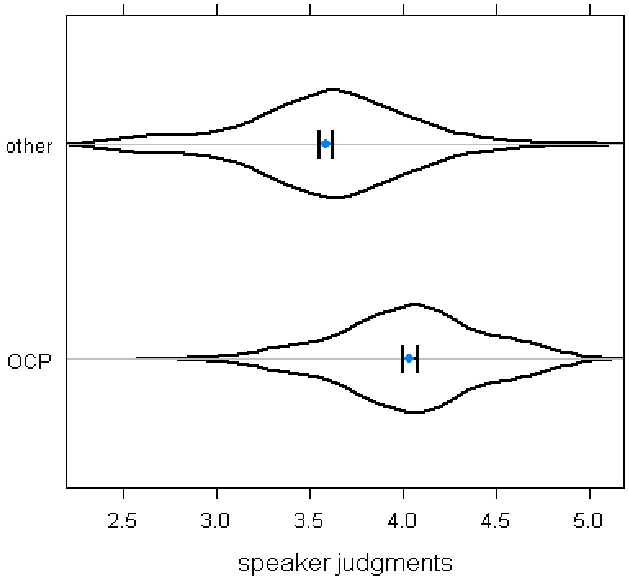

Figure 1: Well-formedness of OCP-violating and non-OCP-violating Amharic verb roots. Curves show estimated judgment distributions; dots and bars are means and standard errors.

3.3.3 Procedures

For each of the two simulations, two models were generated:

i. an automatically learned model with 1000 constraints
ii. a hand-written model designed to determine whether the automatic constraint selection is optimal by comparing the automatic model to a Maxent weighted hand-written grammar.

For both simulations, the hand-written grammar describes the OCP-Place restriction using constraints that are available to the Maxent learner. That set of constraints is then weighted according to the principle of Maximum Entropy. Because of the differing representation of identical consonants, the hand-written grammars were different for simulation 1 and simulation 2:

i. In Simulation 1 (Identical), the hand-written grammar encodes OCP-Place as a restriction over homorganic, but *non-identical*, consonants.
ii. In Simulation 2 (Reduplication), the hand-written grammar describes the more general restriction over homorganic consonants.

Table 3 illustrates the difference between the two hand-written grammars. Note that in simulation 1, the restriction must be described specifically for each individual segment, whereas in simulation 2, the restriction is described generally for each place of articulation.

	condition	Expression of constraint
Simulation 1 Identical: constraints describing OCP-Place restriction for 'b'	left-edge (C_1C_2X)	*[word boundary][b][Labial not b] *[word boundary][Labial not b][b]
	right-edge (XC_2C_3)	*[b][Labial not b][word boundary] *[Labial not b][b][word boundary]
	non-adjacent (C_1XC_3)	*[Labial not b][not Labial][b] *[b][not Labial][Labial not b]
Simulation 2 Reduplication: constraints describing OCP-Place for all Labial consonants	left-edge (C_1C_2X)	*[word boundary][Labial][Labial]
	right-edge (XC_2C_3)	*[Labial][Labial][word boundary]
	non-adjacent (C_1XC_3)	*[Labial][word boundary][Labial]

Table 3: Constraint examples

Table 3 shows that describing a restriction over non-identical homorganic consonants requires more constraints than the general restriction over homorganic consonants. Furthermore, the requirement that grammars use only those natural classes pre-defined by the model imposes further limitations and the final hand-written grammar for simulation 1 contained 384 constraints compared to just 27 for simulation 2. In both simulations, automatically selected constraints were added to the hand-written models until grammar size reached 1000 constraints.

3.3.4 Model evaluation

For both simulations, the automatic model and the hand-written model were evaluated incrementally as constraints were added to the grammar. For the automatic models, each model was evaluated after the acquisition of every new constraint until the grammar size reached 100, and every 20 constraints thereafter. The hand-written models were first evaluated after all the hand-written constraints had been incorporated into the model, and then incrementally in the same manner as for the automatic models.

Grammars were evaluated by (a) computing the log-likelihood (using five-fold cross-validation) of the training data, and (b) evaluating the correlation between maxent scores assigned by the grammar (see figure 2 above) and native speaker well-formedness judgments of nonce verbs.

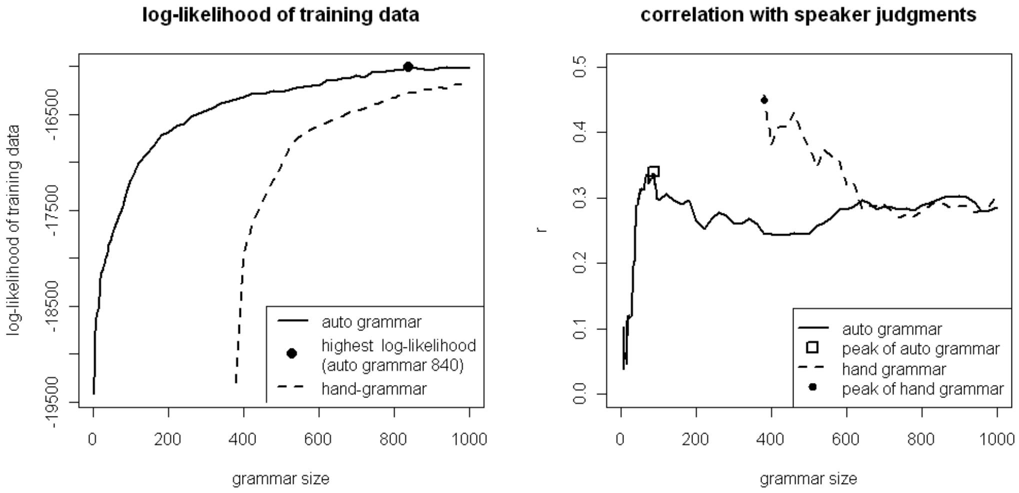

Figure 2: Results from simulation 1

3.4 Results

3.4.1 Simulation 1: Identical

Figure 2 shows the cross-validated log-likelihood of the training data for both the automatic and hand-written models, and the Pearson's correlations between speaker judgments and model Maxent scores. First, we note that the 384 hand-written constraints describing the OCP-Place restriction do not raise the overall log-likelihood of the model as much as the constraints selected early on by the automatic grammar. Second, the log-likelihood of the hand-written model is always lower than the corresponding automatic model with the same number of constraints. Inspection of the constraints that correspond to a sharp rise in log-likelihood shows a large number of constraints over rare and positionally restricted segments, as well as OCP-Place type constraints. For example, both models quickly assigned high weights to constraints against the rare segment [p'] and root-final palatal consonants (for diachronic reasons, palatal consonants are rare in final position[3]).

Overall, the log-likelihood tends to show that the model is performing as we would expect: as constraints are added to the grammar, there is a sharp rise in the log-likelihood and a tapering off as there are fewer powerful constraints to be found. Even with 1000 constraints, there is no evidence of overfitting as there is no fall in log-likelihood.

[3] For the complete grammars, see idiom.ucsd.edu/~colavin/Amharic_results.pdf.

The most important point regarding model/speaker agreement is that the peak in the hand-written model at 384 constraints (r = .45) is higher than the peak in correlation for the automatic model (r = .34 at 80 constraints). This difference is statistically significant (p < .01, with non-parametric bootstrapping). This implies that automatic constraint selection is less optimal than a phonologically motivated grammar that directly encodes the OCP explicitly.

An unexpected result is that for the hand-written model, the correlation with speaker judgments falls off immediately as new constraints are added after the highest peak. This fall in correlation between the hand-written grammar and the speaker judgments between 384 and 1000 constraints is significant (p< .01, with non-parametric bootstrapping).

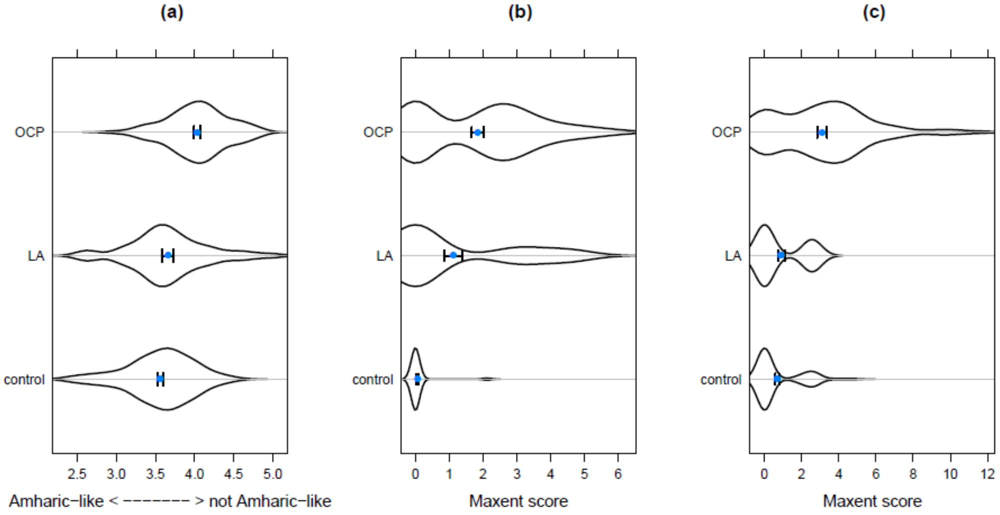

Figure 3: (a) averaged speaker judgments; (b) simulation 1: best automatic model; (c) simulation 2: best automatic model

An analysis of automatically selected constraints in both models shows that the encoding of lexically reduplicated consonants in a manner analogous to ordinary consonants is problematic. One problem arises from Laryngeal Agreement. Laryngeal Agreement (LA) is a restriction requiring stops to agree in the laryngeal features of voice and constricted glottis ([cg]) (Rose & Walker 2004). For example, roots such as [tgr] (stops disagree in voice) and [ftk'] (voiceless stops disagree in [cg]) occur less frequently than would be expected, all else being equal). In Amharic, LA is a weak restriction both statistically (Rose and King 2007) and in speaker judgments. The judgment task data investigated Amharic speaker sensitivity to both OCP-Place violations and LA violations and found that speakers showed no significant dis-preference for nonce forms with LA violations. The results of the judgment task are shown in the violin plot in figure 4a. Note the similarity between the ratings for controls and for LA violations. This result contrasts with figure 4b which shows predicted acceptability ratings for the best automatic model (82 constraints). The predictions for LA violations are clearly stretched more to the unacceptable range than for the speaker judgments.

This over-estimation of the unacceptability of nonce forms with LA violations is directly related to the presence of identical consonants in our training data. Consider the case of the roots [mlt] and [fnk'] and their possible counterparts with identical consonants [mltt] and [fnk'k']. The repeated stops agree in voice and [cg], strengthening adherence to LA; this issue does not arise in the representation without reduplication.

3.4.2 Simulation 2: Reduplication

Turning to the reduplication simulation, figure 4 shows the log-likelihood and correlation with speaker judgments for the automatic and hand-written models, with several points of interest. For log-likelihood:

i) after the hand-written grammar is incorporated into the model, the general shape of both models is very similar. In both cases, the sharp rise in log-likelihood corresponds in many cases to the automatic acquisition of constraints over generally rare segments and positionally restricted segments.

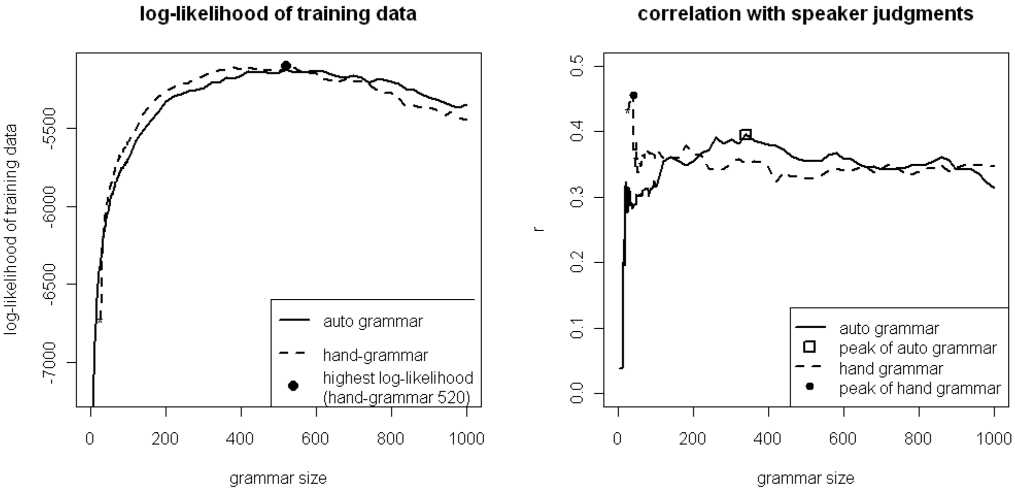

Figure 4: results for simulation 2

ii) for both models, the log-likelihood appears to fall once the grammar size exceeds 550 constraints. This is likely a sign of over-fitting.

iii) contrary to the first simulations, the log-likelihood of the hand-written model is not consistently lower than that of the automatic model.

Figure 4 shows the results for the correlation between model predictions and speaker judgments. Note that the automatic model peaks much later than the hand-written model. The peak of correlation for the hand-written model (r = .45 with 40 constraints) is not significantly better than that of the automatic model (r = .37 at 380 constraints).

This improvement in the automatic model, a direct result of the change in the encoding of reduplicated consonants, is illustrated in Figure 7 which shows the speaker judgments for LA violations and the best automatic model from simulation 2. In simulation 1, model predictions for roots with LA violations were stretched rightwards, to the more unacceptable range, where in simulation 2, the predictions for LA violations are similar to controls.

A second point of interest is the fact that the most predictive model (the hand-written model completed to 40 constraints, r=.45) is significantly (p <.01) better than the model with the highest log-likelihood (hand-written model completed to 520 constraints, r=.33). This result must be interpreted with prudence because although it appears to call into question the linking function between speaker judgments and the log-likelihood of the data, it may well be an artifact of the numerous assumptions that we have made in representing the data. Most obviously, we have assumed that speakers base their judgments of nonce verb roots exclusively on the lexicon of verb roots, and our test data examines only a subset of the full range of possible phonotactic restrictions for the language.

3.5 Discussion

The qualitative improvement in the automatic model's performance over that of Simulation 1 demonstrates that access to information about lexical reduplication allows the learner to generalize from the contents of the Amharic lexicon both more efficiently and more faithfully to the generalizations made by native speakers. Although the hand-written grammar still ultimately achieves a higher raw correlation with native speaker judgment than the automatic grammar, this difference is no longer significant. Furthermore, access to information about lexical reduplication eliminates the discrepancy between the ability of the hand-written versus automatic grammars to achieve high cross-validated log-likeihood on the lexicon.

4. Conclusion

Our study shows that assumptions about data representation have significant consequences for phonotactic modeling. The problem of representation is particularly acute for languages such as Amharic where the definition and behavior of a lexical item can be different from English. Here we considered the role of identical consonants in speaker judgments of OCP-Place, but there are many other variables that have yet to be investigated: root productivity, the number of non-verbs associated with the same verb root, or words without obvious roots. Finally, our study also highlights the importance of speaker judgment data in the evaluation of phonotactic models. The excellent results of Hayes and Wilson in modeling English onsets are based mostly on the binary legal-illegal quality of the Scholes (1986) judgment data that was used to evaluate the model. Our results indicate that speakers are sensitive to OCP-Place violations, but the fit with model predictions is not as strong. Further judgment data designed to explicitly focus on strong distributional restrictions in the language may lead the way to yet better models of human phonotactic knowledge.

References

Albright, Adam. In prep. Natural classes are not enough: Biased generalization in novel onset clusters. MIT

Albright, Adam, and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: computational/experimental study. *Cognition* 90:119–161.

Bailey, Todd M., and Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods. *Journal of Memory and Language* 44(4):568–591.

Bender, M. Lionel and Hailu Fulass. 1978. *Amharic Verb Morphology*. East Lansing, Michigan: The African Studies Center, Michigan State University.

Berent, Iris & Joseph Shimron. 2003. Co-occurrence restrictions on identical consonants in the Hebrew lexicon: Are they due to similarity? *Journal of Linguistics* 39(1),:31-55.

Berkeley, Deborah. 1994. The OCP and gradient data. *Studies in the Linguistic Sciences* 24.59-72.

Buckley, Eugene. 1990.Edge-in association and OCP 'violations' in Tigrinya. *Proceedings of the Ninth West Coast Conference on Formal Linguistics*, pp. 75-90.

Buckley, Eugene. 1997. Tigrinya root consonants and the OCP. *Penn Working Papers in Linguistics* 4.3, 19-51.

Chomsky, M and M Halle. 1968. *The Sound Pattern of English*. Harper and Rowe, New York.

Coetzee, Andries. and Joe Pater. 2008. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language and Linguistic Theory* 26: 289-337.

Coleman, J. S. 1996. The psychological reality of language-specific constraints. Paper presented at the Fourth Phonology Meeting, University of Manchester, 16-18 May 1996.

Frisch, Stefan A. 1996. Similarity and frequency in phonology. Ph.D dissertation, Northwestern University.

Frisch, Stefan A., Nathan R. Large, and David B. Pisoni. 2000. Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language* 42:481–496.

Frisch, Stefan A., Janet B. Pierrehumbert, and Michael Broe. 2004. Similarity avoidance and the OCP. Natural Language and Linguistic Theory 22:179–228.

Frisch, Stefan A., and Bushra A. Zawaydeh. 2001. The psychological reality of OCP-place in Arabic. *Language* 77:91–106.

Gafos, Diamandis. 1998, Eliminating long-distance consonantal spreading. *Natural Language and Linguistic Theory* 16:2, pp. 223-278.

Greenberg, Joseph H. 1950. The patterning of root morphemes in Semitic. *Word* 5.162- 18 1

Hammond, Michael. 2004. Gradience, phonotactics, and the lexicon in English phonology. *International Journal of English Studies* 4:1-24.

Hay, Jennifer, Janet B. Pierrehumbert, and Mary Beckman. 2003. Speech perception, well-formedness and the statistics of the lexicon. In J. Local, R. Ogden and R. Temple (eds.), *Phonetic Interpretation. (Papers in Laboratory Phonology No. 6*) (58-74). Cambridge: Cambridge University Press.

Hayes, Bruce, and Wilson, Colin. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39:379-440

Hudson, Grover. 1979. The Principled Grammar of Amharic Verb Stems. *Journal of African Languages and Linguistics*. Volume 7, Issue 1, Pages 39–58

Hudson, Grover. 1995. Phonology of Ethiopian languages, The *Handbook of Phonological Theory*, John Goldsmith, ed., 782-797 (chapter 27). New York: Blackwell.

Kane, Thomas Leiper, 1990. *Amharic-English Dictionary*. Wiesbaden: Otto Harrassowitz.

Kaye, Alan. 2007. Arabic Morphology. In A. Kaye (ed.) *Morphologies of Asia and Africa*, 211-247. Eisenbrauns,.

Kirby, James, and Alan Yu. 2007. Lexical and phonotactic effects on wordlikeness judgments in Cantonese. In J. Trouvain, & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetics Science,* (pp. 1389–1392). Dudweiler, Germany: Pirrot.

McCarthy, John. 1979. Formal Problems in Semitic Phonology and Morphology. Doctoral Dissertation, MIT, Cambridge, MA.

McCarthy, John. 1983. Consonantal morphology in the Chaha verb. *Proceedings of the West Coast Conference on Formal Linguistics 2*. M. Barlow, D. Flickinger, and M. Wescoat (eds). Stanford, CA: Stanford Linguistics Association.

McCarthy, John. 1986. OCP effects: gemination and antigemination. *Linguistic Inquiry* 17:207-263.

McCarthy, John. 1988. Feature geometry and dependency: a review. *Phonetica* 45:84–108.

Mester, R.A. 1986. Studies in Tier Structure. Doctoral Dissertation, U. Mass, Amherst.

Mikheev, Andrei. 1997. Automatic rule induction for unknown word guessing. *Computational Linguistics* 23:405–423.

Myers, J., & Tsay, J. 2005. The processing of phonological acceptability judgments. *Proceedings of Symposium on 90-92 NSC Projects* (pp. 26-45). Taipei, Taiwan, May.

Ohala, John J., and Manjari Ohala. 1986. Testing hypotheses regarding the psychological reality of morpheme structure constraints. In *Experimental phonology*, ed. John J. Ohala and Jeri J. Jaeger, 239–252. San Diego, Calif.: Academic Press.

Padgett, Jaye. 1995. Partial class behavior and nasal place assimilation. In *Coyote Working Papers in Linguistics*. Tucson: University of Arizona.

Pierrehumbert, J. 1993. Dissimilarity in the Arabic Verbal Roots, *Proceedings of the 23rd Meeting of the Northeastern Linguistic Society*, Graduate Student Association, U. Mass. Amherst. 367-381.

Rose, Sharon. 1997. Theoretical issues in comparative Ethio-Semitic phonology and morphology. Unpublished PhD, McGill University, Montreal.

Rose, Sharon. 2000. Multiple correspondence in reduplication. *Proceeding of the 23rd Annual Meeting of the Berkeley Linguistics Society*, ed. by M. Juge & J. Moxley, 315-326.

Rose, Sharon & Rachel Walker. 2004. A typology of consonant agreement as correspondence. *Language* 80:475-531.

Rose, Sharon and Lisa King. 2007. Speech error elicitation and co-occurrence restrictions in two Ethiopian Semitic languages. *Language and Speech* 50:451-504.

Scholes, Robert. 1966. *Phonotactic grammaticality*. The Hague: Mouton.

Treiman, Rebecca, Brett Kessler, Stephanie Knewasser, Ruth Tincoff, and Margo Bowman. 2000. English speakers' sensitivity to phonotactic patterns. In *Papers in laboratory phonology v: acquisition and the lexicon*, ed. Michael B. Broe and Janet Pierrehumbert, 269–282. Cambridge: Cambridge University Press.

Unseth, Peter. 2002. Bi-Consonantal Reduplication in Amharic and Ethio-Semitic. PhD dissertation, the University of Texas at Arlington.

Vitevitch, Michael S., Paul A. Luce, Jan Charles-Luce, and David Kemmerer. 1997. Phonotactics and syllable stress: implications for the processing of spoken nonsense words. *Language and Speech* 40:47–62.

Vitevitch, M. & P. Luce. 1999. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language* 40:374-408.

Vitevitch, MS, Luce, PA, Pisoni, DB, et al. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language*, 68(1-2), 306-11.