Accidental gaps and surface-based phonotactic learning: a case study of South Bolivian Quechua

Colin Wilson (JHU) & Gillian Gallagher (NYU)

The lexicon of a natural language does not contain all of the phonological forms that are grammatical. This poses a fundamental challenge for the learner, who must distinguish linguistically significant restrictions from *accidental gaps* (Fischer-Jørgensen 1956; Halle 1962; Chomsky & Halle 1965; Pierrehumbert 1994; Frisch & Zawaydeh 2001; Iverson & Salmons 2005; Gorman 2013; Hayes & White 2013). The severity of the challenge depends on the size of the lexicon (Pierrehumbert 2001), the number and relative frequency of sounds (Sigurd 1968; Tambovtsev et al. 2007), and the complexity of the generalizations that learners must entertain (Pierrehumbert 2001; Hayes & Wilson, 2008; Kager & Pater 2012; Jardine & Heinz 2016).

In this squib, we consider the problem that accidental gaps pose for learning surface phonotactic grammars. The empirical basis is the phonotactic pattern of South Bolivian Quechua, with a particular focus on the allophonic distribution of high and mid vowels. We show first that, in describing the vowel distribution, strictly surface-based theories must resort to constraints of greater complexity than would be needed in more traditional analyses that derive outputs from underlying forms. We go on to demonstrate that there are many accidental gaps in the learning data, and show that a modified version of the model proposed in Hayes & Wilson (2008) is effective in distinguishing them from systematic gaps. Extensive evaluation of this model and minimal alternatives provides evidence that two related properties are critical to its relative success: it uses features to state constraints at multiple levels of generality, and it selects constraints of appropriate granularity by statistical comparison of observed and expected frequencies.

1. Vowel height allophony in Quechua

1.1 The pattern and traditional analysis

Descriptively, South Bolivian Quechua (henceforth Quechua; Bills et al. 1971; Laime Ajacopa 2007; Gallagher 2016) has three phonemic vowels /i u a/ with allophonic lowering of /i u/ to [e o] in the vicinity of uvulars /q q^h q'/. Mid vowels occur immediately following or preceding a uvular (1a,b), or preceding a uvular across an intervening coda (1c). High vowels occur in all other consonantal environments (2).

(1) Uvular contexts: [e o] *[i u]

a.	**q'e**pij (*q'ipij)	'to carry'	**q'o**ŋi (*q'uŋi)	'hot'
b.	p**eq**aj (*piqaj)	'to grind'	n**oq**a (*nuqa)	'I'
c.	w**esq'**aj (*wisq'aj)	'to close'	t**oʌq**a (*tuʌqa)	'son-in-law'

(2) Elswhere: [i u] *[e o]

misi (*mese)	'cat'	kuʌku (*koʌko)	'type of bird'

A traditional analysis of this pattern would assign the high vowels *elsewhere* status and specify the contexts in which mid vowels occur. In an SPE-style analysis, only high vowels would appear in underlying representations and mid vowels would be derived (3). (Alternatively, non-low vowels would be underspecified for [high] underlyingly, and [+high] values would result from a default rule.) In OT/HG, complementary distribution would follow by subordinating a general constraint against mid vowels (*E) to specific constraints against high vowels before or after a uvular (4). (Natural classes are abbreviated throughout as follows: K = velar consonant, Q = uvular consonant, I = high vowel, E = mid vowel, V = any vowel. The rules and constraints stated in (3) and (4) apply on the dorsal tier, as specified in table (6) below.)

(3)

UR	/piqaj/	/misi/
I → [-high] // __ Q	peqaj	n/a
SR	[peqaj]	[misi]

(4)

/piqaj/	*QI, *IQ	*E	Id[high]	/mesi/	*QI, *IQ	*E	Id[high]
piqaj	* !			→ misi			*
→ peqaj		*	*	mesi		* !	

1.2 Analysis with surface-based constraints

In phonotactic grammars of the type considered here, constraints refer to a single *surface* level of representation. This avoids the difficulty of determining underlying representations, but has other perhaps surprising consequences. Specifically, in a surface-based analysis of Quechua contextual constraints are required for both high and mid vowels. The restrictions on high vowels can remain as above (*QI, *IQ). However, the constraint *E would be overly broad, as mid vowels do occur in surface forms. While constraints in the relevant models are in principle violable, *E would still be inappropriate: it would penalize all mid vowels equally regardless of context, and thus miss the difference between (for example) ungrammatical *[pena] and grammatical

[qena]. In place of *E, a surface-based analysis requires constraints against mid vowels in the exhaustive set of 'non-uvular' environments.

We provide an example analysis in (5), with constraints again stated on a dorsal tier that contains velar and uvular consonants along with all vowels. This tier is useful in accounting for the non-local interaction between a uvular and a vowel across an intervening coda (e.g., [**orq**o] 'mountain]), and has the welcome side effect of reducing the number of 'non-uvular' contexts that must be enumerated.

(5) a. Surface-true constraints on high vowels in uvular contexts (as in (4))
*Q I *I Q

b. Surface-true constraints on mid vowels in 'non-uvular' contexts

*# E K	*K E K	*V E K
*# E #	*K E #	*V E #
*# E V	*K E V	*V E V

Trigram constraints like those of (5b) are required because a mid vowel can be licensed by a uvular on either side. To exclude mid vowels in unlicensed contexts only, the constraints must ensure that something other than a uvular both precedes and follows.

1.3 Accidental gaps in Quechua

Previous research on surface-based phonotactics has primarily limited constraints to unigrams and bigrams (Adriaans & Kager 2010; Jardine & Heinz 2016; and with restricted exceptions Hayes & Wilson 2008), but Quechua shows that trigrams are necessary in such models.[1] In general, trigram constraints will be required for both-side conditioning patterns such as intervocalic voicing and lenition, but also for many other cases in which a sound can be conditioned by a context on either side.

In general, permitting constraints of greater complexity makes the problem of accidental gaps more severe. To quantify the problem for Quechua, we constructed an exhaust list of hypothetical CV(C)CV(C) forms (> 560,000 items) for which the position-specific unigrams and medial consonant clusters are attested in roots. We then divided this list of hypothetical roots into those that satisfy the known phonotactic generalizations given in (6), and those that violate one or more generalizations. For further discussion of

[1] Many other computational approaches to phonotactics, speech segmentation, and phoneme learning have in practice limited generalizations to unigrams and bigrams (Vitevitch & Luce 2004; Peperkamp et al. 2006; Kirby & Yu 2007; Albright, 2009; Adriaans & Kager 2010; Heinz 2010; Daland & Pierrehumbert 2011; Kempton & Moore 2014; Calamaro & Jarosz 2015; cf. Martin et al. 2013), and would have to be extended to account for Quechua. See Pierrehumbert (1994) and Kager & Pater (2012) for evidence of speaker knowledge of complex phonotactics.

the phonotactic restrictions in Quechua, and of the laryngeal co-occurrence restrictions in particular, see MacEachern (1999) and Gallagher (2011, 2016).

(6) Quechua tiers and phonotactic generalizations[2]

tier	*projected segments*	*phonotactic generalizations*
dorsal	dorsal consonants, vowels	high-mid vowel allophony
C-dorsal	dorsal consonants, +	*K…Q, *Q…K (within morphemes)
laryngeal	stops, affricates, h, ʔ	laryngeal cooccurrence restrictions
segmental	all	*V, *VV, *CCC, *wu, *wo

We compared the tier-based trigrams that occur in the list of legal hypothetical roots with those in a lexicon of 1104 actual roots that was compiled from the Laime Ajacopa (2007) dictionary and verified with a native speaker. On the segmental tier, there are 2966 unique trigram sequences in the hypothetical roots, but only 1472 (49%) of them are attested in the root lexicon. That is, on the segmental tier there are about as many trigram accidental gaps as attested trigrams. The ratio of attested to legal trigrams is higher on the other tiers, but several accidental gaps exist for each one (dorsal tier: 193 attested / 204 legal; C-dorsal tier: 17 / 19; laryngeal tier: 167 / 176).

The unattested legal trigrams typically combine segments that are independently rare. For example, the sequence [eqʼe] is legal but unattested, reflecting the fact that glottalized dorsals are infrequent in medial position generally and that [e] is the least frequent surface vowel in the language. However, it would clearly be incorrect for a learner to conclude that all unattested sequences containing rare parts are accidental gaps. For example, [k^h] is among the rarest segments in Quechua, and the sequence [k^hek] is has zero frequency like [eqʼe], but in this case the gap is principled (see Section 2.1). Considerations of this sort indicate that a surface-based phonotactic learner should induce constraints on a statistical basis, so that it can avoid penalizing sequences like [eqʼe] that are unlikely to occur by chance, and that the learned constraints should be stated at multiple levels of granularity, so that sequences like [k^hek] can be brought under more general restrictions (e.g., *KEK) rather than being ruled out individually.

2. Surface-based phonotactic learning models

2.1 Feature-based statistical model

The maximum entropy phonotactic model developed by Hayes & Wilson (2008) has both of the properties that we have identified as important for overcoming the problem of

[2] The C-dorsal tier contains a morpheme boundary symbol (+), allowing the model to represent the fact that the restriction on uvular and velar consonant cooccurrence holds within morphemes but not across them.

accidental gaps. Constraints in the model are stated over natural classes, and can range from segment-specific (e.g., "no [ʔ] in non-initial position") to very broad (e.g., "no vowels"). Hayes & Wilson (2008) propose that, among other criteria, learning favors constraints with low ratios of observed to expected frequencies. Expected values of hypothetical constraints are calculated from the probability distribution defined by the constraints already in the grammar together with their current weights.

The original maxent model has supported attempts to learn the entire phonotactic pattern of a natural language (i.e., Hayes & Wilson 2008 on Wargamay and Hayes & White 2013 on English). While these attempts were partly successfully, insofar as known generalizations were learned, the model also learned many constraints that penalize accidental gaps. Because the present goal is to learn all and only the phonotactic restrictions of Quechua, we modified the model in two ways.

(i) Initialization. The grammar was initialized with a separate violable constraint for each segment in the Quechua inventory. When appropriately weighted, these constraints are equivalent to a unigram stochastic model of the learning data. The general purpose of this initialization is to prevent the model from learning complex constraints against sequences that contain rare segments (e.g., as in the [eq'e] example from above).

(ii) Gain-based constraint selection. A grammar was induced one constraint at a time by calculating the *gain* of each surface-true constraint. The gain of constraint C is proportional to the highest log probability (of the learning data) that could be obtained by adding C to the grammar while holding all other constraints and their weights fixed (see Della Pietra et al. 1997). This criterion favors constraints that are violated substantially less often than would be expected by chance, given the current grammar, because probability assigned to violators could be profitably reallocated to attested forms. On each round of constraint selection, the constraint with highest gain above a fixed threshold ɣ on each tier was added to the grammar (we evaluated various thresholds and report results for ɣ = 100.0). Learning halted when no constraint had sufficient gain.[3]

For convenience, we refer to this revised maxent model as Maxent-Ftr. In principle, the model could induce gradient phonotactic grammars containing violable constraints. However, the Quechua phonotactics described earlier are categorical, and therefore we required all of the constraints induced by Maxent-Ftr to be surface-true.

[3] Application of the gain threshold is related to L1 regularization (e.g., Perkins et al. 2003), and to the cost of constraints in MDL models (e.g., Rasin & Katzir 2016), because it penalizes the addition of a new constraint regardless of its weight. We also included a term that penalizes large weights, $\lambda \sum_i w_i^2$, but set λ equal to a small constant (1.0e-5) that made this penalty negligible.

2.1 Alternative models

The first alternative we considered, called Maxent-Seg, is identical to Maxent-Ftr except that constraints are stated over segments (equivalently, singleton natural classes). While features are traditionally used in phonotactic descriptions, some recent models eschew them (e.g., Heinz 2010, Heinz & Rogers 2010; cf. Heinz & Koirala, 2010).[4] The comparison of Maxent-Ftr and Maxent-Seg provides a close examination of whether allowing constraints at multiple levels of granularity is key for learning. Maxent-Seg was initialized in the same way as Maxent-Ftr, had access to the same tiers, and used the same statistical criterion for inducing surface-true constraints.

The second alternative is a non-statistical version of Maxent-Seg, referred to as Memory-Seg and inspired by recent formal language research (Heinz et al. 2011; Jardine & Heinz 2016; McMullin 2016; see also de la Higuera 2010). However, while our presentation of the model draws upon that work, the research question addressed here is quite different. We are interested in the grammars that phonotactic models learn from natural 'gappy' data — not in the important but distinct question of what is provably learnable from hypothetical data in which all legal structures are exemplified.

In the Memory-Seg model, a grammar is defined as a set G_t of legal substrings for each tier t. At the onset of learning each G_t is empty. The sets are then updated with the substrings in forms encountered during learning. For example, upon encountering the form [toʎqa] the length-3 substrings added to the dorsal tier set are those in [oqa], namely [#oq], [oqa], [qa#]. In essence, learning involves memorizing the tier-specific segment sequences that are observed. This requires significantly less computation than in Maxent-Ftr or Maxent-Seg, because expected frequencies need not be calculated. There is one free parameter (n), analogous to the maximum constraint length in the Maxent models.

A form is ungrammatical with respect to a Memory-Seg grammar iff it contains at least one substring of length $1 \leq m \leq n$, on at least one tier t, that is not in G_t. For example, a grammar learned from the Quechua lexicon with $n = 3$ would not contain the substring [#ei] on the dorsal tier, and would therefore correctly identify *[mesi] as illegal.

Could there be a non-statistical model like Memory-Seg that learns by memorizing feature sequences (i.e., hypothetical Memory-Ftr)? The problem confronting such a model is that any given segment sequence has may different featural representations. Without a method for deciding *which* representations are relevant for assessing well-formedness — the role that statistics plays in Maxent-Ftr — learning is doomed.

Consider the attested dorsal-tier trigram [oqa], which could be represented with very general classes (e.g., [+syll][-syll][+syll] =VCV), with maximally specific classes (i.e.,

[4] Segmental *n*-gram models are also commonplace in natural language processing (e.g., Jurafsky & Martin 2000), but are typically limited to one segmental tier and contiguous sequences (cf. Ron et al. 1996).

[+syll, -high, -low, +back][-cont, -son, +dorsal, -high, -cg][+syll, -high, +low] = [oqa]), or at intermediate levels of granularity (e.g., [+syll, -high, -low][-cont, -son, +dorsal, -high][+syll, -high, +low] =EQA). If Memory-Ftr judged a sequence as grammatical if it satisfied *any* attested representation, the model would tolerate every VCV trigram and massively overgeneralize. If the model instead required *all* representations to be attested, it would be equivalent to Memory-Seg (as segments correspond to singleton classes).

3. Results

We evaluated the Maxent-Ftr, Maxent-Seg and Memory-Seg models with five-fold cross-validation (e.g., Hastie et al. 2001; Mohri et al. 2012). The complete learning data consisted of the Quechua root lexicon (Section 1.3), and forms derived from the roots by adding representative suffixes (*-nku* '3 pl present', *-spa* 'gerund', *-rqa* '3 sg past') and applying vowel lowering (3) when appropriate. The lexicon was divided into five 'folds' of roughly equal size (approx. 870 forms). Each fold served as a set of legal 'held-out' test forms for models trained on the other four folds. Testing also included the exhaustive set of CV(C)CV(C) nonce roots, each categorized as legal or illegal according to (6).

The models were provided with the same tiers and allowed to learn generalizations up to length three. For the Maxent models, a test form was grammatical iff it satisfied all of the learned constraints. For Memory-Seg, grammaticality was determined as described above. Table 1 shows the proportion of test forms judged grammatical by each model.

	held-out forms	legal nonce roots	illegal nonce roots
Maxent-Ftr	99.8%	82.2%	1.9%
Maxent-Seg	99.7%	71.5%	45.4%
Memory-Seg	96.7%	18.8%	0.1%

Table 1: Proportion of test forms judged grammatical by each model.

The models performed comparably on attested but held-out forms. However, Maxent-Ftr generalized far more successfully to nonce roots, ruling out essentially all of the illegal forms and accepting the great majority of the legals. Nearly all (96%) of the false negatives made by this model involved forms with root-final consonants ([n s r w j x]), which occur rarely in the lexicon and may be phonotactically marginal.[5] Maxent-Seg both undergenerated and massively overgenerated (e.g., it accepted illegal forms such as *[p'ap'a] that violate laryngeal co-occurrence restrictions). This comparison indicates that statistical calculations alone, in the absence of features or the classes they define, do

[5] Word final consonants are frequent in the language as a whole, but they are rare in roots and in our learning set (which did not include any consonant final suffixes).

not suffice for phonotactic learning. Finally, the Memory-Seg grammars showed extreme undergeneralization: by judging forms as ungrammatical if they contained accidental gaps such as [eq'e], these grammars accepted less than 20% of the novel legal roots. The Maxent-Ftr model largely refrained from inducing constraints against [eq'e] and other gaps because the relevant segments are independently rare (i.e., the expected violations of segment-specific constraints are too small) and because more general constraints (e.g., *EQE) are violated by many attested forms (i.e., their observed violations are too large).

4. Conclusion

Most computational research on phonotactic learning has not explicitly considered the problem of accidental gaps. Previous work that does address the problem has focused on specific constraints of English (e.g., Pierrehumbert 1994; Gorman 2013; Hayes & White 2013) or achieved limited results on other languages (e.g., Hayes & Wilson 2008). In this squib, we have shown that the Quechua lexicon is full of accidental gaps — particularly when analyzed at the level of trigrams, as required by a surface-based analysis of vowel height allophony. We have further shown that accurately learning the entire phonotactic pattern of a language is nevertheless within reach of a surface-based model that uses statistical computations to learn constraints over featural representations (the importance of representations in inductive phonotactic models has been previously shown in Albright 2009 and Berent et al. 2012). A minimally-different segmental model overgeneralized, because some parochial constraints were too statistically weak to meet the induction criterion. A model that forgoes both statistics and features had the opposite problem: ignoring the possibility that some sequences will be absent by chance, it incorrectly penalized all unattested trigrams.

We conclude with some general comments about further challenges facing computational models of phonotactic learning. First, all models considered here relied on generalizations stated over tiers, but the tiers were stipulated by the analyst. In principle, the learner could discover both the need for tiers and their contents (e.g., Jardine 2016). Second, the 'sparse data' problem addressed here becomes more severe, and may arise even for bigram constraints, as more detail is represented in surface forms. For Quechua, we transcribed allophonic vowel height, but not the tense/lax vowel distinction conditioned by syllable structure nor the variable production of uvular stops (i.e., /q/ may be [q], [ʁ] or [ɢ]; /q^h/ may be [q^h] or [χ]; /q'/ may be [q'] or [ɢ̰]). Transcribing all such distinctions would increase the number of possible surface sequences and compound the problem of distinguishing linguistically-significant and accidental gaps (e.g., Martin et al. 2013). We hope future work will take up the challenge of learning phonotactics from natural language data that is increasingly detailed and inevitably sparse.

References

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

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.

Bills, G., Rudolph, C.T., Bernardo, V. (1971). *An Introduction to Spoken Bolivian Quechua*. Austin: University of Texas Press.

Calamaro, S., & Jarosz, G. (2015). Learning general phonological rules from distributional information: A computational model. *Cognitive Science*, *39*(3), 647-666.

Chomsky, N., & Halle, M. (1965). Some controversial questions in phonological theory. *Journal of Linguistics*, *1*(2), 97-138.

Daland, R., & Pierrehumbert, J. B. (2011). Learning Diphone-Based Segmentation. *Cognitive Science*, 35(1), 119-155.

De la Higuera, C. (2010). *Grammatical Inference: Learning Automata and Grammars*. Cambridge: Cambridge University Press.

Della Pietra, S., Della Pietra, V., & Lafferty, J. (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(4), 380-393.

Fischer-Jørgensen, E. (1952). On the definition of phoneme categories on a distributional basis. *Acta linguistica*, *7*(1-2), 8-39.

Frisch, S. A., & Zawaydeh, B. A. (2001). The psychological reality of OCP-Place in Arabic. *Language*, *77*(1), 91-106.

Gallagher, G. (2011). Acoustic and articulatory features in phonology–the case for [long VOT]. *The Linguistic Review*, *28*(3), 281-313.

Gallagher, G. (2016). Vowel Height Allophony and Dorsal Place Contrasts in Cochabamba Quechua. *Phonetica*, *73*(2), 101-119.

Gildea, D., & Jurafsky, D. (1996). Learning bias and phonological-rule induction. *Computational Linguistics*, *22*(4), 497-530.

Goldsmith, J., & Riggle, J. (2012). Information theoretic approaches to phonological structure: the case of Finnish vowel harmony. *Natural Language & Linguistic Theory*, *30*(3), 859-896.

Gorman, K. (2013). Generative Phonotactics. Unpublished PhD dissertation, University of Pennsylvania.

Halle, M. (1962). Phonology in generative grammar. *Word*, *18*(1-3), 54-72.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer.

Hayes, B. (2004). Phonological acquisition in Optimality Theory: The early stages. In Kager, R., Pater, J., and Zonneveld, W. (eds.), *Fixing Priorities: Constraints in Phonological Acquisition*, 158-203. Cambridge: Cambridge University Press.

Hayes, B., & White, J. (2013). Phonological naturalness and phonotactic learning. *Linguistic Inquiry*, *44*(1), 45-75.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*(3), 379-440.

Heinz, J. (2010). Learning long-distance phonotactics. *Linguistic Inquiry*, *41*(4), 623-661.

Heinz, J., Rawal, C., & Tanner, H. G. (2011, June). Tier-based strictly local constraints for phonology. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*, 58-64. Association for Computational Linguistics.

Heinz, J., & Rogers, J. (2010, July). Estimating strictly piecewise distributions. In *Proceedings of the 48th Annual meeting of the Association for Computational Linguistics*, 886-896. Association for Computational Linguistics.

Iverson, G. K., & Salmons, J. C. (2005). Filling the gap. *Journal of English Linguistics*, *33*(3), 207-221.

Jardine, A. (2016). Learning tiers for long-distance phonotactics. In Laurel Perkins et al., (eds.), *Proceedings of the 6th Conference on Generative Approaches to Language Acquisition North America (GALANA 2015)*, pp. 60-72. Somerville, MA: Cascadilla Proceedings Project.

Jardine, A., & Heinz, J. (2016). Learning Tier-based Strictly 2-Local Languages. *Transactions of the Association for Computational Linguistics*, *4*, 87-98.

Jarosz, G. (2006). *Rich lexicons and restrictive grammars: Maximum likelihood learning in Optimality Theory*. Unpublished PhD dissertation, Johns Hopkins University.

Jurafsky, D. & Martin, J. (2000). *Speech and Language Processing*. Upper Saddle River: Prentice Hall.

Kager, R., & Pater, J. (2012). Phonotactics as phonology: Knowledge of a complex restriction in Dutch. *Phonology*, *29*(1), 81-111.

Kempton, T., & Moore, R. K. (2014). Discovering the phoneme inventory of an unwritten language: A machine-assisted approach. *Speech Communication*, *56*, 152-166.

Laime Ajacopa, Teofilo. 2007. Diccionario Bilingüe, Iskay Simipi Yuyayk'ancha: Quechua – Castellano Castellano–Quechua. La Paz, Bolivia.

MacEachern, M. R. (1999). *Laryngeal Cooccurrence Restrictions*. New York: Garland.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT press.

Martin, A., Peperkamp, S., & Dupoux, E. (2013). Learning phonemes with a proto-lexicon. *Cognitive Science*, *37*(1), 103-124.

McMullin, K. J. (2016). *Tier-based Locality in Long-Distance Phonotactics: Learnability and Typology*. Unpublished PhD dissertation, University of British Columbia.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning*. Cambridge, MA: MIT Press.

Peperkamp, S., Le Calvez, R., Nadal, J. P., & Dupoux, E. (2006). The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition*, *101*(3), B31-B41.

Perkins, S., Lacker, K., & Theiler, J. (2003). Grafting: Fast, incremental feature selection by gradient descent in function space. *Journal of Machine Learning Research*, *3*, 1333-1356.

Pierrehumbert, Janet. 1994. Syllable structure and word structure: A study of triconsonantal clusters in English. In Keating, P. (ed.), *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*, 168–188. Cambridge: Cambridge University Press.

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech*, *46*(2-3), 115-154.

Prince, Alan, and Bruce Tesar. 2004. Learning phonotactic distributions. In Kager, R., Pater, J., and Zonneveld, W. (eds.), *Fixing Priorities: Constraints in Phonological Acquisition*, 245–291. Cambridge: Cambridge University Press.

Rasin, E., & Katzir, R. (2016). On evaluation metrics in Optimality Theory. *Linguistic Inquiry*, *47*(2), 235-282.

Ron, D., Singer, Y., & Tishby, N. (1996). The power of amnesia: Learning probabilistic automata with variable memory length. *Machine Learning*, *25*(2-3), 117-149.

Sigurd, B. (1955). Rank order of consonants established by distributional criteria. *Studia Linguistica*, *9*(1-2), 8-20.

Tambovtsev, Y., & Martindale, C. (2007). Phoneme frequencies follow a Yule distribution. *SKASE Journal of Theoretical Linguistics*, *4*(2), 1-11.

Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 481-487.