Learning General Phonological Rules from Distributional Information: A Computational Model

Shira Calamaro and Gaja Jarosz

Department of Linguistics, Yale University
Abstract

Phonological rules create alternations in the phonetic realizations of related words. These rules must be learned by infants in order to identify the phonological inventory, the morphological structure, and the lexicon of a language. Recent work proposes a computational model for the learning of one kind of phonological alternation, allophony (Peperkamp, Le Calvez, Nadal, Dupoux 2006). This paper extends the model to account for learning of a broader set of phonological alternations and the formalization of these alternations as general rules. In Experiment 1 we apply the original model to new data in Dutch and demonstrate its limitations in learning non-allophonic rules. In Experiment 2 we extend the model to allow it to learn general rules for alternations that apply to a class of segments. In Experiment 3 the model is further extended to allow for generalization by context; we argue that this generalization must be constrained by linguistic principles.

Keywords: Linguistics, Phonology, Language Acquisition, Computational Model, Statistical Learning
1. Introduction

Phonological rules introduce variation in the phonological realization of phonemes, which results in systematic relationships between classes of sounds called phonological alternations. Such variation can be readily observed when it produces alternations in the realization of sounds in morphologically related words. For example, the [f] in ‘knife’ is pronounced as [v] in the plural form ‘knives’. These phonological alternations pose a challenge to learners because they obscure the relationships between related words. Such processes are common across languages. For example, in Dutch, voicing alternations cause stem-final obstruents in singular nouns to be pronounced differently than in the plural forms. The stem-final obstruent in the Dutch word for ‘earl’ is voiceless [f] in the singular [ɣraf] but alternates with voiced [v] in the plural form [ɣravən] (see e.g. Booij 1995). In order to learn important aspects of the morphology and the lexicon, infants must discover the relationships that hold between sounds in the language, identifying which segments alternate with one another and where these alternations occur.

Alternations can be categorized into two major classes depending on the phonological status of the alternating segments in the language. In English, [t] and [d] are contrastive – they belong to different phonological categories – since they can be used to distinguish words like ‘time’ and ‘dime,’ which only differ in a single segment. The Dutch case above is an alternation that involves such contrasting segments since [f] and [v] can be used to contrast words in Dutch, such as [ɣravən] ‘earls’ vs. [ɣrafən] ‘graphs’. The alternation between [f] and [v] gives rise to a partially overlapping distribution: while [f] and [v] can occur in the same context, such as between [a] and [ə], their distribution is restricted in word-final position, where only [f] occurs. Another type of alternation, allophonic, involves segments that belong to the same phonological category and are not contrastive. Segments involved in allophonic alternations cannot be used to contrast words because their distribution is non-overlapping, or complementary. For example, in English, the sounds written as ‘t’ in ‘Italy’ and ‘Italian’ involve an allophonic alternation that results from a shift in stress placement: the ‘t’ is pronounced as a flap [ɾ] in ‘Italy’ but as an aspirated stop [tʰ] in ‘Italian’. There are no words that contrast [ɾ] and [tʰ] in English because they never occur in the same phonological context. Thus, both types of alternations result in systematic variation, but the distributional regularities they give rise to differ: allophonic
alternations create complementary distributions, while non-allophonic alternations give rise to partially overlapping distributions.

A recent model explores the problem of learning pairs of segments involved in allophonic alternations on the basis of their distributions (Peperkamp, Le Calvez, Nadal, & Dupoux 2006). The model identifies potential allophonic alternations by detecting their complementary distributions. This approach cannot directly extend to alternations involving contrastive segments since, as explained above, contrastive segments have partially overlapping distributions. The present paper proposes extensions to this model that allow it to also learn alternations involving contrastive segments. In addition, the model proposed here includes a second set of extensions: it learns higher-level phonological structure beyond pairs of segments, characterizing generalizations about patterns of alternation as general rules. Sections 2, 3, and 4 introduce the original model, motivate the need for learning general rules, and review related work. Sections 5, 6 and 7 propose extensions to the model and present simulations on Dutch that show the effects these extensions have.

2. The Allophonic Alternation Learning Model

Peperkamp et al 2006 (PLND) propose an algorithm for the learning of allophonic alternations relying on statistical and linguistic information. This section describes PLND’s model, focusing on the components of the model crucial to the rule learning extensions proposed in this paper.

\[ m_{KL}(s_1, s_2) = \sum_{c \in C} P(c|s_1) \log \frac{P(c|s_1)}{P(c|s_2)} + P(c|s_2) \log \frac{P(c|s_2)}{P(c|s_1)} \]

The statistical component of PLND’s model is designed to detect the systematic distributional regularities created by allophonic alternations, namely their complementary distribution. Segments related by allophonic rules, such as [r] and [tʰ] in English, have very different distributions because in any given context, only one or the other segment may occur.

---

1 Where \( P(c|s) = \frac{n(s,c)+1}{n(c)+N} \), \( n(c,s) \) is the number of occurrences of segment \( s \) in the context \( c \) (the number of occurrence of the sequence \( sc \)), \( n(s) \) is the number of occurrences of the segment \( s \), and \( N \) is the number of contexts.
The model applies the Kullback-Leibler dissimilarity measure (Kullback and Leibler 1951) to quantify the distance between the segments’ distributions. The measure defines the dissimilarity $m_{KL}$ for each pair of segments $s_1$ and $s_2$ as the symmetrized, weighted logarithmic difference between the two distributions corresponding to $s_1$ and $s_2$, as defined in (1). At each context $c$, the measure compares the probability of $c$ for each of $s_1$ and $s_2$, assigning higher dissimilarities to segments with more divergent probabilities for $c$. The dissimilarity is zero for identical distributions and increases for more distinct distributions. The extensions of the model proposed in Sections 6 and 7 reformulate this statistical component of the model.

While the statistical component of the model identifies pairs of segments with different distributions, many such pairs are spurious and do not correspond to alternations. For example, consonants and vowels have very different distributions, with vowels usually surrounded by consonants and vice versa. In addition to having distinct distributions, alternating pairs tend to be phonetically similar, and the alternations tend to be phonetically related to the contexts in which they occur. The linguistic component of PLND’s model formalizes these tendencies as filters that are used to eliminate spurious pairs selected by the statistical component\textsuperscript{2}. In order to give the linguistic filters access to phonetic information, PLND represent each segment as a vector of phonetic features with values for place of articulation, sonority, voicing, nasality, and rounding. The first filter removes pairs that are not neighbors in the phonetic space. The second filter requires that for all features, one of the segments in the pair (the allophone) be more similar to its contexts than the other segment (the default)\textsuperscript{3}. PLND formalize the traditional notion that the allophone should have the more predictable distribution of the two in terms of relative entropy. Therefore, pairs that survive both filters are neighbors in phonetic space (filter 1), and one of the segments of the pair has been designated the default and the other the allophone (filter 2). These filters require the more predictable of the two members (the allophone) to also be more similar to its contexts than the default. The extensions we propose in Sections 6 and 7 directly incorporate both of these linguistic filters. In Section 7, we argue that two additional linguistic conditions are needed to constrain context generalization.

\textsuperscript{2} See Appendix A for formal definitions of the Linguistic Filters.
\textsuperscript{3} PLND define the allophone statistically as the segment with higher relative entropy: it is the segment with the more predictable distribution, formalizing the tr. See Appendix A for the formal definition.
3. General Phonological Rules

Before turning to the extensions and experiments, we briefly discuss the motivations for representing phonological alternations as general rules. As described above, PLND’s model represents allographic alternations as pairs of segments. While this identifies the alternating segments, it does not characterize the changes in features that occur and where they occur. More generally, pairs of alternating segments do not indicate how to apply the process or how to extend it to related pairs or contexts. Learning that the pairs [b, p] and [d, t] alternate is an important part of the learning problem, but knowing that they are stops that alternate in voicing in the word-final position is crucial for characterizing the phonological system of the language as a whole. It is this higher-level generalization that learners can use when faced with a related alternating pair, like the [v] in [ɣavən] mapping to the [f] in singular [ɣraf].

Experimental work has shown that learners are capable of making high-level generalizations about the sound patterns of a language. Most relevant is recent work showing that infants and adults can generalize phonetic and phonological patterns to related segments in the same natural class. For example, Cristià and Seidl (2008) found that 7-month-olds exposed to a phonotactic constraint requiring initial segments to be nasals or stops generalized this constraint to novel segments in the same natural class, the class of [−continuant] segments. Finley & Badecker (2009) showed that adult learners generalized vowel harmony alternations to novel vowels belonging to the same natural class. Other work has shown that learners are able to make generalizations about the relationships between segment classes as well as the contexts that condition phonological processes. Specifically, Maye, Weiss, & Aslin (2008) showed that infants can generalize a phonetic contrast (VOT) to segments at a novel place of articulation, and Wilson (2006) found that adult learners familiarized with a phonological process (palatalization) conditioned in one context (mid, front vowels) extended it to a phonologically related context (high, front vowels).

Taken together the experimental evidence indicates that human learning of sound patterns involves generalization about the classes of segments that pattern together, their relationships, and the contexts in which they occur. This is the type of information that is represented in traditional phonological rules of the form A → B / X_Y (see e.g. Chomsky and Halle 1968). For
example, the rule \([+\text{vocalic}] \rightarrow [+\text{nasal}] / \_\_ [+\text{nasal}, −\text{vocalic}]\) paired with the assumption that nasal vowels are not permitted underlyingly, captures the allophonic distribution of vowel nasality in English, accounting for the distinct vowels in words such as ‘kin’ [kɪn] and ‘kid’ [kɪd]. The rule states that a segment that belongs to the class of vowels becomes nasalized when it appears before a segment belonging to the class of nasal consonants. This rule captures a generalization that can be applied to any segment belonging to a natural class. Such a rule reflects three kinds of generalization not captured by pairs alone: generalization identifying the class of segments affected by the alternation (vowels), generalization identifying the phonetic change (nasalization), and generalization about the contexts in which the process applies (nasal consonants). It also encodes information about the direction of the change, specifying how to apply the alternation in any given context. This paper proposes extensions to PLND’s model that enable systematic patterns of alternations to be learned in terms of such general rules.

4. Background and Relation to Previous Work

In this work, we explore a distributional approach to the learning of general phonological rules that builds on the work of PLND. Like PLND, we define the learning task as the discovery of phonological alternations from unstructured phonological data. In this bottom-up approach, the learner must use distributional information to discover a general phonological rule relating \([f]\) and \([v]\) in Dutch without ever being told that forms like \(ɣraf\) and \(ɣravan\) are morphologically related or that they both derive from a stem with an underlying \(/v/\). This distributional approach is consistent with recent experimental findings indicating that one-year-olds can use distributional information to infer phonological alternations (White et al. 2008). These findings suggest that early distributional learning of systematic phonological alternations may provide infants with a way to relate alternating words or morphemes, paving the way for discovery of lexical and morphological structure that is obscured by phonological alternations. Because it is consistent with this bottom-up approach, PLND’s distributional model provides a good starting point for our work on learning of general phonological rules from distributional information.

The present focus on distributional learning distinguishes our approach from other important work on rule learning that requires morphological annotation of the input data (e.g.
These prior models crucially rely on the input data being arranged into morphological paradigms, where the morphological relationships between surface forms are provided. Such morphological annotation dramatically changes the nature of the learning problem by enabling discovery procedures not available to a purely distributional learner. Access to morphologically related pairs such as [ɣraf] and [ɣravan] provides the learner with direct evidence of alternations that the rules must explain. In contrast, a distributional learner has no direct evidence of alternations: the alternations must be inferred from statistical regularities. In principle, learning of morphology could precede and inform learning of phonological alternations, with the output of morphological learning serving as the input to rule learning models like the above. However, as discussed above, there are reasons to believe that infants’ early learning of phonological alternations guides morphological and lexical analysis. Our goal is to pursue this possibility to determine how learning of phonology may proceed from distributional information, without access to morphology. Thus, the task undertaken in the present work is the discovery of relationships between classes of segments, formalized as general rules, that underlie systematic distributional regularities. In this context, unlike in the approaches above, alternations in the realizations of words or morphemes cannot be directly observed because the learner does not have access to morphology. Nonetheless, the learner’s task, like that of the learners above, is to discover general phonological rules governing the systematic phonological variation that gives rise to regularities in the data.

Viewed in a broader learning context, a complete morpho-phonological analysis of the data would include (at least) a morphological analysis of the surface forms, underlying representations for all morphemes, and a set of phonological rules mapping underlying representations to the surface forms. The goal of the previous models described above is to discover the phonological rule component, given information about the morphology or underlying representations. In contrast, the present goal can be viewed as a bottom-up approach to this larger problem of learning morpho-phonology, focusing first on the learning of general phonological rules based only on their observable consequences for the surface sound patterns of the language. As discussed above, experimental results motivate this bottom-up approach,
wherein early learning of phonological alternations guides subsequent learning of the morpho-
phonological system as a whole. If during an early, bottom-up stage of learning, the learner can
infer that, for example, /v/ maps to [f] word-finally, this would provide a strong foundation for
the subsequent inference that [ɣraf] and [ɣravan] may share a stem whose underlying
representation is /ɣraw/. Thus, contextualizing the present work within this broader learning
problem, our goal is to explore the possibility of learning alternations as the first step of morpho-
phonological learning.

Our bottom-up approach is also in line with recent work on the learning of phonotactic
regularities (Hayes and Wilson 2008). Like Hayes and Wilson, our goal is to extract
phonological generalizations from distributional information alone. However, while Hayes and
Wilson focus on the learning of a grammar characterizing static phonotactics, our focus is on
using distributional information to discover phonological rules that characterize systematic
relationships between segment classes, e.g. alternations. Thus, the difference lies in the nature of
the inferred knowledge. Based on distributional information, Hayes and Wilson’s model
constructs a grammar that assigns wellformedness scores to surface representations. In contrast,
given a corpus of unstructured phonological data where one or more alternations have applied,
the present goal is to uncover systematic relationships between sound classes based only on their
distributional consequences.

5. Experiment 1: Allophonic vs. Non-allophonic alternations in PLND
The first experiment tests PLND’s original model on new data, involving different kinds of
alternations and a different method of phonetic transcription. We test PLND’s model on both
allophonic and non-allophonic alternations. Although PLND’s model was designed for the
learning of allophonic alternations only, the results in this section provide an important baseline
for the extensions we subsequently propose. In addition, because the transcription method for our
data differs from that used by PLND in their experiments, this experiment also explores the
model’s learning of allophonic alternations on this new kind of data.
5.1. Data

This experiment relies on data from the Spoken Dutch Corpus, which uses broad phonemic transcription (Oostdijk 2000; CGN). Because this corpus encodes phonetic variation in the transcription of lexical items, it differs from the French data used by PLND, which relied on citation form transcriptions. Although PLND showed using artificially constructed data that the model is robust to some random noise, it is unclear how natural phonetic variability will impact the performance of the model. Thus, in addition to exploring new alternation types, this experiment also examines the original model’s performance on data from a new language and its robustness to phonetic variation of this sort.

Experiment 1a tests the learning of an allophonic alternation, the devoicing of the liquids [r] and [l] before voiceless obstruents, [p], [t], [k], [f], [s], and [x]. Following PLND’s methodology, we artificially applied this allophonic rule to the Dutch data. Since this alternation is allophonic, these pairs are in complementary distribution and never occur in the same environment: liquids are voiceless before voiceless obstruents and voiced elsewhere. This is the kind of alternation for which PLND’s model is designed.

Experiment 1b tests the learning of a non-allophonic alternation in the voicing of obstruents, specifically, the devoicing of obstruents [b], [d], [g], [v], [z], and [ɣ] in word-final position. As above, this rule was artificially applied to the Dutch data. The voiceless variants [p], [t], [k], [f], [s], and [x] are phonemes in the language which means that, unlike the allophonic alternation, the obstruent voicing alternation does not create a complementary distribution: both classes of obstruents may occur in non-word-final position. The following example shows that the contrast between voiced and voiceless obstruents, in this case [v] and [f], is neutralized to voiceless, in this case [f], word-finally:

(2) Word-final devoicing in Dutch (Booij 1995)
   a. γraf ‘graph’ γraʃon ‘graphs’
   b. γraf ‘earl’ γraʃon ‘earls’

As the example also shows, both voiced and voiceless variants occur in other environments, such as preceding a vowel, because voiced and voiceless obstruents, such as [v] and [f] are contrastive
in Dutch. This rule is therefore a new kind of alternation on which the model has not previously been tested.

5.2. Results and discussion

The results of Experiment 1 are summarized in Fig. 1. We report the standard evaluation measures of precision, recall, and f-score at various detection thresholds for the full model with both linguistic filters on. The detection threshold determines a cut-off in the KL dissimilarity score such that pairs above this threshold are returned by the statistical component of the model: it is defined as the Z-score of the KL measure relative to the KL scores across all pairs. Precision is the proportion of pairs returned by the model that are true alternating pairs, and recall is the proportion of true alternating pairs that the model returns. Precision penalizes false positives, recall penalizes false negatives, and the harmonic mean of the two, f-score, balances these two considerations. In Experiment 1a there are two alternating pairs to discover, while in Experiment 1b, there are six. Ideally, for an alternation to be learned successfully, there would be some threshold for which all true alternating pairs are returned (yielding a high recall score), and no false positives are returned (yielding a high precision score). Such an ideal solution would have an f-score of 1.
Figure 1. Experiment 1a and 1b: Performance as a function of detection threshold

As shown in Fig. 1a, the model’s recall is ideal at thresholds between 0 and 1.5, indicating it has successfully found both pairs of allophonic alternations. As the threshold increases, the precision increases because the model posits fewer and fewer pairs, including false positives, until it posits no pairs whatsoever at a threshold of 3. Overall, the model performs quite well: at a threshold of 1.5 it correctly identifies both alternating pairs and posits only one spurious pair [j, u], resulting in an f-score of 0.8. Both linguistic filters improve performance by substantially decreasing the number of spurious pairs at all thresholds\(^4\). For example, at the threshold of 1.5, the precision of 0.67 with both filters drops to 0.50 when just filter 1 is used, to 0.10 with just filter 2, and to 0.02 without either filter. In general, using data from a new language, this experiment confirms that the statistical component of the model can detect allophonically alternating pairs, while the filters are largely successful in eliminating spurious pairs. It also indicates that the limited amounts of phonetic variability in the Dutch corpus do not pose a significant problem for the model. However, while the variability of these data add some degree of phonetic realism, the allophonic rule itself was applied manually, and its exceptionless nature minimizes the potential impact of phonetic variation. The model’s performance is most likely to be impacted by phonetic variation that disrupts the regularity of the allophonic distribution – how pervasive such irregularities are in natural speech and how they affect performance is an important question for future work.

Experiment 1b was less successful. As indicated by the recall of 0.33, the model returned only 2 out of 6 alternating pairs at a threshold of 0. The detected pairs were the two most frequent of the alternating pairs: [d, t] and [z, s]. No pairs were found at thresholds of 1 and above. At all thresholds, the number of false positives outnumbers the number of pairs successfully identified, as evidenced by the low precision. The filters fail to exclude [j, u] and several other spurious pairs. Thus, overall, the statistical component of the model failed to detect four of the six alternating pairs, and the filters were not successful in identifying and eliminating spurious pairs. In other words, the model is not capable of distinguishing regular non-allophonic alternations from spurious pairs.

\(^4\) See Appendix B for the complete results on the effects of the linguistic filters.
Taken together, the results from these experiments indicate that while this model can find allophonic alternations across multiple languages, it is not suited for finding pairs of alternating segments that are non-allophonic. Since the model was not designed for alternations involving contrastive segments, the poor results in Experiment 1b are not terribly surprising. However, it is important to consider the learning of both allophonic and non-allophonic alternations since learners must ultimately acquire both. The original model is equipped to learn allophonic alternations due to their highly complementary distribution. The voiceless liquids only appear before voiceless obstruents, while the voiced liquids appear everywhere else. Because the segments’ distributions are dissimilar across many contexts, the KL score is high overall. The non-allophonic alternations are more difficult to detect because they are not in complementary distribution. As discussed earlier, these segments can be used to distinguish words, so they can occur in the same context. The devoicing alternation creates a partially overlapping distribution, making it less complementary overall than the distribution of allophonic alternations. Out of all possible contexts, the distribution for the pair [b, p] is highly dissimilar only at the word-final context. At all other contexts the distribution between the two segments is very similar, and therefore the overall KL score is too low to be detected by the model as an alternation.

6. Experiment 2: Learning generalized rules
This section presents several extensions to the model, and the extended model is used to learn general rules for the non-allophonic alternations examined in Experiment 1. The new version of the model has two major changes from the original model. First, the model examines how dissimilar a distribution is for a pair of segments at each context, rather than collapsing dissimilarity across all contexts. This is necessary because pairs in non-allophonic alternations are only in complementary distribution in a limited number of contexts. To accomplish this, the new model uses a new formulation of dissimilarity that is context specific. Second, the model makes generalizations about related pairs to learn phonological rules. The phonological rules are represented by the change in features that occurs between alternating pairs and the context in which the alternation occurs.
6.1. Method

The new model calculates a contextualized score for a pair of segments at every context rather than finding a single dissimilarity score for each pair by summing dissimilarities over all contexts. This allows the model to identify the contexts where the two segments pattern differently. As shown in (3), the contextualized score for a pair of segments \( s_1 \) and \( s_2 \) at a context \( c \) is defined in terms of a dissimilarity measure \( D \) that compares the probability of \( c \) given \( s_1 \) to the probability of \( c \) given \( s_2 \). \( D \) is identical to the KL dissimilarity measure defined above in (1) except that no summation is taken over contexts. \( D \) is high when the two segments differ in their relative probabilities of occurring in context \( c \) and low when both segments occur in context \( c \) at similar rates. The final score is calculated by multiplying the dissimilarity \( D \) of a pair at a context by the squared \( Z \)-score relative to all dissimilarity scores for the pair. The \( Z \)-score term increases the score for contexts that have more complementary distributions than other contexts for that pair, partially relativizing the score to the pair of segments. Overall, contextualized scores identify contexts at which two segments behave differently from one another, highlighting possible conditioning contexts. These scores are calculated only for pairs that satisfy PLND’s linguistic filters, that is, for pairs that are neighbors in phonetic space and for which one member can be identified as the default or underlying member\(^5\). This means the two segments’ distributions are consulted first to determine whether they satisfy PLND’s filters, and then the contextualized scores are calculated for each context for the pairs that survive the filters.

\[
\begin{align*}
\text{Contextualized score calculation} \\
Score_{(c,s_1,s_2)} &= D(c,s_1,s_2) \times Z^2 \\
&\text{where } D(c,s_1,s_2) = P(c|s_1) \log \frac{P(c|s_1)}{P(c|s_2)} + P(c|s_2) \log \frac{P(c|s_2)}{P(c|s_1)} \\
&\text{and } Z = \frac{D(c,s_1,s_2)-\mu}{\sigma} \quad \text{\(6\)}
\end{align*}
\]

Contextualized scores are used to learn general contextualized rules by representing the structural change between the two segments as a vector of their featural differences \( \hat{f} \), which encodes how the two segments differ on each of the features. The featural change vectors are then used to discover contextualized rules by finding other pairs that exhibit the same featural

---

\(^5\) PLND’s filters are discussed in Section 2 and defined formally in Appendix A.

\(^6\) Where \( \mu \) and \( \sigma \) are the mean and standard deviation, respectively, of KL scores for pair \((s_1,s_2)\).
change at that context. As shown in (4), the score of the contextualized rule is simply the sum of the scores of pairs that support it.

(4) Score for contextualized rules

\[ \text{Score}_{(c,f)} = \sum_{(s_1,s_2) \text{ where } |s_1-s_2|=f} \text{Score}_{(c,s_1,s_2)} \]

Segments are represented as vectors of values for the feature vector: [place, sonority, voicing, nasality, rounding]. For word-final devoicing, the alternating pair [d, t] has the feature change vector [0, 1, 1, 0, 0] due to the fact that [t] differs from [d] by 1 on the voicing and sonority features. This feature change vector is associated with the context ‘#’, resulting in the contextualized rule (#, [0, 1, 1, 0, 0]) specific to [d, t]. All six voiced-voiceless obstruent pairs have the same feature change vector and can therefore be described by the same rule. This means they can be combined into a single rule whose score is the sum of the contextualized scores of each of the pairs. By summing the scores of supporting pairs, contextualized rules occurring most frequently in the data overall will have the highest scores.

The contextualized rule representation described above is convenient for computational purposes, and we present the format used by the model for transparency. However, it is important to note that the information the model captures corresponds to several kinds of generalizations encoded in traditional re-write rules of the sort discussed earlier. First, the feature vector encodes generalization about the alternating features (voicing, sonority), and second, the context encodes generalizations about where the feature change occurs (#). The model also encodes information about the directionality of the alternation, namely that /d/ should map to [t], /v/ should map to [f], and so on. This information is available in the segment pairs associated with each contextualized rule, which specify the underlying member of each pair (as defined by PLND’s second filter). Thus, the model represents all the information needed to apply devoicing to underlying word-final voiced obstruents.

---

7 The full feature system is summarized in Appendix C.
6.2. Results and discussion

We tested the extended model on the Dutch data with word-final devoicing from Experiment 1b. Fig. 2 shows the highest scoring contextualized rules found by the model, represented as contexts of application and feature change vectors. Alternating features are labeled in the vectors while associated segment pairs are omitted for readability. The highest scoring rule

($(#, [0, 1, 1, 0, 0])$) is word-final devoicing. Its score of 446.3 is more than four times as high as the next highest rule, $(\sigma, [0, 5, 0, 0, 1])$, which derives from the spurious pair $(\gamma, u)$ at context ‘\(\sigma\)’ with a score of 101.7.

The extended model succeeds in learning the word-final devoicing rule. The use of generalization aids in separating robust rules from spurious rules because systematic alternations give rise to higher combined scores. The score of the devoicing rule $(#, [0, 1, 1, 0, 0])$ is the sum of the contextualized scores of the six alternating pairs, $[d, t], [z, s], [b, p], [g, k], [\gamma, x], \text{and} [v, f]$, each of which score between 31 and 150 individually\(^8\). On the other hand, the support for lower-scoring rules comes almost exclusively from individual pairs. As a result, the generalization mechanism also allows the model to distinguish the voicing alternation from the spurious pair $[j, u]$ that resulted in a false positive in Experiment 1. This pair’s highest-scoring contextualized rule $(a, [1, 1, 0, 0, 1])$, ranked third in Fig. 2, cannot compete with the devoicing rule because no other pairs significantly contribute to its structural change and context.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{Figure 2. Experiment 2: Scores of highest scoring contextualized rules and mean}
\end{figure}

The generalization mechanism also helps the model find lower frequency pairs, such as $[v, f]$ with a contextualized score of 31.39, because they belong to a high-scoring contextualized

\[^8\] Interestingly, the devoicing rule is also supported by one low-scoring voiced-voiceless pair $(\gamma, j)$ that was not explicitly included in the applied rule but falls within the same natural class of alternating segments.
rule. These low frequency pairs were not found by the original model in Experiment 1b. PLND report low frequency pairs were also a challenge for the allophonic alternations they examined. In particular, PLND’s model failed to detect two of the lowest frequency alternating pairs in their experiment with sonorant voicing alternations. A follow-up study showed that a similar model tested on Japanese and French also failed to detect some low frequency alternating pairs (Le Calvez, Peperkamp, & DuPoux 2007). The proposed rule generalization mechanism provides a way for low frequency pairs to harness the regularities created by more robust patterns of alternation. In sum, the contextualization and generalization mechanisms favor systematic alternations that create consistent structural changes at consistent contexts while penalizing spurious pairs with idiosyncratic structural changes not relatable to changes in other pairs.

7. Experiment 3: Learning generalized rules and contexts

Experiment 3 addresses several further issues. First, the contextualized rules of the preceding section represent generalization about the alternating features and individual contexts of application; however, as discussed in Section 3, the learner must also be able to generalize over contexts. The model must be able to learn general rules that apply across multiple, related contexts, as in the vowel nasalization rule that applies before the class of nasal consonants discussed earlier. We argue that generalization by context must be constrained by linguistic principles. Another goal of this section is to demonstrate that the resulting model is also able to detect allophonic alternations, like PLND’s model. We therefore test the model’s performance on both allophonic and non-allophonic alternations.

7.1. Method

The extensions build on the model from Experiment 2, adding a context generalizing mechanism that allows contextualized rules with the same structural change to merge, creating generalized rules. Since every structural change gives rise to one contextualized rule for every context, unconstrained merging would simply merge all contexts together, but the goal of context merging is to identify just those contexts that condition the alternation. For example, in the case of vowel nasalization, there is one contextualized vowel nasalization rule for each possible context, but the goal of context generalization is to find the set of contexts, in this case nasal
consonants, that predictably condition nasalization. This is not straightforward since the contexts conditioning the alternation do not necessarily correspond to the highest scoring contextualized rules, and it is not known a priori how many contexts should be merged. Restricting contexts to natural classes cannot constrain merging since there are natural classes of arbitrary size, and the set of all segments is itself a natural class. We propose two constraints on context merging, similar in spirit to the linguistic filters in PLND’s original model, which restrict merges to natural classes of segments conditioning the alternation.

\begin{equation}
\text{Generalized rule score} \\
\text{Score}_{(c,j)} = \sum_{c \in C \text{ where SCC and SVC hold}} \text{Score}_{(c,j)}
\end{equation}

\textit{Shared Change Condition (SCC):} \\
To merge, contexts must share feature values for any non-zero values in $\vec{f}$. 

\textit{Shared Values Condition (SVC):} \\
To merge, contexts must not differ along more than one feature.

First, because phonological alternations are often conditioned by multiple phonetically related contexts, we require contexts to be similar to each other in order to merge. Specifically, the \textit{Shared Values Condition} (SVC), defined in (5), restricts merging to contexts that differ along at most one feature. Second, because the contexts conditioning a phonological change are usually related to the alternating features, we require the contexts to be related to the features specified in the structural change. Specifically, the \textit{Shared Change Condition} (SCC), also defined in (5), requires that for any alternating feature in the feature change vector $\vec{f}$, contexts may merge only if they share values of that feature. In other words, contexts must be consistent with respect to values of alternating features, either by matching those values or by uniformly failing to match them. Intuitively, this condition allows merges for either assimilatory or dissimilatory processes.

These conditions are conservative since merging is allowed only if these strong conditions are met. While these conditions may ultimately need to be relaxed somewhat, it is important to keep in mind that the model does not require alternations to satisfy these conditions in order to be detected. It requires them to satisfy these conditions in order to be generalized
Thus, alternations that fail to satisfy these conditions may still be learned as more specific rules, a point we return to in the discussion.

7.2. Data

The data used in this experiment is a modified version of the Dutch word-final devoicing data, with an additional rule applied in which all simple vowels are nasalized before nasal stops, [n], [m], [ŋ], and [ɲ]. In this experiment, an additional feature [vocalic] is added to the original feature system. This is a binary feature, which has a value of 1 for vowels and 0 otherwise, and is necessary in order for the model to be able to refer to the set of all vowels. The model is now attempting to find the word-final devoicing rule as well as the vowel nasalization rule, which is expected to have a structural change [0, 0, 0, 1, 0, 0] and a set of contexts {n, m, ŋ, ḃ}.

![Diagram](image)

**Figure 3. Experiment 3: Scores of highest-scoring generalized rules by merging conditions**

7.3. Results

Fig. 3 shows the top-scoring generalized rules for four different merging condition combinations: no merging conditions, only SCC, only SVC, and both SVC and SCC. The rules are shown with the structural change vector and a representative set of supporting contexts. The striped bars
highlight the generalized rules with structural changes corresponding to word-final devoicing and vowel nasalization. These rules are the top-scoring generalized rules in each of the four conditions. Thus, the model succeeds in separating robust structural changes from spurious alternations regardless of merging condition. This shows that the extended model is capable of detecting both allophonic (vowel nasalization) and non-allophonic (devoicing) alternations.

Furthermore, the task of the learner also includes making appropriate generalizations about conditioning contexts, and the simulation results indicate that the merging conditions are necessary in order to constrain context generalization. Without either merging condition, the model generalizes each structural change to all contexts: there is nothing preventing the model from summing over all contexts for a given structural change. In contrast, with both merging conditions in place, the model succeeds in properly restricting context generalization. In the case of word-final devoicing, the merging conditions correctly restrict the context to ‘#’ by preventing ‘#’ (defined as a zero-valued feature vector) from merging with other segments.

The merging conditions also correctly restrict merging for the vowel nasalization case, which is more interesting since it requires generalization to a particular set of conditioning contexts: the nasal consonants \{n, m, ŋ, ɲ\}. In this case, the SCC restricts merging by requiring that merged contexts be consistent with respect to the alternating feature [nasal]. As shown in Fig. 3, generalized rules with both nasal and non-nasal contexts are created in the SCC only condition (although the desired nasal context has a higher score). The SVC is needed in order to eliminate the nasal vowels, which are not conditioning contexts for this rule. The SVC requires contexts to differ by at most one feature and therefore prevents the merging of nasal vowels and consonants, which differ on [sonority], [place], and [vocalic]. The SVC also prevents the merging of all non-nasals segments since these differ along a number of features.

7.4. Discussion

In general, the goal of the merging conditions is to allow highly scoring contexts to combine into generalized rules as long as the resulting contexts are sufficiently coherent (SVC) and related to the structural change (SCC). These conditions are successful in constraining context generalization for both the allophonic and non-allophonic alternations discussed above and
therefore represent a significant step forward in the modeling of the learning of phonological alternations. However, these conditions are not flexible enough to allow appropriate generalization of contexts for all phonological rules. For example, when the model is given data just like that of Experiment 3 but also with the liquid devoicing rule of Experiment 1 (not shown), it does not succeed in learning a single general rule of liquid devoicing. This is because the SVC prevents the model from combining all voiceless obstruents into a single class (they differ in sonority as well as place). The SCC, on the other hand, allows a broader set of contexts but fails to rule out voiceless sonorants. The final model therefore fails to merge all the relevant context segments; instead, it splits up the liquid devoicing rule into several less robust rules with overly narrow classes of contexts that do not all individually stand out from the spurious rules.

In a noisy, statistical setting, there is no hard line between conditioning and non-conditioning contexts, and the learner cannot assume conditioning contexts will create perfectly complementary distributions. Instead, the statistical regularities in the data give rise to a continuum of dissimilarity across contexts, and the learner must employ some restraint in order to identify the relevant set of conditioning contexts. In the proposed model, the degree to which an individual context is conditioning for a particular structural change is defined as the contextualized rule score. The learner relies on this score to guide its generalization, favoring generalized contexts that include highly conditioning individual contexts. At the same time, the learner should not generalize indefinitely as this would result in merging all contexts. The solution explored here is to impose hard restrictions on what is allowed to constitute a set of merged contexts, via the merging conditions. As the liquid devoicing example discussed above illustrates, these restrictions are too strong to capture the full range of conditioning contexts observed in phonological systems of the world. However, the conditions are an example of the kinds of linguistic principles that could be used to constrain generalization. In particular, the results presented here suggest that learning biases favoring more coherent classes of contexts and classes of contexts relatable to the structural change are needed. The proposed merging conditions can serve as a foundation for further modeling work. A promising direction for future work is to explore softer versions of the merging conditions, for example by assigning higher scores to classes of contexts with more cohesive members. In this way the balance between
generalization and linguistic principles could be made more flexible.

8. Conclusion
While the original PLND model was able to learn alternating pairs for allophonic alternations, the extensions proposed here have made further progress on the learning of alternations. First, the calculation of contextualized scores extend the scope of the model, allowing it to learn non-allophonic alternations in addition to allophonic ones. Second, rule generalization gives the model the ability to make generalizations about structural changes, to detect low frequency alternations supported by systematic alternations, and to eliminate isolated spurious pairs. Third, context generalization allows the model to learn more general rules, specifically, those occurring at multiple related contexts. Overall, the extensions broaden the scope of the model, improve learning of low-frequency alternations, and provide the model with the ability to generalize about alternations in a way that better corresponds to human learning.

In order to learn general rules of alternation, the extended model incorporates linguistic principles at several levels. It employs Peperkamp et al’s (2006) linguistic filters to restrict attention to phonetically natural alternating pairs, and the extensions we propose make further use of linguistic principles. In reformulating the statistical component of the model, the contextualized rule mechanism incorporates linguistic principles favoring systematic alternations involving shared structural changes in particular contexts. As discussed in detail in the previous section, context generalization crucially relies on linguistic principles implemented via the merging conditions. Finally, all of these mechanisms assume the ability to represent and access abstract phonological features. These principles are integral to the model’s successful learning and generalization of phonological rules and its ability to separate these rules from spurious statistical regularities, suggesting that human learning may be constrained by similar principles.

The proposed extensions were motivated in part by experimental findings on human learning of alternations. An important avenue for future work is to explicitly test the predictions of the model to determine the extent to which human learning and generalization reflects the biases of the model. There is an ongoing debate in the literature regarding the constraints, or lack thereof, needed to explain human learning of phonology (see e.g. Moreton 2008, Seidl &
Buckley 2005). Computational models such as this one enable explicit testing of the predictions made by different hypotheses about constraints on learning. The predictions of the model with and without the various filters and conditions can be compared to experimental results to determine which constraints are needed. As mentioned earlier, one interesting property of the final model is that it represents a kind of intermediate position wherein the learner is biased to favor systematic alternations expressible in terms of natural classes, but it is still capable of learning robust, idiosyncratic alternations as constellations of specific individual rules. There are some similarities between the model’s preference for general, systematic alternations and what Moreton and Pater (to appear) call a complexity bias. The present paper suggests that the preferences for general, systematic alternations are needed for successful learning; an important question for future work is the extent to which the proposed biases can also shed light on the constraints on learning more generally. Exploring the predictions of the model and the biases it incorporates could be used to gain deeper insight into the nature of the biases needed to explain the constraints on human learning.

This work takes a substantial step forward in understanding how general rules of alternation can be learned from distributional information, but much further work is needed to fully understand human learning of phonological structure. In addition to the softening of the linguistic filters and conditions discussed earlier, there are a number of promising directions to explore. These include enriching the model’s phonological representations, especially of contexts, learning of rule interactions, and exploring the relationship between learning of alternations and learning of the lexicon or morphology. The model proposed here provides a foundation for further modeling work of these and other important aspects of phonological learning.

Appendix A: PLND model components

(6) Relative Entropy Definition of Allophone (Peperkamp et al 2006)

\[
In a pair of segments \( s_a \) and \( s_d \), the allophone is defined as:
\[
s_a = \max_{s_a,s_d} \left[ \sum_c P(c|s) \log \frac{P(c|s)}{P(c)} \right]
\]

The other segment, \( s_d \), is defined as the default.
Filter 1 (Peperkamp et al 2006)

Allophonic distributions of segments $s_a$ and $s_d$ are spurious if:

$$\exists s \ [\forall i \in \{1, \ldots, 5\}, v_i(s_a) \leq v_i(s) \leq v_i(s_d) \text{ or } v_i(s_d) \leq v_i(s) \leq v_i(s_a)]$$

*With $v_i(s)$ the $i$th component of the vector representation of $s$. 

Filter 2 (Peperkamp et al 2006)

Allophonic distributions of segments $s_a$ and $s_d$ are spurious if:

$$\exists i \in \{1, \ldots, 5\}, \left| \sum_{s \in \{s_a\}} (v_i(s_a) - v_i(s)) \right| > \left| \sum_{s \in \{s_d\}} (v_i(s_d) - v_i(s)) \right|$$

Appendix B: Effect of Linguistic Filters on False Positives

Appendix C: Dutch feature vectors

Segments are represented as feature vectors with the following values:

- **Place**: bilabial 1, labio-dental 2, dental 3, alveolar 4, post-alveolar 5, palatal 6, velar 7, uvular 8, glottal 9
- **Sonority**: voiceless stop 1, voiced stop 2, voiceless fricative 3, voiced fricative 4, nasal 5, lateral 6, rhotic 7, glide 8, high vowel 9, mid vowel 10, low vowel 11
- **Voicing**: voiceless 0, voiced 1
- **Nasality**: oral 0, nasal 1
- **Rounding**: unrounded 0, rounded 1
- **Vocalic** (added in Experiment 3): non-vowel 0, vowel 1

Figure 4. Experiment 1a and 1b: Effect of linguistic filters and threshold on false positives
References


