Towards an Algorithmic Account of Underlying Forms

Caleb Belth

While phonological theory has often made use of the concept of the underlying form, debates have raged regarding how abstract underlying representations ought to be allowed to be, and a number of arguments have been given that phonology should abandon such representations altogether. In this paper, we consider an algorithmic, learning-based approach to the question. We propose a model that, by default, constructs concrete, surface-true representations of morphemes. When and only when such concrete representations make it challenging to generalize in the face of the sparse statistical profile of language, our proposed model constructs abstract underlying forms that allow for effective generalization. We consider the highly agglutinative language, Turkish, and the heavily-studied Dutch noun voicing alternation as two case studies. We demonstrate that the underlying forms that our model constructs account for the complexities of Turkish phonology resulting from its multifaceted vowel harmony, and enable the highly-accurate prediction of novel surface forms, demonstrating the importance of some underlying forms to generalization. We then argue that the model provides a possible learning-based account of why Dutch-learning children do not productively extend voicing alternations: such alternations are not prevalent enough in a child's input to prevent the construction of productive and accurate morphophonological generalizations from concrete underlying forms.

1 Introduction

A traditional conception of phonological theory involves abstract underlying representations (URs) together with phonological processes (stated as rules or constraints) mapping between this abstract level of representation and a concrete, surface-level representation. Debates in the 1960's and 1970's questioned how abstract URs should be allowed to be (Hyman, 2018, p. 597), with a particularly famous article by Kiparsky (1968) arguing that the positing of non-concrete representations should only be done when motivated. Any perception of this debate as fading in subsequent years is probably better attributed to the field moving on to other questions than it is to a satisfactory resolution of the debate (Anderson, 2021).

Indeed, some phonologists have taken the position that URs should not be used in phonological theory because doing so is "(i) wrong, (ii) redundant, (iii) indeterminate, (iv) insufficient, or (v) uninteresting," as Hyman (2018, p. 591) summarized the objections. Meanwhile, much of the work on learning phonology has either focused on surface restrictions (e.g., Hayes and Wilson 2008) or continued to assume URs (e.g., Tesar and Smolensky 1998; Boersma 1997), abstracting away from the question of how (and if) such representations are constructed (see Jarosz 2019 for a summary).

One of the main justifications for the use of underlying representations is to capture generalizations. For example, the form of the English plural affix—[z], [s], or [əz]—is predictable from the stem-final segment, matching the final segment's voicing and separated from sibilants by [ə] as in (1).

(1) [dɑg-z]
[kæt-s]
[hɔrs-əz]

Positing an underlying /-z/ derived by process into [z], [s], or [əz] allows this generalization to be captured. However such an analysis is not strictly necessary. The allomorphs could each be listed along with a set of sounds each occurs after, or the apparent relationship between singulars and plurals could be ignored altogether and both forms could simply be memorized.

In some cases, experimental and acquisition evidence suggests that learners do have productive, rule-like knowledge. For instance, children overextend morphophonological generalizations (e.g., MacWhinney 1978) and apply them to nonce words (e.g., Berko 1958). But in other cases, experimental and acquisition evidence suggests that learners do not have productive, rule-like knowledge. Some Dutch nouns, for instance, alternate in form between the singular and the plural. Stem-final obstruents are syllabified as a coda in the singular, but as an onset in the plural when the plural is formed with a [-ən] suffix. Due to devoicing of obstruents in final position, this sometimes leads to stem-final obstruents being voiced in the plural and voiceless in the singular, as in (2a), while others are voiceless throughout the paradigm, as in (2b).

(2) a. [bɛt] ‘bed’ [bɛdən] ‘bed’-PL
b. [pɛt] ‘cap’ [pɛtən] ‘cap’-PL

The underlying forms of alternating Dutch nouns are often interpreted as having underlyingly voiced obstruents that are neutralized in final position; for instance, /bɛd/ is taken as the underlying form of ‘bed,’ which is transformed into [bɛt] by a productive devoicing rule. However, acquisition and experimental evidence shows no evidence of Dutch-learning children extending the voicing alternation productively (Zamuner et al., 2006; Kerkhoff, 2007; Zamuner et al., 2012). For instance, children have difficulty recognizing or producing [slɑt] as the singular form of the plural nonce word [slɑdən].[1]

How then are we to choose from the possible analyses of (1)? Is the desire to capture a generalization sufficient motivation to choose the /-z/ analysis? In this chapter we propose a learning-based approach to this question. Specifically, we propose a computational model that assumes, by default, that underlying forms are concrete.[2] We present evidence for this in § 1.1. The model attempts to form morphological generalizations out of necessity to deal with the sparse statistical profile of language (Yang 2016, ch. 2; Chan 2008), which we review in § 1.2. The question then becomes learning-based: when does surface-alternation of a morpheme prevent the learner from forming morphological generalizations from concrete representations? In some—but critically not all—cases, surface-alternations are pervasive enough to drive the learner to resort to abstract URs in order to effectively generalize. We present the model in § 2.

We evaluate the model on natural-language corpuses of the highly agglutinative language Turkish, demonstrating both when abstract URs are necessary for generalization and when they are not (§ 3). When combined with the model from Belth (2023b) for learning local and non-local alternations, the proposed model achieves high accuracy generalizing to held-out test words (§ 3.4). We then evaluate the model on a natural-language corpus of Dutch nouns (§ 4). The alternating nouns are not prevalent enough to drive our model to construct abstract URs, and thus allows for highly accurate generalization to held-out test words without a productive generalization for the alternation. This provides a possible learning-based account of the fact that Dutch-learning children show no evidence of having productive knowledge of the voicing alternation.

[1]See § 4.1 for a detailed discussion.

[2]By *concrete* we mean that allomorphs are not collapsed into a single underlying category.

1.1 Studies Pertaining to Lexical Representations

The view that representations of words are, at least initially, minimally abstract receives support from a range of theoretical perspectives. Kiparsky (1968) observed that in the absence of alternation, children have no reason for constructing an abstract representation of a morpheme. This has been incorporated into Optimality-Theoretic-approaches to learning, which usually posit that children's starting point is to construct underlying forms identical to their observed surface realization (Hayes, 2004; Tesar, 2013). The same position is proposed by Ringe and Eska (2013), who observe that historical sound changes support the position. Below, we interpret lexical and experimental studies, which we view as providing support for this widely-adopted position.

1.1.1 Lexical Studies

In a detailed study of the Providence corpus (Demuth et al., 2006) of the CHILDES database, Richter (2018, 2021) studied the acquisition of the English flap [ɾ] as an allophone of an abstract underlying /T/ phoneme. Richter found a U-shaped development curve, which is characteristic of the construction of a linguistic generalization. Up to 3;0, children produce the flap [ɾ] in the intervocalic contexts that it would be expected in (i.e. where the following vowel is unstressed). However, between 3;0 and 5;0, children show in increased rate of producing *[t] instead of [ɾ] in those contexts. The children eventually return to accurate production of [ɾ] in the expected contexts by around the time they learn to read.

It thus appears that children are initially positing underlying /ɾ/ for words with surface [ɾ]. The fact that children begin to produce *[t] in lieu of [ɾ] suggests that children eventually construct an abstract underlying form that either includes both [t] and [ɾ] (e.g., /T/), or that replaces /ɾ/ with /t/, such that contexts with surface [ɾ] must be derived by a productive process.

1.1.2 Experimental Studies

Studies of early infant linguistic development have found that infants show an initial sensitivity to contrasts (e.g. [b] ∼ [p]) regardless of whether a given contrast is linguistically significant in their native language, but that the sensitivity to non-native contrasts declines with age and linguistic experience (Werker and Tees, 1984; Kuhl et al., 1992). This change appears to be mediated by linguistic experience, which improves the recognition of acoustically non-salient native contrasts

(Narayan et al., 2010) and does not degrade perception of non-native contrasts that do not conflict with the child's developing phonology (Best et al., 1988).

The ability to detect two sounds as distinct requires that children record the acoustic feature(s) that differentiate them. This supports the conclusion that children's phonetic representations initially record acoustic information concretely. The fact that the shift in perception of contrasts with age is mediated by linguistic experience suggests that organization into abstract structures is not the default, but rather a response to linguistic exposure.

When we turn to experiments pertaining to knowledge of alternations, results suggest that learners' default is to assume that novel words do not alternate. Dutch, in particular, has received a great deal of attention. In our case study of Dutch in 4.1, we discuss experiments by Zamuner et al. (2006, 2012) in some detail, which support the hypothesis that children initially store words concretely.

Coetzee (2009) performed a similar experiment in an artificial language experiment, and its results corroborate the natural language results from Zamuner et al. (2006, 2012). Coetzee's artificial language alternation was inspired by another alternation in Dutch. In some noun paradigms, a vowel that is short in the SG lengthens in the plural, when a stem-final consonant is re-syllabified as the onset of the syllable containing the plural suffix, which leaves the vowel in a stressed, open syllable. Examples from Coetzee, p. 110 are shown in (3a). However, not all monosyllabic SG nouns with short vowels exhibit the alternation, as shown in (3b), and whether a noun is alternating or non-alternating is not phonologically-predictable.

(3)	a.	[xɑt]	'hole'	[ˈxaː.tən]	'holes'
		[spɛl]	'game'	[ˈspeː.lən]	'games'
	b.	[kɑt]	'cat'	[ˈkɑ.tən]	'cats'
		[stɛl]	'set'	[ˈstɛ.lən]	'sets'

When presented with singular-plural pairs from an artificial language modeled after the Dutch vowel-lengthening alternation, in which half of the words alternated and half did not, Coetzee found that learners were able to learn which words alternated, but did not extend the alternation to novel words. However, when a substantial majority of the exposure data were alternating paradigms, the learners did begin to extend the alternation to nonce paradigms, suggesting that sufficient evidence of surface alternation can lead learners to abandon the default creation of concrete underlying forms.

Together, these natural language and artificial language experimental results suggest that learners by default construct concrete lexical representations, but sometimes move away from this default when substantial amounts of surface alternation are present in the linguistic data they are exposed to.

1.2 Sparsity Necessitates Generalization

1.2.1 The Frequency Distribution of Language Necessitates Generalization

The statistical distribution of words in a corpus of reasonable size consistently follows Zipf (1949)'s "law," which states an inverse relationship between the rank frequency of a word type and its token frequency. Specifically, Zipf's law predicts that the second most frequent word (rank 2) is roughly one half as frequent as the most frequent word, the third most frequent word (rank 3) is roughly one third as frequent as the most frequent word, etc. An example relationship between frequency and rank of word types, in log scale, is shown in Fig. 1, which we derived from the child-directed speech in Roger Brown (1973)'s acquisition study of English. The substantive implication of this robust empirical observation is that most of a language's words occur very infrequently and only a few occur frequently. It follows from this that children's early mental lexicons—over which they learn the core of their language's morphophonology—contain only a small fraction of the entire vocabulary of the language. Thus, in order to effectively use the long-tail of language, children must aggressively generalize from their early mental lexicons.

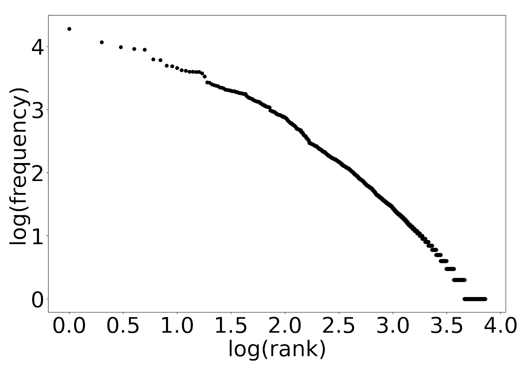


Figure 1: The log frequency of child-directed word types in Brown (1973)'s acquisition corpus approximately follows the famous Zipfian distribution, in which a few words occur very often and most words occur only a few times.

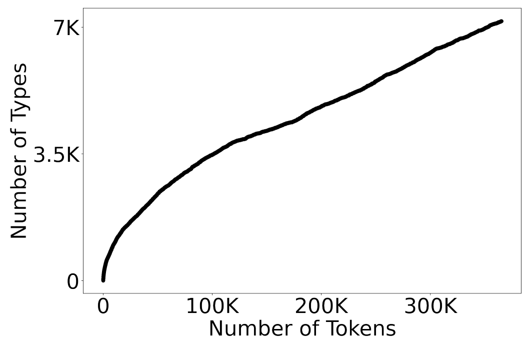


Figure 2: The rate of new child-directed word types in Brown (1973)'s acquisition corpus decreases as the number of tokens increases, but never approaches an asymptote.

1.2.2 The Indefinite Size of Language Necessitates Generalization

Another major statistical observation about corpuses is Heap's law (sometimes called Herdan's law) (Herdan, 1960), which gives the number of word types as a function of the number of tokens, and has the intuitive interpretion of *diminishing returns*: As the number of tokens increases, the number of types increases, but at a decreasing rate. However, this growth function does not have an asymptote. It thus appears that vocabulary size does not have a non-arbitrary limit (Chan, 2008, p. 24). This can be seen in Fig. 2, where the number of word types in Brown (1973)'s child-directed corpus increases rapidly for the first tokens, but the rate of new types slows down as the number of tokens increases, while never appearing to approach an asymptote.

This slowing down of new word types entering the child's mental lexicon after the most frequent types corroborates the implications of Zipf's law: children must generalize from their early mental lexicons in order to accommodate the new words that gradually and persistently enter at later points.

1.2.3 Paradigm (Un)saturation Necessitates Generalization

A major manifestation of the fact that most word types occur very infrequently and new words enter the vocabulary at decreasing rates is that morphological paradigms are generally incomplete. Chan (2008) introduced the concept of *paradigm saturation*, which measures, for each lemma, the fraction of the total number of inflectional categories in the corpus for which the lemma's inflectional form is attested in the corpus (Yang, 2016, p. 21). For example, English verbs have six inflectional categories, as exemplified with the verb *walk* in (4).

(4) Infinitive: To <walk>
First and second person present tense: I/you <walk>
Third person singular present tense: She <walks>
Progressive: They are <walking>
Past tense: They <walked>
Past participle: They have <walked>

If a corpus (or a child's mental lexicon) contains, e.g., four of these six forms, then the paradigm saturation for the lemma <walk> would be 2/3.

Chan (2008, Tab. 4.4) reports that the average saturation across 16 corpuses of different languages was only 71.3%. The situation seems quite dire for languages with rich morphologies. For example, Chan's table reports the *maximum* saturation of any stem in Finnish—an agglutinative language—to be 40.3%. This makes concrete the abstract observation from § 1.2.1-§ 1.2.2 that the sparsity of language requires the child to generalize from their early mental lexicon to the long tail of the linguistic distribution. Suppose a child knows some singular and some plural nouns, as shown in Tab. 1. For some words, the child knows both the singular and the plural forms (e.g., *dog* and *dogs*); for others they know only one (*cat* but not *cats*). At this stage, a word like [kæt] has the same intrinsic status as *Wug* in Berko (1958)'s seminal study: namely, the child knows the SG, but to use the PL must produce a novel form. Here, the motivation for predicting the plural form arises on the occasion of needing to use that form but finding an empty paradigm cell. The statistical profile of language—in particular the results from Chan (2008)—demonstrate that such a scenario is pervasive, and perhaps the driving force behind the construction of a morphophonological grammar.

1.2.4 Summary of Sparsity's Implications

The previous subsections demonstrate that the empirical conditions of language acquisition require children to generalize beyond their early vocabulary, thus constructing productive generalizations. This observation forms the basis for our model (§ 2). In § 2.4, we introduce the Tolerance Principle (Yang, 2016), which provides a precise, cognitively-grounded tipping point at which generalization becomes preferable to memorization. The derivation of the tipping point, presented

Table 1: A hypothetical state of a child's mental lexicon for some English nouns. The child only knows both the SG and PL forms for some words.

SG	PL	Gloss
[kæt]	??	cat ~ cats
[dɑg]	[dɑg-z]	dog ~ dogs
[hɔrs]	[hɔrs-əz]	horse ~ horses
??	[bɝd-z]	bird ~ birds
[wɛb]	??	web ~ webs
[seɪf]	[seɪf-s]	safe ~ safes
??	[mæp-s]	map ~ maps

in Yang (2016, ch. 3), is grounded in the statistical profile of language data reviewed above.

2 Model

2.1 Model Input

The input to the model is a set of morphologically-analyzed surface forms. An example input of nine forms is shown in Tab. 2. These word forms are processed by the model incrementally, modeling the growth of a learner's lexicon.

While morphological segmentation is an important area of study in its own right, we believe it is a justified assumption given experimental evidence that infants can effectively morphologically segment nonce words. These results have been observed for French-learning 11mo-old (Marquis and Shi, 2012) and English-learning 15mo-old (Mintz, 2013) infants. The finding is corroborated by results for 15mo Hungarian-learning infants, despite the high-level of agglutination in Hungarian (Ladányi et al., 2020).

2.2 Model Output

The output of the model is a lexicon, which contains a representation for each morpheme, and a lexicalized list of any input word forms not decomposable into those morphemes. The representation of a morpheme may be concrete or abstract, but we will refer to the representation constructed in the lexicon as a UR, regardless of

Surface Form	Morphological Analysis
1. [buz-lɑr]	‘ice-PL’
2. [kɯz-lɑr]	‘girl’-PL
3. [el-ler]	‘hand-PL’
4. [jer-ler-in]	‘place’-PL-GEN
5. [søz-ler]	‘word-PL’
6. [dɑl-lɑr-ɯn]	‘branch’-PL-GEN
7. [sɑp-lɑr]	‘stalk-PL’
8. [jyz-yn]	‘face’-GEN
9. [ip-ler-in]	‘rope’-PL-GEN

Table 2: An example Turkish input consisting of morphologically-analyzed surface forms.

its abstractness. We treat surface and underlying representations, whether concrete or abstract, as sequences of segments, where each segment is a set of distinctive features.

As discussed by Ettlinger (2008, sec. 4.3.4), the term *abstract* is not always used consistently. A UR is sometimes called *abstract* because it lacks the phonetic detail of an actual speech sound (e.g., /D/ as an alveolar stop lacking a voicing specification), or because it contains different segments from a surface form. In our view, concrete long-term memory representations can be used via direct access while non-concrete underlying forms introduce discrepancies from surface forms, which thus necessitates the construction of a generalization (e.g., rule or constraint) to non-trivially map between these levels of representation. We thus consider a UR *concrete* if all word forms where the morpheme surfaces as something other than the UR are lexicalized. Conversely, we consider a UR *abstract* if it is the result of collapsing the morpheme's surface forms into an underlying category whose realization must be inferred from a generalization (rule or constraint). For the time being, our treatment does not differentiate degrees of abstractness. For example, our use of *abstract* includes both /d/ derived into [t] by devoicing generalization /d/ → [t] / [−voi] __ and /D/ derived into [d] / [t] via a voice assimilation generalization, while /D/ would usually be thought of as *more* abstract than /d/ if [d] sometimes surfaces while /D/ never does. Future work will consider the question of degrees of abstractness.

We assume, following prior work (§ 5), that each morpheme has a single UR. Future work will consider scenarios where this may not be the case. Future work will also consider what changes are necessary to handle nonconcatenative morphology.

2.3 Model Description

By default, the model creates a concrete UR for each morpheme. Prior work (§ 5) often resorts to phonological processes to produce the various surface forms of a morpheme at the first instance of surface alternation. Our model differs from this approach by treating underlying forms as concrete even after the first instance of surface alternation. Instead of immediately collapsing surface forms into a single, abstract UR, our model simply lexicalizes all word forms in which a morpheme occurs as something other than its most frequent form. It is only when the resulting lexicalization becomes unsustainable (see § 2.4) that the model then constructs abstract underlying forms from which the surface realizations are derived by morphophonological process.

The pseudocode for the algorithm is shown in (5).[3] As discussed in § 2.1, the input to the model is an incremental stream of morphologically-analyzed surface forms. Whenever the model receives a new surface form (5; step 1), it initially creates a concrete underlying form for each morpheme, storing the most frequent[4] form of the morpheme concretely (5; step 3), and lexicalizes any wordforms that contain a different form of the morpheme (5; step 8). However, if too many wordforms in the lexicon are exceptions—where the measurement of "too many" occurs as described in § 2.4—the model instead constructs an abstract UR (5; Step 5) and then learns a phonological process, via a separate model (see § 2.6), to account for the resulting alternation.

(5) **Input:** Incremental stream of morphologically analyzed SRs

1. **While** surface form in input **do**
2. – **For** morpheme in segmentation **do**
3. — Morpheme UR ← most freq form
4. — **If** too many alternative forms **do**
5. —– Construct abstract UR
6. —– Learn phonological process
7. **— Else do**
8. —– Lexicalize exceptions

[3]Code is available at https://github.com/cbelth/underlying-forms-SCiL

[4]If two forms are equally frequent, the choice of UR is arbitrary; we used lexicographic ordering to make the ordering complete.

Meaning	UR	Plural Form
PL	/lɑr/	N/A
‘ice’ ‘girl’	/buz/ /kɯz/	**Stem-PL**
‘hand’	/el/	/el-ler/

Table 3: When the first three words from Tab. 2 enter the lexicon, the stems and plural affix are all stored concretely (left two columns). The plural form of the ‘ice’ and ‘girl’ stems are predictably decomposable into their concrete stems and the PL affix (denoted with the boldface concatenation), so those forms need not be stored in the lexicon. However, with /-lɑr/ as the UR of the plural, the plural form of ‘hand’ cannot be so decomposed, so it is instead lexicalized.

For example, consider the PL suffix after the first 2 (of 9) inputs listed in Tab. 2 have entered the learner’s lexicon. At this point, the model will be storing the only attested surface form [-lɑr] as the concrete UR /-lɑr/.

When the third word enters the lexicon, our model will lexicalize the form ‘hand-PL’ as /el-ler/, rather than immediately constructing an abstract PL morpheme. This is shown in Tab. 3, where each stem and the plural affix have concrete underlying forms, and the plural form of ‘ice’ and ‘girl’ are formed by suffixing the plural to the stem, but the plural form of ‘hand’ is lexicalized.

By the time all 9 words enter the lexicon, however, there will be 4 instances of [-lɑr] and 4 of [-ler], making it no longer sustainable to keep a concrete underlying form. The difference between these two scenarios and, more generally, the decision of when to create an abstract underlying form, is made by the Tolerance Principle (Yang, 2016), as described next.

2.4 When is Abstraction Needed?

In order to detect what amount of surface alternation prohibits generalization from concrete representations, the model uses the Tolerance Principle, proposed by Yang (2016). The Tolerance Principle is a cognitively-grounded tipping point, which hypothesizes that children form productive generalizations when the number of exceptions to a proposed generalization results in a real-time processing cost lower than that without the generalization. The exact derivation of the Tolerance Principle is provided by Yang (2016, ch. 3), but rests critically upon the empiri-

Morphemes		Word Forms		
Meaning	UR	PL Form	GEN From	PL, GEN Form
PL	/lɑr/	N/A	N/A	N/A
GEN	/in/	N/A	N/A	N/A
'ice'	/buz/		??	??
'girl'	/kɯz/	**ROOT-PL**	??	??
'stalk'	/sɑp/		??	??
'hand'	/el/	/el-ler/	??	??
'word'	/søz/	/søz-ler/	??	??
'face'	/jyz/	??	/jyz-yn/	??
'place'	/jer/	??	??	/jer-ler-in/
'branch'	/dɑl/	??	??	/dɑl-lɑr-ɯn/
'rope'	/ip/	??	??	/ip-ler-in/

Table 4: The left two columns contain morphemes—meaning and form (UR); the right three columns contain word forms. Boldface denotes word forms that can be predictably decomposed into concrete underlying forms, while '/-/' notation denotes word forms that must be lexicalized. The '??' denotes word forms that are unknown. Once all nine words from Tab. 2 enter the lexicon, most forms (6 of 9) cannot be predictably decomposed into concrete underlying forms, so the model constructs abstract URs, as described in § 2.5.

cal observation of linguistic sparsity. The Tolerance Principle has had much prior success in computational modeling and lexical studies (Yang, 2016; Richter, 2018; Van Tuijl and Coopmans, 2021; Richter, 2021; Belth et al., 2021; Payne, 2022), and has been evaluated and supported experimentally (Schuler et al., 2016; Koulaguina and Shi, 2019; Emond and Shi, 2021).

Our model's default treatment of underlying forms as concrete can be stated as a morpheme-specific rule. In the example above, where only the first 2 words of Tab. 2 have entered the lexicon, the rule for the PL form would be (6), which predicts that the PL morpheme is realized as [-lɑr].

(6) If PL then [-lɑr]

The Tolerance Principle threshold, which evaluates a linguistic rule (generalization), is given in (7), where n is the number of items the rule applies to and e is the number of exceptions to the rule.

(7)

$$e \leq \frac{n}{\ln n}$$

Thus, our model tracks—for each morpheme—the number of observed words in which the morpheme appears (n) and the number of those where surface alternation leads the morpheme to be realized as something other than its hypothesized concrete form (e).

If the (7) threshold is met, then the UR remains concrete and the word forms where the suffix is realized as something else are lexicalized[5] as exceptions. For example, when the 3rd item in Tab. 2 enters the lexicon, the realization of PL as [-ler] violates (6). However, with only three word forms containing PL this one exception can be lexicalized, since $1 \leq 3/\ln 3$.

On the other hand, if the (7) threshold is violated—i.e., $n > \frac{n}{\ln n}$—then the model constructs an abstract underlying form. For example, when the 9th item of Tab. 2 enters the lexicon, the realization of PL as [-ler] becomes the 4th of 8 forms in which PL is realized as [-ler] instead of the [-lɑr] predicted by (6). Because $4 > 8/\ln 8$, the model will construct an abstract UR for the PL morpheme.

This is shown in Tab. 4, where the plural is realized as [-lɑr] in 3 plural forms and 1 plural, genitive form, but there are 4 forms that must be lexicalized because

[5]By lexicalization, we mean that the word form is stored in the lexicon verbatim instead of being decomposed into the underlying morphemes. See Tab. 3 for an example.

they instead have the [-ler] form.[6]

Constructing abstract URs introduces discrepancies between URs and SRs for any word forms containing the morpheme, so our model then passes the (UR, SR) pairs implicit in its lexicon[7] to a model that learns phonological alternations to account for the newly-introduced discrepancies. The process of constructing abstract URs is described in § 2.5 and the process of learning what conditions the alternations is described in § 2.6.

2.5 Constructing Abstract URs

The model's first step in constructing an abstract UR for a morpheme is to create the set of forms that the morpheme is realized as. For example, the forms of the GEN affix attested in Tab. 2 are [-in] / [-ɯn] / [-yn], and of the PL affix are [-lɑr] / [-ler].

Next, the model aligns each of the forms. This is trivial for fixed-length affixes (e.g., the case of the PL affix). If the length of the forms are not all the same, then the model counts the lengths of the morpheme's realizations. For example, the dative affix can be realized as [-ɑ] or [-e], but may contain an affix-initial [j] when attaching to a morpheme that ends in a vowel. The model thus counts the number of words in which [-ɑ] or [-e] (length 1) is the realization, and the number in which [-jɑ] or [-je] is the realization (length 2), and chooses the most frequent length as the length of the UR. If a shorter length is chosen, the extra segment(s) are treated as epenthesized; if the longer is chosen, they are treated as deleted. For simplicity, we assume that these segments epenthesize or delete on the left for suffixes and the right for prefixes. This process is not guaranteed to generalize to other languages, so future work will develop a more robust alignment process by more tightly combining the problems of abstract UR construction and rule construction.

Once the forms are aligned, the UR is constructed one segment at a time. Each segment is set to match in features where all realizations of the affix match; features that alternate across forms are unspecified underlyingly. For example, [-lɑr] / [-ler] will lead to /-lAr/, where A is the low, unround vowel with backness unspecified, because both forms agree in the initial and final segments, but the vowel alternates on backness. Similarly, [-in] / [-ɯn] / [-yn] will result in /-Hn/, where H is the high

[6]Note that the PL, GEN of 'branch' is lexicalized because the GEN affix is realized in a form other than [in], not because of the PL affix, which is why that form does not get counted as an exception in the Tolerance Principle calculation for the PL affix.

[7]See § 2.6 for a description of how the set of (UR, SR) pairs is computed.

vowel with backness and roundness unspecified, since [i] and [y] differ in backness from [ɯ] while [i] and [ɯ] differ from [y] in roundness.

2.6 Learning Alternations

When the number of words where the morpheme's surface alternation requires the word be lexicalized becomes too great, the model constructs an abstract UR for the morpheme. This abstract UR introduces a discrepancy between the abstract UR and its surface realization. The model thus constructs a set of (UR, SR) pairs from the lexicon, which it passes to a model that learns a phonological process to derive the various surface forms.

For example, when the 9th item from Tab. 2 causes /lɑr/ to no-longer be sustainable as the PL affix UR, the lexicon is as described in Tab. 4. The surface form for the PL forms of the roots 'ice', 'girl', and 'stalk' are computed by concatenating /lɑr/ to the stem (i.e., Stem-PL), and the remaining six known surface forms, which were lexicalized, are extracted directly from the lexicon. Since the PL is being collapsed into /lAr/, each word's UR is computed by replacing the surface realization of the PL affix with this new UR. Thus, the (UR, SR) pairs at this point would be {(/buzlAr/, [buzlɑr]), (/kɯzlAr/, [kɯzlɑr]), (/sɑplAr/, [sɑplɑr]), (/ellAr/, [eller]), (/søzlAr/, [søzler]), (/jyzyn/, [jyzyn]), (/jerlArin/, [jerlerin]), (/dɑllArɯn/, [dɑllɑrɯn]), (/iplArin/, [iplerin])}.

Learning phonological processes from UR-SR pairs is an active area of study, and many models have been proposed (see Jarosz 2019 for an overview). In this paper we chose the model from Belth (2023b), which is a cognitively-grounded model that provides a unified ability to learn local and non-local alternations, which is important, given Turkish's non-local vowel harmony combined with local processes like voicing assimilation (see § 3.1).

The Belth (2023b) model learns rules to predict the surface form of alternating segments—in this case those that are underlyingly abstract. To do so, the model tracks only dependencies between alternating segments and the segments adjacent to them. If these adjacent segments fail to allow the surface form to be accurately predicted, the model deletes any adjacent segments that prevent the surface form from being predicted, and repeats. The iteratively deleted segments accumulate into a deletion set, the complement of which is interpreted as a tier. The learned rules are applied locally over the tier projection. Because segments are deleted only when adjacent dependencies fail to make the surface form predictable, local processes are a special case, and thus local and non-local processes are learnable by a unified model.

3 Turkish Case Study

This section provides a case study of our proposed model on the highly agglutinative language, Turkish. In § 3.1 we describe some relevant details of Turkish. We then describe the setup of our evaluation in § 3.2. Finally, we present qualitative results in § 3.3 and quantitative results in § 3.4.

3.1 Turkish

Turkish phonology receives attention often because of its apparently complex vowel harmony system. It exhibits both primary front/back harmony and secondary rounding harmony, which is parasitic on height: only [+high] vowels harmonize for roundness. Moreover, Turkish has a number of exceptional suffixes whose vowels do not participate in harmony, and even half-harmonizing suffixes, which have multiple vowels, some of which harmonize and some of which do not. These harmony processes occur alongside other processes, such as local voicing assimilation. The Turkish vowel inventory is shown in (8).

(8)

	front		back	
	unround	round	unround	round
high	i	y	ɯ	u
low	e	ø	ɑ	o

The primary harmony process is observed in affix vowels that alternate between [+back] when the preceding vowel is [+back] and [−back] when it is [−back], as in (9) (examples from Nevins 2010, p. 28; Kabak 2011, p. 3).

(9)	[dɑl-lɑr-ɯn]	branch-PL-GEN
	[jer-ler-in]	place-PL-GEN
	[ip-ler-in]	rope-PL-GEN

The secondary rounding harmony involves suffixal [+high] vowels matching in roundness to the vowel to the left, as in (10) (examples from Nevins 2010, p. 29; Kabak 2011, p. 3). This harmony occurs regardless of whether the vowel to the left is [+high] (10a) or [−high] (10b).

(10)

a. [ip-in] rope-GEN
[jyz-yn] face-GEN
[kɯz-ɯn] girl-GEN
[buz-un] ice-GEN

b. [el-in] hand-GEN
[søz-yn] word-GEN
[sɑp-ɯn] stalk-GEN
[jol-un] road-GEN

Some suffixes have vowels that do not participate in harmony. For example, the suffix [-ki] can attach to a temporal or spatial nominal root to yield adjectival forms as in (11), where the suffix surfaces with the vowel [i] regardless of the final vowel of the stem (examples from Oflazer 1994, p. 144). The PL suffix, which alternated in (9), here harmonizes with the [i] vowel (11b), thus surfacing as [e].

(11) a. [ønʤe-ki] '(the one) before'
[jɑrɯn-ki] '(the one) tomorrow'

b. [ønʤe-ki-ler] '(the ones) before'
[jɑrɯn-ki-ler] '(the ones) tomorrow'

The situation gets more complex, as some suffixes are *half harmonizing*, meaning they have two vowels with one harmonizing and one not.[8] An example is shown in (12a), where the first vowel of the abilitative (ABIL) suffix harmonizes with the vowel to the left, but the second vowel is always [−back] [i] even when the first vowel is [+back] (Kornfilt, 2013). The aorist (AOR) suffix vowel then harmonizes with the abilitative's non-harmonizing second vowel [i] in (12a). The example (12b) demonstrates that the AOR suffixal vowel surfacing as [i] in (12a) is indeed due to harmony, as it harmonizes in (12b) with [o].

(12) a. [jɑz-ɑbil-ir] 'write'-ABIL-AOR
[jyz-ebil-ir] 'swim'-ABIL-AOR

b. [ol-ur] 'become'-AOR

Vowel harmony often goes in hand with other phonological processes, such as voicing assimilation. This can be seen, for example, in the locative (LOC) suffix, which exhibits vowel harmony, but begins with an alveolar stop, which assimilates

[8]The term *half harmonizing* is from Nevins (2010), but in principle, other fractions of vowels (e.g. 1 of 3) could potentially harmonize in vowel harmony systems.

in voicing to the segment to its left, as in (13) (examples from Dobrovolsky 1982; Çöltekin 2010; Kornfilt 2013).

(13) [byro-dɑ] 'office'-LOC
[ev-de] 'house'-LOC
[ʤep-te] 'pocket'-LOC

In the remaining subsections, we demonstrate how our proposed model elegantly accounts for these complexities in Turkish (§ 3.3), and how this allows for novel surface forms to be accurately predicted (§ 3.4). First, though, we introduce the setup and data we used for our experiments (§ 3.2).

3.2 Setup and Data

To simulate learning in Turkish, we constructed two Turkish datasets consisting of frequency-annotated and morphologically-analyzed surface forms (see below). To simulate one learning trajectory, we sampled words with replacement from the corpus, weighted by frequency. Each time a new word form is sampled, the learner adds it to its lexicon. We then investigate the underlying forms of each morpheme, seeing which are concrete and which are abstract (§ 3.3). We then evaluate how accurately the model, combined with a model for learning alternation rules, allows novel surface forms to be predicted (§ 3.4).

We constructed two datasets, called MorphoChallenge and CHILDES. The first used data from MorphoChallenge (Kurimo et al., 2010), which contains a large Turkish corpus annotated with word frequencies. To generate morphological analyses of words, we used Çöltekin (2010, 2014)'s finite state morphological analyzer, which is designed for Turkish. This is similar to the process used in the MorphoChallenge, but is publicly available.[9] We dropped any word in MorphoChallenge that had fewer than 25 occurrences or for which the morphological analyzer failed to provide an analysis. We also removed forms with affixes that are analyzed by Çöltekin (2010, 2014) as having multiple underlying forms. For example, the highly irregular aorist suffix is sometimes described as having four underlying forms: /-Ar/, /-Hr, /-z/, /-null/. Future work will consider scenarios where multiple URs are necessary. This resulted in 22,315 frequency-annotated and morphologically-analyzed surface forms, which we transcribed into IPA.

The second dataset is derived from the child-directed speech in the Aksu (Slobin, 1982) and Altinkamis corpuses of the CHILDES database (MacWhinney, 2000).

[9] https://github.com/coltekin/TRmorph

We computed the frequency of each word in the corpuses and used the same process as above to morphologically analyze each word. This dataset is much smaller, so we did not exclude words with low corpus counts from this dataset. This resulted in 1,727 frequency-annotated and morphologically-analyzed surface forms, transcribed into IPA.

Note that some Turkish suffixes exhibit deletion/epenthesis to avoid CC or VV clusters. These additional processes are at present ignored, because the implementation provided by Belth (2023b) was designed for harmony and disharmony. Future work will extend the implementation to epenthesis and deletion by incorporating Belth (2023a), which handles such processes.

3.3 Suffixes: Abstract and Concrete

Remarkably, the apparent complexity of Turkish vowel harmony, discussed in § 3.1, receives an account when we investigate the output of our model.[10] As before, we will let A denote the Turkish low, unround vowel with backness unspecified (extensionally, {e, ɑ}) and H be the Turkish high vowel with both backness and height unspecified (extensionally, {i, y, ɯ, u}). Moreover, we will use D to denote the alveolar stop with voicing unspecified (extensionally, {d, t}).

We will walk through the complexities exemplified by (9)-(13) one-by-one. First, the PL suffix in (9), which has a low unrounded vowel, participates in front/back harmony, but not rounding harmony because it is not a [+high] vowel. Our model constructed the underlying form /-lAr/ for this suffix, capturing the fact that it only harmonizes for backness.

The GEN suffix in (9)-(10) has a [+high] vowel and participates in both primary and secondary harmony. Our model constructed the underlying form /-Hn/ for this suffix, which captures the surface alternation of this morpheme.

Next, the [-ki] suffix in (11) does not participate in harmony, and our model consistently represents it with a concrete form /-ki/.

For the abilitative suffix in (12), our model abstracts the first, harmonizing vowel, but keeps the second, non-harmonizing vowel concrete -/Abil/.

Lastly, the UR for the locative suffix in (13) is constructed with both segments abstract /-DA/, capturing both the voicing assimilation of the initial alveolar stop and the vowel harmony of the second segment.

These underlying forms lead Belth (2023b)'s model to learn two rules, which

[10] This analysis is performed on a random, frequency-weighted 80% sample of the MorphoChallenge dataset.

allow for the accurate prediction of novel surface forms. On the resulting (UR, SR) pairs, the model learns a vowel harmony rule, which targets both /A/ and /H/ vowels, and enforces harmony with respect to their unspecified values: [back] for /A/ and both [back] and [round] for /H/ (14a). The model automatically constructs a vowel tier and enforces harmony locally over that tier. The model also learns a local voice assimilation rule, which causes /D/ to take its [voi] value from the segment to its left (14b).[11]

(14) a. $\textsc{Agree}([?\text{back}], \{\text{back}, \text{round}\})/[-\text{cons}]\ \_\_ \circ \text{proj}(\cdot, [-\text{cons}])$
     b. $\textsc{Agree}([?\text{voi}], \{\text{voi}\})/[*]\ \_\_$

It is worth noting that others—in particular Nevins (2010)—have similarly argued that Turkish vowel harmony can be elegantly accounted for with an underspecification approach. Our model builds on Nevins (2010)'s observations by providing an explicit computational model that constructs underlying forms, which are consistent with this analysis.

As a further analysis, we show the 10 most frequent affixes in a 1K word sample of the CHILDES corpus in Tab. 5, along with the UR that our model constructed for each. Of the 10 affixes, 7 have been collapsed into abstract forms. However, there are 3 forms (P1S, lH, P2S) that were quite frequent, but are still able to be stored concretely. The P1S and P2S affixes do not have alternating segments in Turkish, so it is expected that these would be concrete. The "lH" affix, as captured by its name, can surface with any high vowel. However, in the training data, the [-lɯ] form occurs 25 out of 32 times, so the 7 words where it surfaces as something else are lexicalized ($7 <= 32/\ln 32$).

3.4 Quantitative Evaluation

We also evaluated how the model enables generalization, when paired with a model for learning phonological alternations. We used our model to learn to map a stem and morphological analysis of a surface form to an actual surface form. For example, given the stem [dɑl] and morphological analysis Stem-PL-GEN, our model's underlying forms for -PL and -GEN are concatenated to the stem to form a UR, to which the generalizations learned by Belth (2023b)'s model can then be applied to predict a surface form, such as [dɑllɑrɯn].

[11]The notation [*] denotes any segment.

Affix	UR	Abstract
PL	/-lAr/	Y
P3S	/-H/	Y
P1S	/-m/	N
GEN	/-Hn/	Y
DAT	/-A/	Y
ACC	/-H/	Y
LOC	/-DA/	Y
VN:INF	/-mA/	Y
lH	/-lɯ/	N
P2S	/-n/	N

Table 5: Top 10 most frequent affixes in a random, frequency-weighted sample of 1K words from the CHILDES dataset, and the URs that our model learned. See http://coltekin.net/cagri/trmorph/trmorph-manual.pdf for a description of affix names.

3.4.1 Setup

We ran the model on both datasets, simulating incremental learning by sampling words with replacement and weighted by frequency, and adding them to the lexicon when sampled. As this process incrementally adds words to the lexicon, our model operates as described in (5). In 250-word increments (i.e., every time the lexicon grows by 250 unique words) for MorphoChallenge and 100-word increments for CHILDES, we evaluated the model by using the rules learned by Belth (2023b)'s model—on our learned underlying forms—to predict the surface form of all the words not in the lexicon. We carried out 5 simulations on each dataset, using different random seeds for sampling on each.

As a comparison, we used a transformer-based (Vaswani et al., 2017) seq-to-seq model. At each training increment (250 words for MorphoChallenge and 100 for CHILDES), we created a random 80/20 train/dev split of the training data to tune hyperparameters. We search over the hyperparameters shown in (15) and choose the combination with the highest accuracy on the dev set. We then trained a model with the best hyperparameters on the entire training set (i.e. re-merging the 80/20 split).

(15) learning rate $\in [0.0001, 0.01]$
number of epochs $\in \{10, 11, 12, ..., 29, 30\}$

embedding dimension $\in \{16, 32, 64, 128, 256, 512\}$
hidden dimension $\in \{16, 32, 64, 128, 256, 512\}$
number of attention heads $\in \{1, 2, 4, 8\}$
number of encoder layers $\in \{1, 2, 3, 4\}$
number of decoder layers $\in \{1, 2, 3, 4\}$

3.4.2 Results

The results are shown in Fig. 3, where the x-axis shows the incremental growth of the learner's lexicon (i.e., the training size), and the y-axis shows the accuracy at predicting novel surface forms at that point during training. The accuracy is computed over all surface forms not currently in the training data. Each subfigure is for one of the two datasets. The MorphoChallenge results (Fig. 3a) are reported up to a size of 3K words, so the test results are on 10s of thousands of novel words.

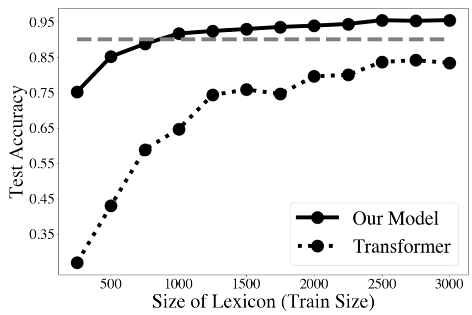
Our model's accuracy is higher than that of the transformer model at all training sizes. The dip in transformer accuracy at 800 training size in CHILDES is likely due to high variance across the 5 runs, and the curve would likely smooth with more simulations.

Our model's performance appears to be consistent with acquisition studies. Altan (2009) found that Turkish-speaking children as young as 2;0 extend vowel harmony to nonce words and overextend it to (adult) lexical exceptions. Studies across languages reveal that a child's vocabulary is quite modest at this age, with an upper bound around 1K words (Fenson et al., 1994; Hart and Risley, 1995; Szagun et al., 2006; Bornstein et al., 2004). The model's performance on both datasets is above 90% accuracy (gray, dashed line) when its vocabulary contains 1K words.

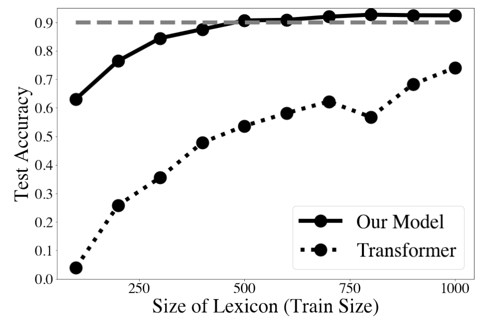
3.4.3 Error Analysis

Of the errors, around 52% result from the model having a concrete form of an affix, which it then errantly predicts for a novel word that exhibits alternation in that affix. For example, there are insufficient forms in the training data to make /ɯp/ as the concrete CV:IP affix prohibitive ($e = 5 <= n = 13/\ln 13$), even though vowel harmony leads it to sometimes surfaces with other high vowels. As a result, novel words like [gel-ip], which take the [ip] form of the affix are mispredicted.

About 47% of the errors are the result of vowel harmony or consonant assimilation being predicted for a novel form that exceptionally does not involve harmony. For example, the word [saat-ler] 'watch-PL' is predicted by our model to be [saat-lar] because the UR for the plural suffix is /lAr/, as it systematically har-



(a) MorphChallenge



(b) CHILDES

Figure 3: Our proposed model's accuracy (averaged over 5 simulations) at predicting novel surface forms. The *x*-axis shows the growth of the learner's lexicon (i.e., the training size). The gray, dashed line marks the 0.9 accuracy point.

monizes. According to a Wiktionary search,[12] the root [saat] is of Arabic origins. Because Arabic has a different vowel system, vowels in Arabic loan words may conform to the Turkish vowel system when entering Turkish, and thus sometimes behave oddly. Indeed, Altan (2009) observed that children may overextend vowel harmony to such words.

The remaining 1% of errors result from very low frequency affixes which are simply unattested in the training data.

4 Dutch Case Study

In this section, we provide a case study of our proposed model on voicing alternations in Dutch noun paradigms. In § 4.1 we describe the relevant details of Dutch. In § 4.2 we describe our experimental setup. Then, in § 4.3 we discuss how the model's results may account for experimental results discussed in § 4.1, and demonstrate that the model's output allows accurate prediction of novel forms in § 4.4.

4.1 Dutch

Dutch, like many other languages (e.g., German and Polish), exhibits a phonotactic restriction against syllable-final voiced obstruents. Beyond the distribution of

[12]https://en.wiktionary.org/wiki/saat#Turkish

voiced obstruents, a primary indication of this restriction comes from certain informative noun paradigms. Many Dutch plural nouns are formed by suffixing [-ən]. When this plural suffix attaches to a stem ending in an obstruent, the obstruent is syllabified as the onset of the syllable containing the plural suffix syllable. In some noun paradigms—such as (16a; data from Zamuner et al. 2012, p. 482)—the stem-final obstruent is voiced in the plural form, but unvoiced in the singular, where it occurs in syllable-final position. In other paradigms, the obstruent is voiceless throughout the paradigm—e.g., (16b).

(16) a. [bɛt] 'bed' [bɛdən] 'bed'-PL
b. [pɛt] 'cap' [pɛtən] 'cap'-PL

Because not all nouns with stem-final obstruents alternate, researchers have taken interest in whether Dutch learners are aware of which nouns alternate and which do not (i.e., are they sensitive to the alternation for nouns that they know?), and, if so, whether learners generalize productively so as to expect stem-final voiced obstruents in novel plural nouns to be voiceless in the singular.

Dutch nouns also frequently occur in the diminutive (both singular and plural), with especially common occurrence in child-directed speech. In fact, some plural diminutive forms of a noun are more frequent than their non-diminutive counterparts. For example, Kerkhoff (2007, p. 113) found that *eendjes* 'ducklings' occurs more times (eight) in her corpus of child-directed speech than does the *ducks* (three). This is an important fact because nouns with voiced stem-final segments in the (non-diminutive) plural are unvoiced in both the singular and plural diminutives, as in the (non-diminutive) singular. For example, in (17) the stem-final obstruent of *paard* 'horse' surfaces as [t] in all forms except for the non-diminutive plural (17b).

(17) a. [part] 'horse'
b. [pardən] 'horses'
c. [partjə] 'horsie'
d. [partjəs] 'horsies'

If the child learns (17c) 'horsie' and (17d) 'horsies' before learning (17b) 'horses'—a plausible scenario given the prevelance of dimnuitves in child-directed speech—then the child receives no evidence of an alternation. This highlights the importance of considering the details of the child's input in understanding what representations and generalizations they may construct.

Many other languages exhibit syllable-final devoicing that manifests in voicing alternations. For example, German exhibits a descriptively similar alternation, in which devoicing of obstruents in final position appears in the form of voicing alternations like [hʊnt] 'dog' ~ [hʊndə] 'dogs.' Because of their descriptive similarity, these cross-linguistic patterns and processes are usually treated as a group. However, what is cross-linguistically descriptively similar may not have the same cognitive status across speakers of the languages.

In an age-controlled experiment, Buckler and Fikkert (2016) found German 3-year-olds show more advanced knowledge of which nouns in their mental lexicon alternate in comparison to 3-year-old Dutch children. The authors argue that this is likely attributable to language-specific factors, including the fact that German uses voicing contrastively over a broader range of obstruents than Dutch does and that Dutch exhibits both progressive and regressive voicing assimilation while German only commonly exhibits progressive voicing assimilation. These cross-linguistic differences may make the voicing alternating in noun paradigms harder to acquire for Dutch learners than German learners. Perhaps more significantly, Buckler and Fikkert (2016) performed a corpus study of German and Dutch nouns in CHILDES (MacWhinney, 2000), and found that the analyzed German child-directed speech contained a much greater number of alternating noun paradigms than that of Dutch, whether counted in type frequency (72 in German vs. 22 in Dutch) or token frequency (1572 in German vs. 158 in Dutch).

Several other studies have investigated whether knowledge of the Dutch voicing alternation is productive for children, and found evidence that it is not. Zamuner et al. (2012) performed a reverse wug test with 2.5-year-old and 3.5-year-old children. A reverse wug test presents the plural form of a nonce noun and asks participants to produce the singular, which flips the classical setting originating from Berko (1958), in which the participants are asked to produce the plural form from the singular nonce. For example, the children were presented with nonce noun plurals where the stem-final obstruent for some was voiced [d], as in (18a), and for others was voiceless [t], as in (18b).

(18) a. [slɑdən]
[kɛdən]
b. [klatən]
[jɪtən]

The nonce-paradigms with the [d] plural forms alternate, since the [d] will occur in final position in the singular; the paradigms with [t] plurals do not alternate.

In general, children at both ages were not very good at producing a singular form at all, perhaps showing the difficulty of the task for a young child. However, when children did produce a singular form, they showed significantly better performance at both ages when the stem-final obstruent was voiceless—i.e. the nonce noun paradigm did not alternate—than when it was voiced.

Another study by Zamuner et al. (2006) corroborates these results. The structure of Zamuner et al. (2006)'s study was very similar to that of Zamuner et al. (2012), but it tested children's comprehension rather than production. Again, 2.5 year old and 3.5 year old children were tested. The children were presented the same stimuli as in the production study, but instead of being asked to produce the singular form, an experimenter read the singular form and asked the children to point to the correct image. There were three image choices: the plural image of the nonce, the singular image of the nonce (the correct choice), and a distractor (a plural image of a different nonce). For example, a child would be shown a plural picture and told that it represents multiple [slɑdən], and then asked to point to the picture of a [slɑt].

Children showed better performance at this comprehension task than production-based tasks. Moreover, children at both ages were much more accurate at identifying the correct singular image when the nonce-paradigm did not alternate than when it did. The authors also flipped the experiment, performing a standard singular-to-plural (comprehension) wug test. Again, the children were significantly better at identifying the correct plural for non-alternating paradigms.

These results are perhaps not surprising. Buckler and Fikkert (2016) found Dutch 3-year-olds to show little knowledge of which nouns in their mental lexicon alternate; lacking such robust knowledge, there would be little information for children to have used to construct a productive generalization for the voicing alternation.

Kerkhoff (2007, ch. 6) found that adult speakers of Dutch were able to produce singulars for nonce plurals from both alternating and non-alternating paradigms. This may constitute evidence that the voicing alternation eventually emerges as productive. However, such productivity may be aided by orthography: Kerkhoff (2007) observes that, e.g., the orthographic rendering of (16a) *bed* ~ *bedden* makes the status of the final [t] in [bɛt] as a neutralized /d/ transparent, and children continue to show little evidence of productively generalizing the alternation even at age 6.

In the remainder of this chapter, we will demonstrate how our model gives a possible learning-based explanation for these results. When run on a realistic child-directed speech corpus, the prevalence of voicing alternations in noun paradigms is rare enough that alternating forms can be lexicalized as the exceptions to productive morphological rules. Consequently, no phonological process is constructed.

Before proceeding, it is important to address one remaining fact about Dutch nouns. Some nouns are formed by suffixing [-s], not [-ən], to the stem; (19) gives examples.

(19) [voːɣəl] 'bird' [voːɣəls] 'birds'
[vɪŋər] 'finger' [vɪŋərs] 'fingers'

The [-s] plural is usually used after syllables that do not have main stress (Booij, 1999, p. 82). Because [ə] does not bear stress in Dutch (Booij, 1999, p. 5), a very strong indicator of the [-s] plural suffix is a stem ending in a syllable with a [ə] nucleus. The [-s] suffix is less frequent than the [-ən] suffix, does not often attach to a stem with a final obstruent, and does not lead to a stem-final consonant being re-syllabified. Thus, the [-s] suffix does not directly contribute to the Dutch voicing alternation, and nouns taking it are usually ignored in corpus and experimental studies. However, a complete story of Dutch nouns must account for these, and our own dataset (see below) retains them.

4.2 Setup and Data

We extracted nouns from the child-directed speech in the Van Kampen (Van Kampen, 1994, 2009) corpus of the CHILDES (MacWhinney, 2000) database. We used all sessions with children up to age 3.5 years, following Buckler and Fikkert (2016)'s corpus study, and the age of children in Zamuner et al. (2006, 2012)'s experiments. We used the TreeTagger[13] part-of-speech tagger (Schmid, 1999, 2013) to extract nouns from the corpus. The tagger also specifies whether the noun is singular or plural. We tagged diminutive nouns post-hoc based on the *-je* suffix (the TreeTagger provides lemmas for tagged words). To get IPA transcriptions, we took the intersection of the resulting set of nouns with the Dutch part of *wikipron* (Lee et al., 2020). Finally, we computed the frequency of each word in the corpus, and dropped words with only a single occurrence. This resulted in a set of 887 nouns—606 singular, 107 plural, 124 singular diminutive, and 50 plural diminutive.

[13]https://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

Meaning	UR	Abstract
PL	/-ən/	N
DIM	/-jə/	N
‘horse’	/part/	N
‘book’	/buk/	N

Table 6: The underlying forms of the plural and diminutive suffixes, as well as example roots ‘horse’ (which alternates) and ‘book’ (which does not alternate) after running on the model on the 887 Dutch nouns from the CHILDES dataset.

4.3 Generalization Without Abstraction

We ran our model on the 887 nouns in our dataset. The underlying forms of the plural (PL) and diminutive (DIM) affixes, as well as the root nouns ‘horse’ and ‘book’ are shown in Tab. 6. All four morphemes are concrete. Even though the noun ‘horse’ occurs in singular, plural, singular diminutive, and plural diminutive forms, the root occurs as [part], with voiceless [t], in all but the plural [pardən] (see. 17). Since $1 \leq 4/\ln 4$, the plural form is lexicalized. The root contains voiceless /t/ because 3 of 4 forms realize the obstruent as that. Similarly, the UR for ‘book’ is [buk] because this is the realization of the root in all four forms.

In short, then, surface alternation is rare enough in our sample of Dutch child-directed speech that concrete forms are sustainable for root morphemes, even if the final segment is an alternating obstruent. It is possible, however, that the number of alternating noun paradigms could grow large enough that it prohibits the formation of morphological (re-)inflection rules. For example, suppose a learner is trying to learn a generalization from SGs to PLs and posits the rule (20), stating that the plural form of a noun can be derived by suffixing [-ən] to the root form.

(20) PL = ROOT-[ən]

Alternating noun paradigms will appear to be exceptions to this generalization, since suffixing [ən] to a root form that ends in a voiceless obstruent will fail to yield the appropriate plural form in such cases (e.g., producing *[partən] instead of [pardən]). This is not a feature of *particular* morphemes, and hence does not lead our model to posit abstract underlying forms. It is, however, a possible second stage at which abstraction could become necessary.

To determine whether this is the case for Dutch, we ran the *Abduction of Tolerable Productivity* (ATP) model from Belth et al. (2021), to learn morphological

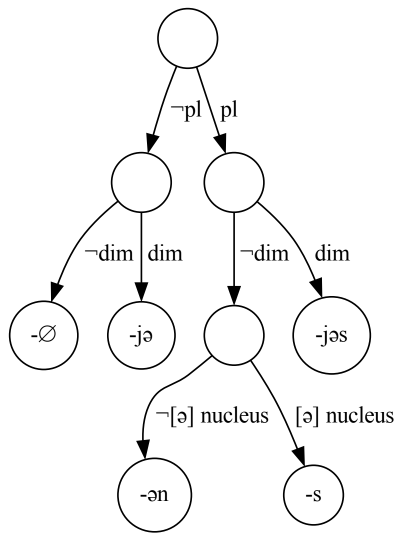


Figure 4: Morphological Inflection Rules

inflection rules from the underlying forms our model constructs. The model learns to map from root forms and features (morphosyntactic and/or phonological) to an inflected form, recursively subdividing root forms based on their features until a rule (branch in the implied decision tree) becomes productive under the Tolerance Principle, or until there are no more features to split on.

If the model fails to construct a productive rule for deriving the plural form from concrete roots because of the voicing alternation, this is evidence that abstraction is necessary to generalization. Future work will combine these into a single model.

ATP's output is shown as a decision tree in Fig. 4. The left-most leaf shows that (non-diminutive) singulars add no suffix (-$\emptyset$ denotes the null suffix) to the lemma. The [-jə] and [-jəs] leaves show that singular and plural diminutives are formed by suffixing [-jə] and [-jəs] to the lemma, respectively. The remaining two leaves show how plurals are formed: [-s] if the lemma ends in a syllable with a [ə] nucleus, and [-ən] otherwise.

In (21), we show the Tolerance Principle counts for each of ATP's rules. The rules are shown as an ordered list from most specific (deepest in tree) to least specific.

(21) a. PL $\wedge\neg$DIM $\wedge\neg$[ə] nucleus $\longrightarrow$ [-ən] (n = 90, e = 10 $\leq$ 90/ln 90 = 20.0)
b. PL $\wedge$ $\neg$DIM $\longrightarrow$ [-s] (n = 17, e = 2 $\leq$ 16/ln 16 = 6.0)
c. PL $\longrightarrow$ [-jəs] (n = 50, e = 0 $\leq$ 50/ln 50 = 12.8)
d. $\neg$DIM $\longrightarrow$ -$\emptyset$ (n = 603, e = 0 $\leq$ 603/ln 603 = 94.2)
e. Elsewhere $\longrightarrow$ [-jə] (n = 124, e = 6 $\leq$ 124/ln 124 = 25.7)

Since alternations only occur for paradigms where a stem-final obstruent is re-syllabified in the onset of the [-ən] PL suffix, the alternating nouns fall in the third-from-left branch of the tree, whose counts are shown in (21a). Since 10 $\leq$ 90/ln 90, the rule is productive despite the alternation and the alternating nouns can be lexicalized. Moreover, of the 10 exceptions, only 8 are due to the voicing alternation.[14] The parent node of the plural suffixes (21a)-(21b) had $n = 90 + 17 = 107$ nouns, of which 16 take the [-s] suffix and only 8 alternate. Thus, subdividing to isolate the 16 nouns that take [-s] instead of subdividing to isolate the 8 alternating nouns that take [-ən] is expected. Note that if the underlying forms of alternating noun stems contained voiced obstruents instead of voiceless obstruents, the exceptions would instead have gone to the left-most branch (21d) (since devoicing would be necessary to produce the correct singular form), but 8 $\leq$ 94.2, so the conclusion would hold even more strongly.

It appears, then, that Dutch-learning children may lack productive knowledge of voicing alternations because such knowledge is not important to their developing linguistic system. Our model demonstrates that highly productive morphological generalizations are possible without knowledge of voicing alternations, and perhaps child learners are content with these imperfect but fully sufficient generalizations.

To make this point more clear, consider our model's behavior on the nonce nouns from Zamuner et al. (2006, 2012). We consider the comprehension reverse wug test, in which children were presented with a nonce plural followed by its singular form and were then asked to identify the picture corresponding to the singular. Since our model did not need to learn a voicing alternation, when we present the plurals to our model, the predicted singular form is simply the plural with the [-ən] suffix removed. For example, our model predicts [slɑd], [kɛd], [klat], and [jɪt] as the singular for the alternating nonce words in (18), repeated in (22).

[14]The others are due to one stem that takes [-s] and one stem that has an irregular vowel change in the plural.

Table 7: Both humans and our model perform much better at identifying the singular of a nonce plural when the noun does not alternate than when it does. In comparison, a trigram language model does not reflect this asymmetry.

	Non-Alternating	Alternating
Humans	0.61	0.48
Our Model	1.00	0.33
Trigram	1.00	1.00

(22) a. [slɑdən]
[kɛdən]

b. [klatən]
[jɪtən]

The predictions will match the singular spoken by the experimenter, which always ends in [t], for non-alternating forms (22b) but not for the alternating forms (22a). When the model's prediction does not match the singular, we suppose that the image is chosen at random (from three choices). The resulting predictions are shown in Tab. 7, where our model predicts the expected form for non-alternating nouns but performs at-chance for alternating nouns. For reference, we report the average performance of the children in Zamuner et al. (2006)'s study.[15] Clearly the children performed much worse overall than the model, perhaps indicating the overall difficulty of the task for children. However, the trend is correct. Contrast this with a trigram language model fit to the surface forms of the 887 Dutch nouns. Because voiced obstruents never occur in syllable final position, the [t] singular form spoken by the experimenter always has higher probability than the same noun with a final voiced [d]. This highlights how knowledge of the phonotactic restriction against voiced obstruents in final position does not automatically lead learners to create a productive voicing alternation generalization.

[15]Zamuner et al. (2006)'s results were broken down by age, but no significant effect was found for age, so we aggregated the results into a single number for each type of noun by taking the weighted average of the two age groups' results.

4.4 Quantitative Evaluation

4.4.1 Setup

As for Turkish, we evaluated how the model enables morphophonological generalization. However, as discussed above (§ 4.3), the voicing alternation is not pervasive enough to require abstract underlying representations. This was established by running our model on the entire dataset of 887 nouns. It is important to confirm that the these observations are true throughout (simulated) development of the mental lexicon. Secondly, it is important to know that when the calculation of sufficiently accurate generalization from concrete representations (via the Tolerance Principle) does not require abstraction, that morphophonological generalizations constructed over concrete representations still achieve high accuracy generalizing to new words.

Thus, we ran our model on the Dutch dataset, simulating incremental learning. As for Turkish, each simulation sampled words with replacement and weighted by frequency, and added them to the lexicon when sampled. In 100-word increments up to a vocabulary size of 700 nouns, we confirmed that all underlying forms were still concrete and that productive morphological inflection rules could still be extracted via ATP. At each of these increments, we also evaluated the accuracy of the resulting morphological inflection rules over 187 test words not seen during training. We performed 30 simulations using different random seeds.

We compared to a transformer-based (Vaswani et al., 2017) seq-to-seq model. At each training size, we created a random 80/20 train/dev split of the training data and tuned hyperparameters—the same as Turkish (15)—by choosing the combination with the highest accuracy on the dev set, and then training a model with the best hyperparameters on the entire training set (i.e. re-merging the 80/20 split).

4.4.2 Results

Across the 30 simulations, the voicing alternation was never pervasive enough to require the model to construct abstract underlying forms nor to prevent ATP from learning productive morphological inflection rules. The accuracy generalizing to held-out test words, shown in Fig. 5, approaches 0.95 by the time the vocabulary contains 700 nouns. The performance surpasses that of the transformer seq-to-seq model at all training sizes.

Despite not creating abstract underlying forms for alternating paradigms—and hence not learning a productive voicing alternation—the high overall accuracy demonstrates that accurate generalizations are nevertheless possible.

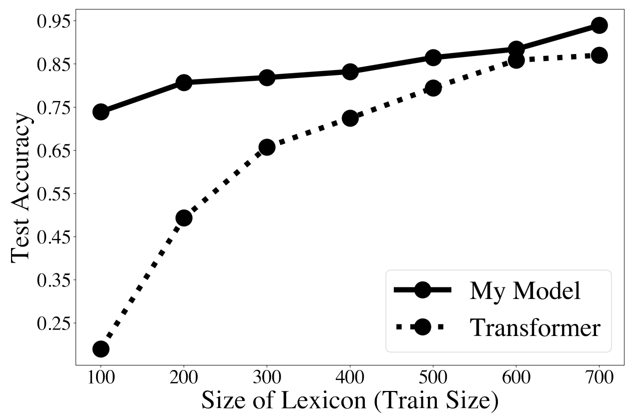


Figure 5: Accuracy generalizing to held-out test words.

5 Prior Work

Tesar (2013) and Hua et al. (2020) focus on theoretical analyses of the nature of the problem of learning URs. O'Hara (2017); Rasin et al. (2018); Ellis et al. (2022) proposed computational models, but evaluate on small, phonology-textbook-like data, not large, natural-language corpora.

Cotterell et al. (2015) also predominately models textbook-like problems, but presents some limited analysis on more realistic corpora. However, these corpora only involve very simple morphological paradigms involving a single suffix, and present to the model a fairly curated subset of the corpus that isolates the relevant morphophonological process.

Richter (2021) studies the question of when allophonic surface segments are collapsed into an abstract underlying segment, focusing on the English flap [ɾ] allophone of /T/. While Richter (2021) focuses on allophones, our proposed model is inspired by it and can be viewed as extending the same principles to morphophonological alternations.

Of these prior models, we were only able to get access to code for Rasin et al. (2018), which we were unable to get to run on our large datasets. In future versions of this work, we intend to implement some of these existing models in order to compare their performance and behavior to that of our proposed model.

Table 8: The logic of model's functioning, and how high generalization accuracy is achievable both when abstraction is needed and when it is not.

Language	Concrete Generalization?	Abstract URs	High Accuracy
Turkish	✗	✓	✓
Dutch	✓	✗	✓

6 Conclusion

This paper begins developing an algorithmic, learning-based account of underlying forms, taking the highly agglutinating language of Turkish and an apparent lack of productivity in Dutch as two case studies. The proposed model starts with concrete underlying representations and constructs abstract URs only in cases where doing so is needed to form generalizations that deal with the sparsity of morphological forms in the learner's input.

The model constructs abstract underlying forms only when they are critical for generalization, and thus retains concrete forms when abstraction is unnecessary. This flexibility is at the core of the model's success, as evidenced by the constructed underlying representations of Turkish suffixes in § 3.3 and alternating Dutch noun roots in § 4.3. For example, the half-harmonizing suffixes consist of concrete segments except for the single, harmonizing vowel. Similarly, exceptional, non-harmonizing suffixes remain fully concrete.

When combined with a model for learning local and non-local alternations, the proposed model achieves at least 95% accuracy predicting the surface form of held-out test words.

In the case of Turkish, abstract URs enabled the learning of accurate morphophonological generalizations because of the pervasive amount of surface alternation. On the other hand, the lack of surface alternation in Dutch did not drive the learner to construct abstract underlying forms, yet accurate, productive morphophonological generalizations were still able to be extracted. The situation is summarized in Tab. 8. When generalization from concrete URs is not possible (Turkish), our model constructs abstract URs, and high generalization accuracy is achieved from these. On the other hand, when generalization from concrete URs is possible (Dutch), our model need not construct abstract URs, but high generalization accuracy is still achieved, directly from the concrete representations.

6.1 Limitations and Future Directions

The results in this paper are promising, but more work is needed on the problem of underlying forms, and future work will need to evaluate the model on other languages.

We identified two stages at which underlying abstraction can become necessary. First, the number of forms that a morpheme takes across word forms can become large enough that underlying abstraction is necessary to predict its surface form across these words. Second, alternations across a set of morphemes can accumulate so as to prevent the formulation of productive morphological inflection rules. The first was prevalent enough in the heavily-agglutinating language Turkish to motivate underlying abstraction, but neither force necessitated abstraction in the case of Dutch nouns. An important direction for future work is to figure out how these two possible reasons for abstraction relate, including whether they can be combined.

Furthermore, we have not yet considered the question of *degrees* of abstractness. For instance, how can we provide a learning-based decision for whether to treat a Turkish alternating affix vowel as underspecified /A/ derived into [ɑ] and [e] via a harmony rule, versus /ɑ/ derived into [e] or /e/ into [ɑ]?

Lastly, we have assumed to this point that morphological segmentations are available to the learner. As discussed in § 2.1, there is experimental evidence that children are able to perform morphological segmentation. However, the task of constructing URs is intertwined with the segmentation problem, as one cannot hope to determine the UR of a morpheme without at least knowing which segments belong to its own surface realizations and not those of another morpheme. Thus, in future work we will pursue a learning-based account of segmentation and attempt to bring the problems together, jointly segmenting surface forms, learning underlying forms, and morphophonological grammars.

References

Altan, A. (2009). Acquisition of vowel harmony in Turkish. *Dilbilim 35. Yıl Yazıları*, pages 9–26.

Anderson, S. R. (2021). *Phonology in the Twentieth Century*. Number 5 in History and Philosophy of the Language Sciences. Language Science Press, Berlin.

Belth, C. (2023a). A learning-based account of local phonological processes. *Phonology*. In Press.

Belth, C. (2023b). A learning-based account of phonological tiers. Accepted.

Belth, C., Payne, S., Beser, D., Kodner, J., and Yang, C. (2021). The greedy and recursive search for morphological productivity. In *CogSci*.

Berko, J. (1958). The child's learning of english morphology. *Word*, 14(2-3):150–177.

Best, C. T., McRoberts, G. W., and Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by english-speaking adults and infants. *Journal of experimental psychology: human perception and performance*, 14(3):345.

Boersma, P. (1997). How we learn variation, optionality, and probability. In *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, volume 21, pages 43–58. Citeseer.

Booij, G. (1999). The phonology of dutch.

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., Pascual, L., Pêcheux, M.-G., Ruel, J., Venuti, P., and Vyt, A. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, dutch, french, hebrew, italian, korean, and american english. *Child development*, 75(4):1115–1139.

Brown, R. (1973). *A first language: The early stages*. Harvard University Press, Cambridge, MA.

Buckler, H. and Fikkert, P. (2016). Dutch and german 3-year-olds' representations of voicing alternations. *Language and Speech*, 59(2):236–265.

Chan, E. (2008). *Structures and distributions in morphology learning*. PhD thesis, University of Pennsylvania.

Coetzee, A. W. (2009). Learning lexical indexation. *Phonology*, 26(1):109–145.

Çöltekin, Ç. (2010). A freely available morphological analyzer for turkish. In *LREC*, volume 2, pages 19–28.

Çöltekin, Ç. (2014). A set of open source tools for turkish natural language processing. In *LREC*, pages 1079–1086.

Cotterell, R., Peng, N., and Eisner, J. (2015). Modeling word forms using latent underlying morphs and phonology. *Transactions of the Association for Computational Linguistics*, 3:433–447.

Demuth, K., Culbertson, J., and Alter, J. (2006). Word-minimality, epenthesis and coda licensing in the early acquisition of english. *Language and Speech*, 49(2):137–173.

Dobrovolsky, M. (1982). Some thoughts on turkish voicing assimilation. In *Calgary Working Papers in Linguistics*, volume 7, pages 1–6.

Ellis, K., Albright, A., Solar-Lezama, A., Tenenbaum, J. B., and O'Donnell, T. J. (2022). Synthesizing theories of human language with bayesian program induction. *Nature communications*, 13(1):5024.

Emond, E. and Shi, R. (2021). Infants' rule generalization is governed by the Tolerance Principle. In Dionne, D. and Vidal Covas, L.-A., editors, *Proceedings of the 45nd annual Boston University Conference on Language Development*, pages 191–204.

Ettlinger, M. (2008). *Input-driven opacity*. University of California, Berkeley.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., and Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, pages i–185.

Hart, B. and Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children.* Paul H Brookes Publishing, Baltimore, MD.

Hayes, B. (2004). Phonological acquisition in optimality theory: the early stages. *Constraints in phonological acquisition*, pages 158–203.

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

Herdan, G. (1960). *Type-token mathematics*, volume 4. Mouton.

Hua, W., Jardine, A., and Dai, H. (2020). Learning underlying representations and input-strictly-local functions. In *Proceedings of the 37th West Coast Conference on Formal Linguistics*.

Hyman, L. M. (2018). Why underlying representations? *Journal of Linguistics*, 54(3):591–610.

Jarosz, G. (2019). Computational modeling of phonological learning. *Annual Review of Linguistics*.

Kabak, B. (2011). Turkish vowel harmony. *The Blackwell companion to phonology*, pages 1–24.

Kerkhoff, A. O. (2007). *Acquisition of morpho-phonology: The Dutch voicing alternation*. PhD thesis, LOT.

Kiparsky, P. (1968). *How abstract is phonology?* Indiana University Linguistics Club.

Kornfilt, J. (2013). *Turkish*. Routledge.

Koulaguina, E. and Shi, R. (2019). Rule generalization from inconsistent input in early infancy. *Language Acquisition*, 26(4):416–435.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255(5044):606–608.

Kurimo, M., Virpioja, S., Turunen, V., and Lagus, K. (2010). Morpho challenge 2005-2010: Evaluations and results. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 87–95. ACL.

Ladányi, E., Kovács, Á. M., and Gervain, J. (2020). How 15-month-old infants process morphologically complex forms in an agglutinative language? *Infancy*, 25(2):190–204.

Lee, J. L., Ashby, L. F., Garza, M. E., Lee-Sikka, Y., Miller, S., Wong, A., McCarthy, A. D., and Gorman, K. (2020). Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France. European Language Resources Association.

MacWhinney, B. (1978). *The acquisition of morphophonology*. Monographs of the Society for Research in Child Development. University of Chicago Press.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. transcription format and programs*, volume 1. Psychology Press.

Marquis, A. and Shi, R. (2012). Initial morphological learning in preverbal infants. *Cognition*, 122(1):61–66.

Mintz, T. H. (2013). The segmentation of sub-lexical morphemes in english-learning 15-month-olds. *Frontiers in psychology*, 4:24.

Narayan, C. R., Werker, J. F., and Beddor, P. S. (2010). The interaction between acoustic salience and language experience in developmental speech perception: Evidence from nasal place discrimination. *Developmental science*, 13(3):407–420.

Nevins, A. (2010). *Locality in vowel harmony*, volume 55. Mit Press.

Oflazer, K. (1994). Two-level description of turkish morphology. *Literary and linguistic computing*, 9(2):137–148.

O'Hara, C. (2017). How abstract is more abstract? learning abstract underlying representations. *Phonology*, 34(2):325–345.

Payne, S. (2022). *When collisions are a good thing: the acquisition of morphological marking*. Bachelor's thesis, University of Pennsylvania.

Rasin, E., Berger, I., Lan, N., and Katzir, R. (2018). Learning phonological optionality and opacity from distributional evidence. In *Proceedings of NELS*, volume 48, pages 269–282.

Richter, C. (2018). Learning allophones: What input is necessary. In *Proceedings of the 42nd annual Boston University Conference on Language Development.* Cascadilla Press.

Richter, C. (2021). *Alternation-Sensitive Phoneme Learning: Implications For Children's Development And Language Change*. PhD thesis, University of Pennsylvania.

Ringe, D. and Eska, J. F. (2013). *Historical linguistics: Toward a twenty-first century reintegration*. Cambridge University Press.

Schmid, H. (1999). Improvements in part-of-speech tagging with an application to german. *Natural language processing using very large corpora*, pages 13–25.

Schmid, H. (2013). Probabilistic part-ofispeech tagging using decision trees. In *New methods in language processing*, page 154.

Schuler, K. D., Yang, C., and Newport, E. L. (2016). Testing the tolerance principle: Children form productive rules when it is more computationally efficient to do so. In *CogSci*, volume 38, pages 2321–2326.

Slobin, D. I. (1982). Universal and particular in the acquisition of language. *Language acquisition: The state of the art*, 57.

Szagun, G., Steinbrink, C., Franik, M., and Stumper, B. (2006). Development of vocabulary and grammar in young german-speaking children assessed with a german language development inventory. *First Language*, 26(3):259–280.

Tesar, B. (2013). *Output-Driven Phonology: Theory and Learning*. Cambridge University Press, Cambridge.

Tesar, B. and Smolensky, P. (1998). Learnability in optimality theory. *Linguistic Inquiry*, 29(2):229–268.

Van Kampen, J. (2009). The non-biological evolution of grammar: Wh-question formation in germanic. *Biolinguistics*, 3(2-3):154–185.

Van Kampen, N. J. (1994). The learnability of the left branch condition. *Linguistics in the Netherlands*, 11:83–94.

Van Tuijl, R. and Coopmans, P. (2021). The productivity of dutch diminutives. *Linguistics in the Netherlands*, 38(1):128–143.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Werker, J. F. and Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development*, 7(1):49–63.

Yang, C. (2016). *The price of linguistic productivity: How children learn to break the rules of language*. MIT press.

Zamuner, T. S., Kerkhoff, A., and Fikkert, J. (2006). Acquisition of voicing neutralization and alternations in dutch. In *Proceedings of the 30th annual Boston University Conference on Language Development*, pages 701–712.

Zamuner, T. S., Kerkhoff, A., and Fikkert, P. (2012). Phonotactics and morphophonology in early child language: Evidence from dutch. *Applied psycholinguistics*, 33(3):481–499.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley, Cambridge, MA.