When are Alternations Learned?

Caleb Belth

Because the acquisition of phonotactics begins before the acquisition of morphology, it is possible that phonotactic knowledge underlies the acquisition of alternations, at least when the two align. An example is final devoicing. The typical analysis is that alternating obstruents are underlyingly voiced but devoiced in final position, avoiding a phonotactic violation. If phonotactic knowledge underlies alternation learning, acquisition of alternations should follow rapidly after the acquisition of the morphology that produces violations. Yet different languages with voicing alternations show developmental differences: Dutch-learning children's developing knowledge is more restricted than that of their German-learning peers. I hypothesize that children begin attending to alternations, and form non-surface-faithful representations, not as a reflex of phonotactic knowledge, but in response to being unable to morphologically generalize without doing so. Implemented as an incremental learning model, the proposal explains observed developmental trajectories, with implications for our understanding of alternations and non-surface-faithful representations.

1 Introduction

The place of alternations, phonotactics, and abstract representations has varied across phonological theories. In SPE (Chomsky & Halle, 1968), phonological processes were central and phonotactics were interpretable as little more than the surface structures not produced. SPE's preference for minimal numbers of symbols encouraged compressing any regularity in the lexicon into increasingly abstract representations and corresponding generalizations. That path, however, led to arguably excessively abstract representations. Subsequent critiques, perhaps most famously by Kiparsky (1968), urged better justification for abstract representations and emphasized the importance of productive alternations in that justification. However, the lack of a general theory of productivity and a shift towards constraint-based theories (Prince & Smolensky, 1993) accompanied a near inversion of focus to restrictions on surface structures, with alternations being viewed as the (nearly) automatic consequence of morphological processes creating phonotactically illicit structures. For instance, Hayes & Wilson (2008:p. 424) propose that when putting morphemes together violates phonotactic restrictions that are already learned, the need to learn the corresponding alternation becomes "immediately apparent" to the learner. This view, marshaling the observation that phonotactics appear to be acquired earlier in development than much of morphology (Jusczyk *et al.*, 1994;

Sundara *et al.*, 2022), is now often adopted (Hayes, 2004; Pater & Tessier, 2005; Hayes & Wilson, 2008; Chong, 2021; Kuo, 2024; Wang & Hayes, 2025).

Consider a cross-linguistically prevalent example: final devoicing. Final devoicing produces alternations when a vowel-initial suffix leads to an obstruent being syllabified in both final and onset positions across a paradigm, like examples from Dutch (1a)-(1b) (Booij, 1999) and German (1c)-(1d) (Wiese, 1996) show. The vowel-initial Pl suffixes result in a voicing alternation in some paradigms (1a)-(1c) but not others (1b)-(1d).

(1) a. [bɛt] 'bed' [bɛdən] 'beds'   c. [moːnt] 'moon' [moːndə] 'moons'
    b. [pɛt] 'cap' [pɛtən] 'caps'   d. [bʁoːt] 'bread' [bʁoːtə] 'breads'

A typical analysis involves positing non-surface-faithful URs for the alternating stems—for instance, /bɛd/ as the stem UR for (1a). The final [t] in the Sg must then be interpreted as something other than /t/—it is not surface faithful.

This alternation avoids a violation of language-specific phonotactics, as both languages lack final voiced obstruents, and of general markedness. By age 3, well after a sensitivity to language-specific phonotactics emerges (Sundara *et al.*, 2022), Dutch- and German-learning children have acquired the basic morphology that triggers alternations (Mills, 1986; Elsen, 2002; Kerkhoff, 2007). Thus, if acquisition of phonotactics is carried over to alternation learning, we ought to expect Dutch- and German-learning children to show similar knowledge of the alternation.

Yet the developmental facts vary sharply between Dutch and German. Because prior acquisition research on voicing alternations has focused on noun paradigms, our discussion centers on nouns. Dutch, where development has been studied in the greatest detail (see § 4), shows a slow developmental trajectory, with limited productivity of the voicing alternation well into childhood—at least age 6-7, the oldest age of children in studies I am aware of—(Zamuner *et al.*, 2006; Kerkhoff, 2007; Zamuner *et al.*, 2012). There is little evidence even of productive devoicing through at least 4;8. The voicing alternation appears to have highly-limited productivity even for adults, who extend it to wugs only about 2% of the time (Kerkhoff, 2007).

In contrast, German-learning children are extending voicing alternations to novel nouns productively by age 5 (the earliest age in studies I am aware of: van de Vijver & Baer-Henney 2014), and in an age-controlled experiment with 3-year-old Dutch- and German-learning children, German-learning children appeared more advanced in their knowledge of stem-final obstruent voicing than Dutch-learning children (Buckler & Fikkert, 2016). Buckler & Fikkert also note that, despite formal similarity, the distributional properties of the voicing alternation in corpus data are markedly different between Dutch and German, with greater evidence of the alternation being available in German.

In this article, I propose a theory of alternations in which representations are constructed and alternations learned in order to support productive generalizations in the face of a sparse linguistic input. Specifically, I hypothesize that learners attend to learning voicing alternations—and set up non-surface-faithful representations—only when they fail to construct sufficient morphological inflection rules without doing so. This hypothesis is formalized in an implemented computational learning model for Dutch and German noun morphophonology. The model extends Belth *et al.* (2021)'s recursive learning model, ATP, to the case of voicing alternations. The core of the model is a recursive application of the Tolerance Principle (Yang, 2016), which is used to propose and verify rules, thereby serving as a threshold for when learners will begin to attend to, and learn, the alternation.

I apply the model to Dutch and German child-directed speech, and compare its generalization behavior to a range of results from prior corpus and experimental studies with Dutch- and German-learning children. The model accounts for the developmental differences between Dutch and German, closely matching voicing error rates, mispronunciation recognition, wug test, and reverse wug test results.

I begin with a discussion of the distributional facts of Dutch and German noun voicing alternations (§ 2), followed by a description of the general problem of morphophonological learning and a presentation of my proposed learning model (§ 3). I then present the developmental facts of Dutch and German, comparing the model and alternative proposals to them at each turn (§ 4). I conclude with a discussion (§ 5).

2 Distributional Facts of Dutch and German Noun Voicing Alternations

Dutch (§ 2.1) and German (§ 2.2) both have phonotactic constraints against final voiced obstruents and both have vowel-initial nominal suffixes that lead stem-final consonants to be syllabified in an onset in some inflected nouns. Whether a noun is alternating or non-alternating is lexically specific (typically treated as a difference in underlying voicing). The distributional patterning of the alternation is markedly different between the languages.

2.1 Dutch

In Dutch, the vowel-initial suffixes are the plural allomorph [-ən] and the diminutive allomorph [-ətjə], though the latter is rare and is usually the last of five Dim allomorphs to be acquired (Van Tuijl & Coopmans, 2021). Both trigger alternations in some obstruent-final paradigms but not others (2).

(2)	a.	[bɛt]	'bed'	[bɛdən]	'bed'-PL
	b.	[pɛt]	'cap'	[pɛtən]	'cap'-PL
	c.	[slɑp]	'slab'	[slɑbətjə]	'slab'-DIM
	d.	[pɔp]	'doll'	[pɔpətjə]	'doll'-DIM

The Dutch stops that contrast in voicing in initial and/or medial position are /t/ ~ /d/ and /p/ ~ /b/. There is a contrast between fricatives /f/ ~ /v/, /s/ ~ /z/, and /x/ ~ /ɣ/, but it is weakening for many speakers (Slis & van Heugten, 1989; Booij, 1999).

Some plurals are marked with [-s], though the [-ən] allomorph is more frequent. The [-s] suffix does not create the environment for an alternation (3).

(3)	[voːɣəl]	'bird'	[voːɣəls]	'bird'-PL
	[vɪŋər]	'finger'	[vɪŋərs]	'finger'-PL

The choice between [-ən] and [-s] is largely predictable from prosodic features: in particular, nouns with unstressed final syllables usually take [-s] (Booij, 1999:p. 82).

Dutch has four additional DIM allomorphs—[-jə], [-tjə], [-pjə], [-kjə]—none of which are vowel initial, and the PL DIM is categorically formed by adding [-s] to the DIM form. Moreover, diminutives are very common in Dutch and particularly in child-directed speech, often occurring at higher token frequencies than their non-DIM counterparts (Kerkhoff, 2007). For example, Kerkhoff:p. 113, reports that, in a corpus of child-directed speech, *eendjes* "ducklings" occurs more times than *eenden* "ducks." Thus, even for an alternating paradigm, a child who knows the DIM PL of a noun but not the PL, has no evidence of that alternation. In a paradigm like (4), only one of four forms (4b) provides evidence for the alternation.

(4)	a. [paːrt]	'horse'		c.	[paːrtjə]	'horse'-DIM
	b. [paːrdən]	'horse'-PL		d.	[paːrtjəs]	'horse'-DIM-PL

In our dataset (see § 4.1.1), the DIM PL is more frequent than the non-DIM PL in 42.3% of paradigms. Together these facts indicate that the voicing alternation, at least descriptively, occurs in a rather small corner of the Dutch noun system.

2.2 German

German has a complex noun system, involving five plural suffixes and case markings for some nouns (Wiese, 1996). The plural suffixes are predictable from complex combinations of features of the noun, including its grammatical gender and phonological ending. Several studies have investigated the productivity and acquisition of these suffixes (Clahsen *et al.*, 1992; Marcus *et al.*, 1995; Elsen, 2002; Zaretsky & Lange, 2015; Yang, 2016; McCurdy *et al.*, 2020; Belth *et al.*, 2021), but do not typically discuss the voicing alternation.

The plural suffixes are -(e)n [-ən]/[-n], -e [-ə], -∅, -er [-ɐ], -s [-s]. The vowel-initial [-ən], [-ə], and [-ɐ] all create possible environments for voicing alternations (5).

(5)	[geːgənt]	'area'	[geːgəndən]	'area'-Pl
	[mɔnt]	'moon'	[mɔndə]	'moon'-Pl
	[bɪlt]	'picture'	[bɪldɐ]	'picture'-Pl

In comparison to Dutch, where only labial and coronal stops contrast voicing and the contrast is weakening for fricatives, German stops /t/ ~ /d/, /p/ ~ /b/, and /k/ ~ /g/ all contrast voicing, as well as the fricatives /f/ ~ /v/ and /s/ ~ /z/ (Wiese, 1996:sec. 2.1).

Some Sg nouns also mark case, using various suffixes (Pounder, 1996). Gen case is sometimes marked, mostly for neuter and masculine nouns, with [-s], [-əs], [-n], or [-ən]; [-əs] and [-ən] can both create environments for voicing alternations (6a). Some nouns (mostly masculine) mark Acc case with [-ən] or [-n], with the former creating environments for voicing alternations (6b). Similarly, some nouns mark Dat case with [-ə], [-ən], or [-n], also creating some environments for alternations (6c).

(6)	a.	[t͡suːk]	'train'	[t͡suːgəs]	'train'-Gen
		[bɪlt]	'picture'	[bɪldən]	'picture'-Gen
	b.	[hɛlt]	'hero'	[hɛldən]	'hero'-Acc
	c.	[t͡suːk]	'train'	[t͡suːgə]	'train'-Dat
		[hɛlt]	'hero'	[hɛldən]	'hero'-Dat

Any Pl noun that does not already end in [n] or [s] marks Dat with [-n]. This does not directly introduce an alternation, but in alternating paradigms it adds an additional form with a voiced final obstruent, as (7) shows for the last two nouns in (5).

(7)	[mɔnt]	'moon'	[mɔndə]	'moon'-Pl	[mɔndən]	'moon-Pl-Dat'
	[bɪlt]	'picture'	[bɪldɐ]	'picture'-Pl	[bɪldɐn]	'picture'-Pl-Dat

Thus, in comparison to Dutch, German has ample environments for the alternation.

3 Learning Model

I begin by describing some important aspects of the empirical conditions of morphophonological acquisition (§ 3.1) and how these motivate the main idea of the proposal (§ 3.2), before turning to a description of the model (§ 3.3).

3.1 Empirical Conditions of Morphophonological Acquisition

The acquisition of morphophonology (and language more broadly) is characterized by sparsity. It is well known that the frequency distributions of many linguistic structures—including words and lemmas—have long tails. Plotting rank frequency against token frequency in log scale roughly forms a downward-sloping line, like in Figure 1: the Zipfian distribution (Zipf, 1949). Figure (1a) was generated from the child-directed words in the Van Kampen (Van Kampen, 2009) and Groningen (Bol, 1995) CHILDES Dutch corpuses (MacWhinney, 2014), and Figure (1b) from the child-directed words in the German Leo corpus (Behrens, 2006). Since the plots approximate downward-sloping lines, token frequency is approximately inversely proportional to rank frequency. The density of the plot increases as it moves downward (movement is exponential since the scale is logarithmic). The upshot is that a small number of words dominate the upper-end of the frequency distribution, while most are low frequency.

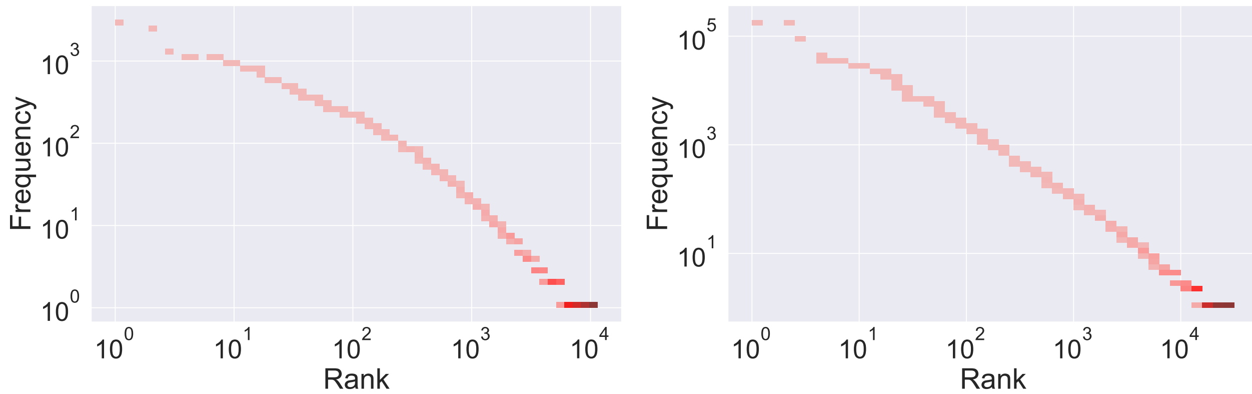
Because this distribution characterizes a child's input, they will only have heard a small subset of a language's vocabulary enough times to have the possibility of learning their form. In order to use the vast number of word forms never (or rarely) encountered in the input, the child must generalize from the words they know: morphophonological generalization is a necessary response to sparsity.

Moreover, paradigms are characteristically incomplete. Chan (2008:ch. 4) introduced *paradigm saturation*, which measures, for a given lemma, the fraction of forms of that lemma that are attested in a corpus, out of the total number of possible forms for the lemma. Chan analyzed verb paradigms in 16 corpuses of different languages, and reported that only English had even one verb with 100% paradigm saturation. Figure 2 plots saturation values for Dutch and German noun paradigms in the child-directed speech used for the present study (see § 4.1.1 for dataset details). I computed saturation values by sampling nouns until the sample included forms from 1000 unique paradigms. Each saturation value is a distribution over 30 samples. Dutch nouns can appear in four forms (Sg, Pl, Dim, Dim Pl) while German nouns can appear as Sg or Pl in any of four cases (Nom, Acc, Dat, Gen; see § 2). Case is only marked for some German nouns, so I include only forms that differ from the Nom in the denominator of the saturation value. Most noun paradigms are incomplete.

3.2 The Main Idea

These results suggest that sparsity drives learners to generalize from the words that they know to construct a productive grammar. However, it is not in general possible to form perfect generalizations; learners may be content with generalizations that cover a sufficient amount of the regularity and tolerate the remaining input as irregularity.

As a precise criterion of sufficient generalization, Yang (2016) introduced the Tolerance/Sufficiency



(a) Dutch **(b)** German

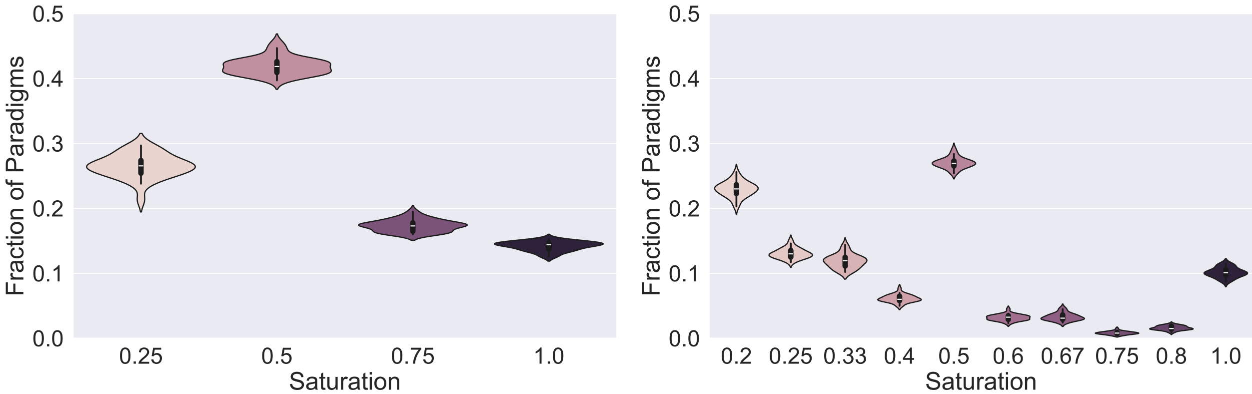
Figure 1: Frequency vs. Rank plots (log-log scale) of words in child-directed speech for Dutch (a) and German (b). The distributions are Zipfian: most words are low frequency.

Principle (TSP). The TSP evaluates a generalization/rule, R, based on the number of items in its scope (n) and the number of those that follow the rule (m), or the number that violate the rule ($e = n - m$). The threshold, given in (8), is a proposed psychological tipping point for how many exceptions a rule can *tolerate*—alternatively, what a *sufficient* number of rule-followers is—for a learner to accept the rule and apply it productively to new items. The threshold $n/\ln n$ is denoted θ_n.

(8) **Tolerance/Sufficiency Principle**: a rule R is productive *iff* $e = n - m \leq \dfrac{n}{\ln n}$

The TSP threshold was derived (Yang, 2016:ch.3) from considerations of the Zipfian character of linguistic data and its psychological impacts on processing. Experimental evaluation of the TSP has supported it as a categorical generalization tipping point for human learners (Schuler *et al.*, 2016; Koulaguina & Shi, 2019; Shi & Emond, 2023), in particular children, and it has been extensively used in corpus and computational studies of acquisition (Yang 2016; Kodner 2020; Belth *et al.* 2021; Björnsdóttir 2021; Pearl & Sprouse 2021; Richter 2021; Van Tuijl & Coopmans 2021; Belth 2023b; Henke 2023, among others). The TSP evaluates a rule, but Yang (2016:ch.3) described a recursive approach to both proposing and evaluating rules. Belth *et al.* (2021) developed a computational implementation of such a recursive approach, which is the basis of my model (§ 3.3).

Voicing alternations in noun paradigms form exceptions to morphological generalizations that do not account for them. For example, if a Dutch-learning child forms the generalization that plurals are formed by adding [-ən], then an alternating noun like [bɛt] ~ [bɛdən] is an exception because of the [t] ~ [d] alternation. However, because lexical exceptions are inevitable, some degree of alternating nouns may be tolerated without causing the rule to be refined to incorporate the alternation. If, at some point during acquisition, enough nouns alternate that they can no longer



(a) Dutch **(b)** German

Figure 2: Saturation values for Dutch (a) and German (b) noun paradigms plotted against the fraction of paradigms with that saturation, after sampling words until 1000 unique lemmas were represented. The plot is generated from 30 samples, yielding a distribution for each saturation value. Most paradigms are incomplete (saturation < 1.0).

be tolerated as exceptions, then this triggers the learner to refine the generalization. This is the core of my proposal: learning the voicing alternation is triggered when the relative number of alternating nouns rises to the level that productive morphological rules (as measured by the TSP) can no longer be constructed without them capturing the alternation. I turn now to making this proposal concrete, in the form of a learning algorithm.

3.3 The Learning Model

The proposed model, AlterTP (I will publish code along with the article), learns incrementally to do morphophonological (re)inflection from a growing lexicon. From the lexicon, it attempts to learn a grammar to predict the phonological forms of empty paradigm slots. Following the morphological (re)inflection literature (e.g., McCurdy *et al.* 2020; Belth *et al.* 2021; Kodner & Khalifa 2022; Goldman *et al.* 2023), the learning data is represented as (src, feats, tgt) triples, where src is a phonological form (a lemma or inflected form), feats is a set of features of the src (morphosyntactic, prosodic, etc.), and tgt is a phonological form inflected from the src. For instance, Dutch input may include triples like (9), where "F-stress" means the src has final stress.

(9) ([paːrt], {Dim, F-stress}, [paːrtjə])
([paːrt], {Pl, F-stress}, [paːrdən])
([paːrtjə], {Dim, Pl, F-stress}, [paːrtjəs])

The feats are presumed to be those available from earlier stages of development: in particular, Dutch-learning children know the main features of stress before age 3 (Fikkert, 1994) and by the same age German children are robustly sensitive to case, but are still figuring out how it is marked

(Tracy, 1986; Eisenbeiss *et al.*, 2006; Dittmar *et al.*, 2008). AlterTP interprets any concatenative difference between the src and tgt as a suffix.[1] I will typically refer to the tgt as only the suffix: for instance, ([paːrt], {Dim, F-stress}, [-jə]) for the first triple in (9). If the tgt is not equal to the src plus the inferred suffix, then the tgt is labeled "+alter." This only means that AlterTP knows that, in addition to taking a suffix, the stem undergoes *some* change (it alternates), though the nature of that change is yet to be learned (see § 3.3.4).[2] For instance [paːrdən] adds [-ən] to [paːrt] but does not equal [paːrt] + [-ən], so it is represented ([paːrt], {Pl, F-stress}, [-ən] +alter). Moving forward, I refer to the tgt as represented within AlterTP (e.g., "[-ən]" or "[-ən] +alter") rather than the full phonological form.

AlterTP is based on the model ATP, developed by Belth *et al.* (2021) for recursively applying the TSP to construct morphological inflection rules. We can think of a linguistic rule as a structural description, which is a conjunction of properties that an item must match in order for the rule to apply, together with a property that is predicted to be true of any items matching the structural description. For instance, (10) predicts that any triple that has the feats 'Pl' but not 'Dim' is inflected with [-ən].

(10) $\textsc{Pl} \wedge \neg\textsc{Dim} \implies$ [-ən]

By selecting the set of items that match its structural description, a rule tacitly also defines the complement of that set. In other words, the structural description *partitions* the set of items into the set of those that match the rule and the set of those that do not.

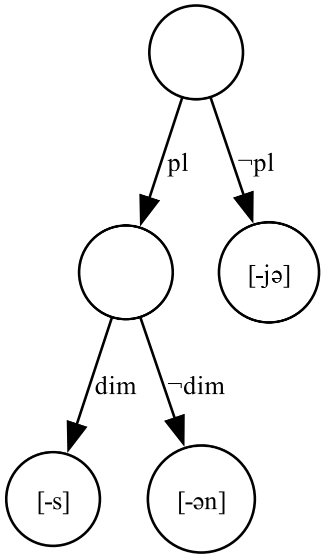
The TSP evaluates a rule of this form, where n is the number of items matching the rule's structural description and $m = n - e$ is the number of matching items about which the prediction is true. If n items match the structural description of a rule but $e > \theta_n$ of those items do not match the prediction, the rule can be refined by adding an additional condition to the structural description in an attempt to filter out the exceptions. This process of evaluating a rule under the TSP and, if necessary, adding a condition to its structural description can be carried out recursively. Consequently, we can represent a collection of rules as a binary[3] branching *decision tree*, like that in (11), where each node represents a recursive evaluation of and, if not a leaf, a refinement of a rule.

[1] While sufficient for present purposes, extensions would be necessary to consider the left edge (prefixation) or to apply the approach to non-concatenative morphology.

[2] Suffix alternations are directly captured. The Dutch and German suffixes all involve allomorphy (§ 2); AlterTP learns the conditions in which each applies. Belth *et al.* (2021) provide further cases.

[3] Building one feature at a time is motivated by experimental results that adults and children sort items into categories by attending to one dimension (feature) at a time (Smith, 1979; Nelson, 1984; Medin *et al.*, 1987).

(11) Example Inflection Tree



This grammar predicts that PL DIM nouns take [-s], while PL non-DIM nouns take [-ən] and non-PL (DIM) nouns take [-jə].

ALTERTP constructs such a grammar, which is a (partial) mapping from (SRC, FEATS) pairs to TGT's. The grammar partitions the triples: each leaf picks out a subset of triples subject to the rule whose structural description is encoded in the path to that leaf. The majority TGT of the triples is predicted for those triples and evaluated by the TSP. For example the left-most path in (11) picks out the PL DIM nouns, and predicts that they take [-s]. Because the grammar is a partition, every triple falls at exactly one leaf.

As any recursive procedure, the model can be specified by describing the conditions of recursion—the *recursive case(s)*—and the conditions in which recursion stops: the *base cases(s)*. The recursive procedure applies to a set of triples, but § 3.3.3 will describe how the recursive procedure fits into ALTERTP learning incrementally.

I will use the triples in (12) as an example input to a recursive step. The triples are a small snapshot of Dutch plurals curated to demonstrate the model.

(12) a. ([buːk], {PL, F-stress}, [-ən]) ([blum], {PL, F-stress}, [-ən])
([pɛt], {PL, F-stress}, [-ən]) ([vut], {PL, F-stress}, [-ən])

b. ([bɛt], {PL, F-stress}, [-ən] +alter)
([hœys], {PL, F-stress}, [-ən] +alter)
([hant], {PL, F-stress}, [-ən] +alter)

c. ([voːɣəl], {PL}, [-s]) ([lɛtər], {PL}, [-s])
([vɪŋər], {PL}, [-s]) ([mudər], {PL}, [-s])

The model description will work through (12) as an example, ultimately producing the grammar (20) in § 3.3.3. You may wish to follow along with Figure 3, which provides a diagrammatic summary.

3.3.1 *Recursive Step*

The input to a recursive step is a set of n triples, such as (12). AlterTP computes the most frequent tgt (i.e., the tgt occurring with the largest number of those triples), whose frequency is m. AlterTP checks whether the path to the current node is a *sufficient* condition for that most frequent tgt under the TSP: if $n - m \leq \theta_n$, then the current node is a sufficient condition for being in the set of triples with the most frequent tgt at the node. This is a *base case*, and recursion stops, here yielding a productive rule captured by the path to the current node.

On the other hand, if $n - m > \theta_n$, AlterTP chooses a feature to form a binary split of the triples at the node, as described in § 3.3.2. The feature is chosen to cover a *necessary* number (under the TSP) of the triples with the most frequent tgt.[4] The binary split creates two child nodes, and AlterTP recurses on each (the *recursive case*).

In the example input (12), [-ən] and [-s] both have frequency of 4, as they occur with 4 triples each. However, being at the current node is not a sufficient condition for either because, of the 11 triples at the node, 7 have a different tgt and $7 > \theta_{11} = 4.6$. Thus AlterTP attempts to subdivide the input based on some feature. Before turning to this, it will be helpful to define the ordinary notions of *necessity* and *sufficiency* as interpreted under the TSP and in the terms of the present work.

(13) A feature is **necessary** for (or covers a necessary subset of) triples T if m of the $n = |T|$ triples have the feature and $n - m \leq \theta_n$. In this case, being a triple in the set T implies, up to the TSP threshold, having the feature.

- Example: $m = 7$ of the $n = 11$ triples in (12) have the feature F-stress; since $4 \leq \theta_{11}$, F-stress covers a necessary subset of those triples.

(14) A (conjunction of) feature(s) is **sufficient** for being in a set of triples, T, if of the n triples matching the feature(s), m of them are in T and $n - m \leq \theta_n$. In this case, having the feature(s) implies, up to the TSP threshold, being in T.

- Example: $m = 4$ of the $n = 7$ triples with feature F-stress in (12) take [-ən] and are

[4] If no split is possible, then recursion stops (another *base case*) and the rule represented by the path to this node fails to be productive. Because, in this scenario, recursion stops without finding sufficient conditions, AlterTP learns a *partial* mapping from (src, feats) pairs to tgt's. Such cases are predicted non-productive conditions (e.g., Björnsdóttir 2021).

not subject to an alternation; since $3 \leq \theta_7$, F-stress is sufficient for being one of those triples.

3.3.2 Choosing a Split Feature

When a node fails to be productive, AlterTP seeks to subdivide the triples at the current node into two sets based on whether or not they have a certain feature. The main idea is (15).

(15) **Summary:** If the current node contains more than one suffix, AlterTP chooses a feature to separate the triples with the most frequent suffix. However, if a suffix has been isolated but is unproductive due to alternating nouns, AlterTP turns to looking for a feature to separate the alternating nouns. I will call the subset of triples that AlterTP attempts to separate from the rest, T.

AlterTP first searches for a feature shared by the triples with the most frequent suffix. In counting suffix frequency, AlterTP includes triples that alternate with those that do not, such that, for instance, "[-ən]" and "[-ən] +alter" triples both count towards the frequency of [-ən]. In (12), [-ən] occurs with 7 triples while [-s] occurs with 4, so AlterTP will first seek a feature shared by [-ən] triples. So AlterTP tries to separate triples T = (12a)-(12b) with the [-ən] suffix from triples (12c) with [-s].

AlterTP considers two kinds of features: features that are contained in a triple's feats and properties of the src's phonological form, which I will call FormFeats. The former are part of the input, believed to be acquired by children before the age under investigation. The FormFeats are learned on-the-fly by AlterTP.

FormFeat construction is based on the phonological learning model D2L (Belth, 2024), which uses a domain-general adjacency mechanism (Saffran *et al.*, 1996, 1997; Aslin *et al.*, 1998; Santelmann & Jusczyk, 1998; Saffran *et al.*, 1999; Fiser & Aslin, 2002; Gómez, 2002; Gómez & Maye, 2005) to identify segments that are sufficiently predictive of an alternation under the TSP and, when not possible, to automatically construct a tier representation by iteratively removing segments that are *not* sufficient predictors of the alternation on the current representation. This basic mechanism has formal support (Jardine & Heinz, 2016; Burness & McMullin, 2019; Yin, 2025) and accounts for developmental and experimental results (Belth, 2023b, 2025a,b). In the present work, D2L's adjacency-mechanism is used to identify segments that are predictive of the triples tgt, automatically constructing tiers as needed.

Using this D2L mechanism, AlterTP tries to construct a necessary (and if possible sufficient) FormFeat. A FormFeat created by this process (described in more detail in appendix A) matches a triple if it matches the suffix-adjacent segment of its src, possibly after projection onto a tier. For example, a feature like (16a) matches suffix-adjacent coronals while a feature like (16b) matches the suffix-adjacent segment after constructing a tier that deletes/excludes the segments in del(·). For example (16b) matches [ə] when suffix-adjacent on a vowel tier.

(16) a. [+cor] __

b. [ə] __ / del([−syl])

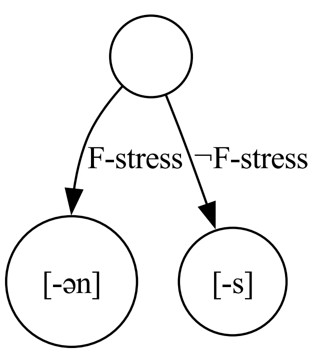
I will call the features AlterTP considers at a particular recursive step, F, which contains the features from the node's triples' feats and any constructed FormFeat. In finding a split feature, AlterTP seeks a feature that covers a *necessary* chunk of T, as defined in (13): being a triple in T implies, up to the TSP threshold, having the feature.[5]

With the feature options F in hand, AlterTP chooses a feature from F using the criteria in (17). Any features in F that would yield vacuous splits (i.e., all triples at the node share/lack the feature) or are not necessary for T are excluded from consideration. If no feature is necessary, then AlterTP proceeds to the next most frequent suffix.[6]

(17) Any features in F that are necessary and sufficient are preferred over those that are only necessary. Of the resulting features, AlterTP chooses the one that covers the most triples in T, breaking ties based on the number of non-T triples it removes, then lexicographically if still tied.

For (12), F = {Pl, F-stress, [−lat] __}. Feature Pl is shared by all triples and is thus vacuous. F-stress covers a necessary proportion of T (all of them) and because all F-stress triples are in T, it is also sufficient. Because F-stress removes more non-T triples (all 4) than does [−lat] __ (only 1), AlterTP splits on F-stress, yielding the tree (18).

(18) Grammar after first split



[5] AlterTP allows negated features, in which case necessity means being in T implies *not* having the feature. The only difference is T triples go down the right (not left) branch.

[6] Any suffixes occurring ≤ 5 triples are skipped, because the TSP is ill-defined for very small n (Belth *et al.*, 2021).

Down the F-stress branch, the next node starts a new recursive step. The most frequent TGT at that node is [-ən], with 4 of the 7 triples at that node having that TGT (3 have "[-ən] +alter"). Since $3 \leq \theta_7 = 3.6$, this rule is productive and recursion stops (the alternating nouns are lexicalized). Down the ¬F-stress branch, 4/4 triples have the [-s] TGT, so this rule is also productive.

When no split is found for a suffix, either because the remaining exceptions are due to the as-of-yet-unlearned alternation or because no further separation of alternative suffixes is possible, AlterTP turns its attention to separating alternating from non-alternating nouns. In this case, T comprises the triples that take the most-frequent suffix *and* alternate (e.g., all triples with a "[-ən] +alter" TGT). If no necessary feature is found for the alternating triples, AlterTP tries to find one for the non-alternating triples. The choice of feature, however, happens identically in all cases. The only difference is what triples are in T.

3.3.3 Incremental Update

AlterTP is incremental, learning one triple at a time. Suppose the next triple is (19).

(19) ([paːrt], {Pl, F-stress}, [ən] +alter)

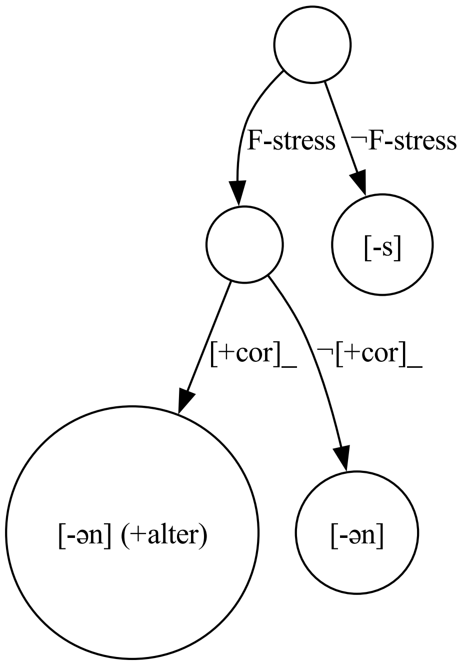
Upon receiving a new triple, AlterTP identifies which leaf (rule) the new triple matches and adds it to that leaf. If the rule represented by the leaf remains productive, then nothing further is needed. If the rule is not productive[7], AlterTP restarts recursion from the node (3.3.1). This is why the recursive step applies to a set of triples even though AlterTP is incremental. If all of the triples at the node fall under some rule produced by this recursion, then the grammar has been successfully updated. If some triples fall under unproductive conditions, the model traverses up the tree to the current node's parent and tries again. This continues either until the tree update is successful or the root node is reached.

In our example, because (19) has the feature F-stress, the left leaf of (18) matches. After adding the new triple to it, the node has $n = 8$ triples, $e = 4$ of which alternate. Since $4 > \theta_8 = 3.8$, productivity is broken. This triggers the recursive step.

All 8 triples take [-ən] as a suffix, so the 4 exceptions are due to the unlearned alternation. Thus, AlterTP seeks a feature to separate the alternating from the non-alternating triples. The features F-stress and Pl are both vacuous. However, the FormFeat constructed is "[+cor]__", which covers $m = 3$ of the $n = 4$ alternating triples while excluding all but the 2 [t]-final non-alternating triples. Since $1 \leq \theta_4 = 2.9$ this feature is necessary for the alternating triples and is used to form a new split (20).

[7] The triple could break rule productivity or it could already have been unproductive.

(20) Grammar after second split



Down the new branch, the TGT "[-ən] +alter" is now productive, because only $e = 2$ of the $n = 6$ F-stress, coronal-final nouns do not alternate, which is below the TSP threshold of $\theta_6 = 3.3$. To this point however, AlterTP only knows that something happens in addition to taking [-ən]. This is thus the point at which AlterTP turns its attention to learning the alternation.

3.3.4 How Alternations are Learned

Thus, AlterTP seeks to separate alternating from non-alternating nouns when the alternating nouns cannot be lexicalized as exceptions. Once a TGT that requires learning an unknown alternation (a "+alter" TGT) becomes productive (contingent on the alternation) at a leaf, AlterTP constructs non-surface-faithful representations for alternating segments at that leaf and turns to learning the alternation.

AlterTP sets up non-surface-faithful URs for each alternating stem at the node using the SRC and inflected forms. It aligns the two forms of the stem at their right edge and forms a new sequence of segments. Any segments that are identical between the aligned forms are copied; any that are not identical are merged into a set. This yields underlying-surface form pairs. In our example, (21) shows the aligned pairs and their derived URs, where /D/ = {[t], [d]} and /Z/ = {[s], [z]}.

(21)	[bɛt]	[hœys]	[hant]	[paːrt]	[pɛt]	[vut]
	[bɛdən]	[hœyzən]	[handən]	[paːrdən]	[pɛtən]	[vutən]
	/bɛD/	/hœyZ/	/hanD/	/paːrD/	/pɛt/	/vut/
	/bɛDən/	/hœyZən/	/hanDən/	/paːrDən/	/pɛtən/	/vutən/

Thus, alternating stems at this node now receive non-surface-faithful URs. The UR-SR pairs are added to a global set shared by all nodes where non-surface-faithful URs are introduced so that the alternation is not learned individually at each node where it is predicted productive. Over these pairs, AlterTP learns a phonological grammar using PLP (Belth, 2023a). PLP learns phonological rule(s) to predict the surface form of alternating segments—e.g., in what contexts /D/ → [d] and in what contexts /D/ → [t]—by incorporating adjacent segments into a rule and iteratively increasing its window of attention beyond those until the rule passes the TSP.[8] See Belth (2023a) for details on how learning multiple alternations and interactions is handled in the PLP framework. In our example (and the experiments), the right-adjacent context is predictive of the alternation, which PLP incorporates into the learned rule on the first iteration (22).

(22) {D, Z} → [−voi] / __ #
→ [+voi] / __ /ə/

AlterTP then checks whether one of the forms in each set /D/ and /Z/ is an elsewhere/default form. To do so, it tries replacing /D/ and /Z/ with their [+voi] or [−voi] members and checks whether either resulting rule, /d, z/ → [−voi] / __ # or /t, s/ → [+voi] / __ [ə], makes the same predictions as (22). Here, since both [d] and [t] occur before [ə] but only [t] before #, the [+voi] forms are learned as the elsewhere forms and reorganized into the UR by replacing /D/ with /d/ and /Z/ with /z/ in both the URs and the rule. This results in a familiar devoicing rule (23).

(23) /d, z/ → [−voi] / __ #

Since new nouns falling at this node are predicted to alternate, such stems receive non-surface-faithful URs: /d/ for any final [t] and /z/ for any final [s].

In summary, any noun that falls at a leaf where the alternation is productive, like the left-most leaf in (20), is predicted to be an alternating noun, and hence is given a non-surface-faithful UR. In other words, predicting the alternation amounts to predicting the underlying voicing of stem-final obstruents. Any non-alternating nouns at the leaf are lexicalized. AlterTP predicts learners to have robust knowledge of the voicing alternation for nouns at these nodes, as well as to extend the alternation productively to new words falling at it. Nouns falling at other leaves are not predicted to alternate; any such alternating nouns will be lexicalized. However, because phonological rules are learned globally, devoicing applies productively to any new nouns. For instance, if a learner encounters a wug Pl [slɑdən], removing the suffix will strand [d] in final position, subject to devoicing (23), even if the same learner instead encountering the wug Sg [slɑt] might not predict it to alternate. As shown in (§ 4.3-4.4), this asymmetry—where the productivity

[8] If PLP fails to learn a rule, the node is deemed unproductive, though this never occurs in the experiments as the alternation is predictable from the right-adjacent context.

of the alternation is highly restricted while devoicing eventually comes to be applied productively—aligns with a developmental asymmetry between these two kinds of knowledge.

3.3.5 Accessing Known Forms and Predicting Novel Forms

Inflecting a noun. Given a noun form with some features, AlterTP can predict the form of that noun with an additional inflection feature (e.g., predict the Pl from a Sg or the Pl Dim from a Dim). To do so, AlterTP traverses the learned tree based on the noun's features and the inflection feature until it reaches a leaf. Since the grammar partitions the nouns, exactly one leaf will match. If the leaf is productive, it is applied to predict the inflected form. If the alternation is productive at that node, the stem-final segment is given a non-surface-faithful-UR. Any phonological processes then apply to predict the SR for the inflected UR. If the leaf is not productive, the input form is returned uninflected, mirroring the fact that in wug tests, children who do not know how to inflect a wug often repeat it uninflected (e.g., Berko 1958). For example, if the tree (20) is applied to the Sg wug [slɑt], which has final stress and ends in a [+cor] segment, the left-most leaf matches. The surface [t] will be parsed to /d/, and the inflection will yield /slɑdən/. AlterTP can also inflect a noun lemma into any paradigm slot by applying the same method iteratively for each feature characterizing the target paradigm slot.

Uninflecting a noun. AlterTP can also *un*inflect a noun for a particular feature (e.g., predict a Sg from a Pl). To do so, AlterTP searches the learned inflection tree's leaves for any productive leaf that predicts the input word's ending (suffix). The suffix predicted by that leaf's rule is then removed. Any phonological rules (e.g., devoicing) that have been learned are then applied to predict the SR. For example, if the tree (20) is applied to the Pl wug [slɑdən], the [-ən] suffix matches and is removed, yielding [slɑd], which would then be devoiced to [slɑt]. The same would apply to a wug like [klatən], only devoicing would not apply, yielding [klat].

Known Nouns. At each leaf, exceptions are lexicalized, meaning each (src, feats) pair is stored with its full tgt form. To produce the inflected form of a known src, AlterTP identifies the appropriate node (as above) and checks the exceptions. If the src exists, the stored tgt is returned; otherwise, the rule is applied. Stems, unless lexical exceptions, are stored in the lexicon with the underlying form predicted by the grammar (i.e., surface faithful unless predicted to alternate). AlterTP does not predict that children produce voiced final obstruents in known words. Each alternating noun either falls at a leaf where the alternation is productive or one where it is not. If unproductive, then the stem form, stored surface-faithfully, is produced. If productive, then devoicing has been learned and applies. Either way, the form produced will not violate the phonotactic restriction. For discussion of predictions on wugs, see § 4.3.

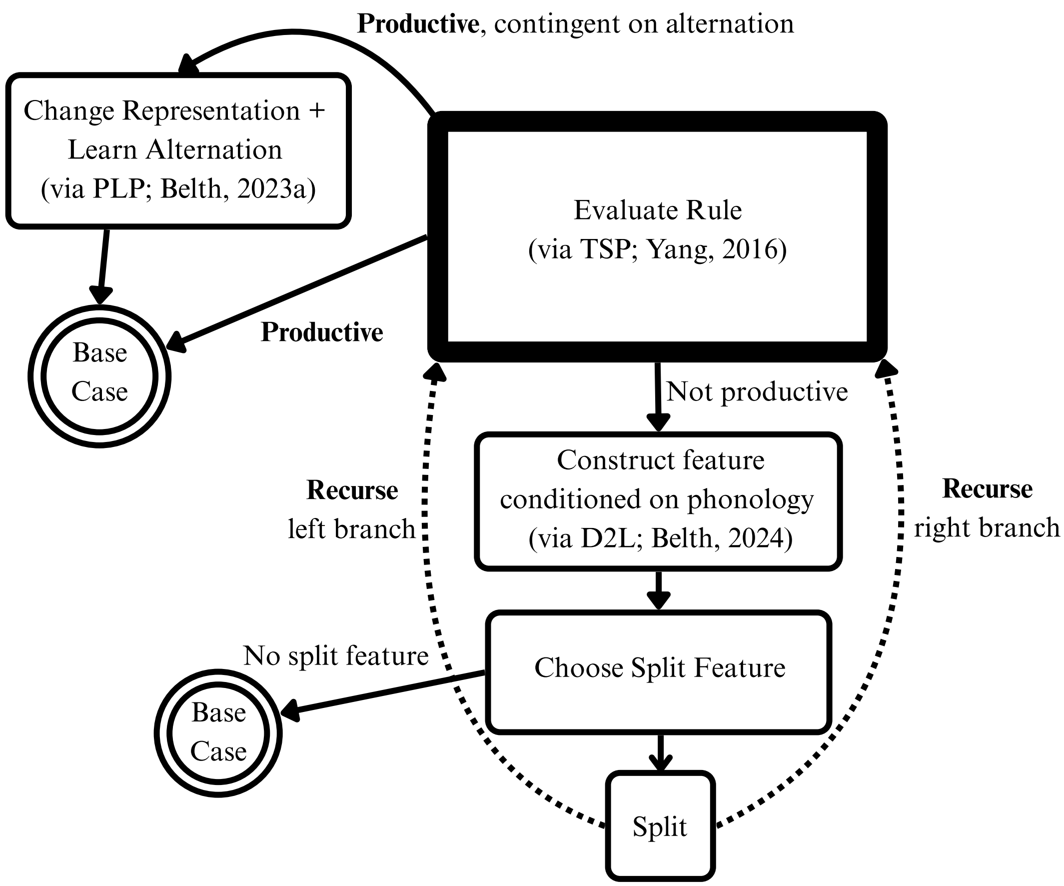


Figure 3: Diagram of AlterTP's recursive logic.

3.4 Model Summary

The model is summarized in Figure 3. AlterTP recursively tries to separate triples that take the same TGT from those that take a different one. Each split is necessary for the set of triples the model is currently trying to account for, and recursion stops when the conditions leading to the current recursive state (node) are sufficient for the TGT trying to be explained (or when no further splitting is possible). Thus, the model learns necessary and sufficient conditions for each TGT. When trying to explain a particular suffix, if enough of the triples taking that suffix alternate to prevent the suffix from being sufficient, AlterTP subdivides to separate the alternating from the non-alternating triples; otherwise, the alternating exceptions are lexicalized. When AlterTP finds necessary and sufficient conditions for a TGT that includes stems alternating, AlterTP then turns its attention to learn the alternation for those stems.

Because AlterTP incrementally partitions the input into subsets, represented by the leaves of the tree, the alternation might be learned down some but not all branches, as in (20). That is, not all nouns alternate, and AlterTP learns the structural conditioning of that lexical variation. Thus, even though the TSP is categorical, productivity is not all-or-nothing at the lexicon level; rather, it is conditioned by each subset of the partitioned lexicon. This is in the spirit of other work on generalizations over sublexicons (e.g., Itô & Mester 1999; Gouskova *et al.* 2015).

4 Dutch and German Development

I now turn to how AlterTP explains results from prior studies on the acquisition of Dutch and German voicing alternations. I first describe the datasets and setup (4.1).

4.1 Data and Setup

4.1.1 Datasets

I constructed a Dutch dataset from the Van Kampen (Van Kampen, 2009) and Groningen (Bol, 1995) CHILDES corpuses (MacWhinney, 2014), which together contain child-directed speech (CDS) for 9 Dutch-learning children, aged 1;05-6;0. After processing the dataset as described in appendix B, it consists of 1562 (src, feats, tgt) triples, falling into 1051 unique noun paradigms. The src and tgt are transcribed in IPA and the feats are Pl/Dim and prosodic features (the placement of stress in the lemma and whether the lemma is monosyllabic) believed to be acquired prior to the ages being modeled. The triples are also annotated for token frequency.

For German, I extracted all CDS nouns from the Leo (Behrens, 2006) CHILDES corpus, which contains CDS from 1;11 to 4;11. After processing the dataset (also described in appendix B), it contains 9390 triples falling in 4195 noun paradigms. The src and tgt were again transcribed in IPA and the triples annotated for token frequency. The feats consist of grammatical gender, number (Pl), and case.

4.1.2 Simulating Acquisition Trajectories

To simulate developmental trajectories for multiple learners, I took 30 random samples (using 30 seeds for pseudorandom initialization) of triples from each dataset; each sample was weighted by triple token frequency. The sampled triples were added to a list of training items in the order they were sampled, and AlterTP was trained incrementally on this. Timepoints were inserted from 100 to 1500 in increments of 100 to produce a sequence of developmental timepoints for evaluating learning models periodically during training. These increments were also used for batch learning by the non-incremental comparison models (§ 4.1.3). The training set at each training size was a proper subset of the next-larger training set (e.g., training size 400 contains all the words of training size 300, plus 100 more). For each seed, the triples not sampled are test triples, available for evaluating generalization behavior (§ 4.5). This sampling reflects the fact that, due to sparsity § 3.1, paradigms are not typically complete.

In the results below, I report evaluation metrics against training size. The experiments primarily compare to children age 2;6 and above. Szagun *et al.* (2006) estimated that German-learning children know on average 429 (± about 100) words by age 2;6 and Bornstein *et al.* (2004) estimated that across several languages (including Dutch), children's vocabularies of 201-500 words contain on average 54% nouns. Thus, 200 training triples is the chosen starting size for the experiments. Due to the size of the Dutch dataset, the largest training size is 1500.

4.1.3 Comparison Models

To compare to the view of alternation learning that early acquisition of phonotactics underlies the learning of phonological alternations, I used the UCLA phonotactic learner (Hayes & Wilson, 2008) to learn a set of phonotactic constraints from the surface forms present in the training data. In § 4.3, I will describe how I used this model to evaluate a phonotactic-driven theory of alternations against my proposal in reverse wug tests that evaluate Dutch-learners' knowledge of devoicing. Since the UCLA phonotactic learner is a model for learning phonotactics, it is not applicable to the experiments involving morphological inflection; the results speak to the view of alternation learning as arising from phonotactic knowledge, not the UCLA learner as a model of phonotactics.

I also chose a neural network to compare to, as an exemplar of connectionist approaches to acquisition (Rumelhart & McClelland, 1986): a character-level, transformer-based encoder-decoder model augmented to incorporate morphological features based on Wu *et al.* (2021)'s model, which is commonly used in morphophonological tasks (e.g., Kodner & Khalifa 2022; Goldman *et al.* 2023). I will call this model Neural. I describe the implementation and hyperparameter tuning in appendix C.

4.2 Lexical Knowledge of Alternation

One aspect of voicing alternation acquisition is *lexical knowledge*. This includes the ability to correctly produce the stem-final obstruent of Pl nouns and the ability to recognize mispronunciations of such obstruents. The former has been investigated in Dutch through corpus analysis of child speech and elicitation studies, while the latter has been studied experimentally in an age-controlled experiment comparing Dutch- and German-learning children. I will begin with the Dutch studies.

4.2.1 Voicing Error Rates on Alternating and Non-Alternating Plurals

Kerkhoff (2007:ch. 4) studied the rates of mispronunciations of stem-final [t] and [d] in Pl nouns in the CLPF corpus (Fikkert, 1994; Levelt, 1994), which contains productions from twelve Dutch-learning children age 1;0-2;11. Kerkhoff reports that of 43 Pl productions with stem-final [t] targets, 0 were incorrectly produced as *[d]. In contrast, 16/66 (24.2%) [d] targets were incorrectly

Table 1: Summary of voicing error rates by Dutch-learning children for target [t] and [d] in producing noun plurals. Column averages are given in the final row.

Source	Type	Age	*[d] (target [t])	*[t] (target [d])
Kerkhoff (2007:tab. 17)	Corpus	1;0-2;11	0.0%	24.2%
Zamuner *et al.* (2012)	Elicitation	2;7-3;7	2.0%	15.5%
Kerkhoff (2007:ch. 5)	Elicitation	3;0-7;0	3.1%	41.6%
Average			1.7%	27.1%

produced as *[t].

Zamuner *et al.* (2012) elicited [t] and [d] plurals from Dutch-learning 2;7 and 3;7 children. They found 1% and 3% error rates for the respective ages for [t]-target plurals, compared to 24% and 7% for [d]-target plurals. Kerkhoff (2007:ch. 5) performed a similar elicitation study, and found 2.8% (age 3;0), 4.6% (age 5;0), and 1.7% (age 7;0) voicing error rates for target [t], compared to 42.2%, 41.4%, and 40.8% for target [d].

These results are summarized in Table 1 (results are averaged across the age groups of children in each study). Erroneous productions of *[d] for target [t] are rare, while productions of *[t] for target [d] are comparatively common. The dominance of *[t] for target [d] errors could at least in part be due to children's lack of knowledge of the voicing alternation, perhaps indicating the overapplication of a plural "add [-ən]" morphological rule that does not encode the alternation.

Given the substantial number of voicing errors, it is worth discussing the possible role of the acquisition of the voicing contrast in the subsequent acquisition of the voicing alternation. Dutch is a prevoicing language while German is an aspirating language (Buckler & Fikkert, 2016). In production, Dutch-learning children accurately produce medial [t] by age 2;11 (van der Feest, 2007) and medial [d] by age 4;5 (Kuijpers, 1993) (perhaps by 3;6, Zamuner *et al.* 2012). The Dutch-learning children modeled below fall in the range 2;7-7;2, except for Kerkhoff's study of voicing error rates for age 1;0-2;11 children in the CLPF corpus. Based on the same CLPF corpus, van der Feest observes that for all medial [t]'s and [d]'s (not necessarily plurals), children more often produce an incorrect *[t] for a medial [d] (28.4%) than the reverse (1.5%). This is the same pattern of errors described above for stem-final [t]/[d] when in medial position in a plural. Thus, it is possible that the earliest asymmetry in error types reported in the studies in Tab. 2 could at least in part be due to the acquisition of the voicing contrast, rather than a reflection of the acquisition of the voicing alternation. However, Kerkhoff's elicitation study of plurals spans children age 3;0-7;0, well past the age that children's production of medial voicing is reported to be acquired. In fact, Kerkhoff grouped 3-year-olds, 5-year-olds, and 7-year-olds, but found no effect of age, indicating that the error asymmetry between medial [t] and [d] in plurals cannot simply be attributed to the general medial voicing errors observed early in development.

Furthermore, van der Feest concluded that, while voicing error rates in production occur, children are representing the voicing contrast (evidenced by a perceptual study) by age 2;0, though the perceptual study was limited to the contrast in initial position. Kerkhoff (2007) concludes that the voicing contrast is acquired by around age 3. In the studies discussed below, Dutch children's knowledge (or lack thereof) of the voicing alternation persists well past this age, indicating that the acquisition of the voicing alternation cannot simply be attributed to the acquisition of the voicing contrast.

As far as I am aware, the acquisition of the voicing contrast in German has only been studied in initial position. This prior research suggests that the contrast is acquired early, by around age 2 (Kager *et al.*, 2007). Thus, the voicing contrast is probably acquired earlier in German than in Dutch. However, the asymmetrical developmental trajectory of the voicing alternation persists well past the age the contrast is acquired in both languages, even to adulthood (§ 4.4).

4.2.2 Knowledge of Which Nouns Alternate

In an age-controlled experiment, Buckler & Fikkert (2016) tested Dutch- and German-learning 3-year-olds' sensitivity to mispronunciations in alternating and non-alternating plurals. They used an eye-tracking experiment that allows for inferring relative sensitivity to mispronunciations based on the speed and duration of children's looks to target items. They found a greater degree of sensitivity to mispronunciations of both alternating and non-alternating plurals among the German-learning children than the Dutch-learning children. This could indicate that German 3-year-olds have more advanced knowledge of which nouns alternate and which do not than their Dutch peers.

4.2.3 Evaluating Model Lexical Knowledge of Alternation

To compare to Dutch child error rates, I probed each model at each training size and seed to predict the Pl of the 30 most-frequent /t/- and /d/-final nouns in the Dutch dataset. The children's error rates could in theory be comprised of a mixture of Pl nouns that they know from experience and those that they have predicted the form of despite not having learned it from experience. Evaluating over a fixed set of Pl nouns allows for simulating this, since some may be training items and some unseen. Moreover, children's rates were computed over *tokens* and some tokens may be retrieved in whole rather than be derived via the grammar (Bybee, 1995). Using the fraction of errors rather than absolute error rates allows for comparison between models and humans while factoring out these aspects of production, which affect both types of errors. Thus, I computed the fraction of predicted voicing errors that are /d/ → *[t]. Lastly, the studies of children did not find any significant trends in errors with age, so I average error rates across training sizes and seeds.

Table 2 shows the average fraction of predicted voicing errors in Dutch that are /d/ → *[t] for each model. Errors producing a *[t] for a target [d] amounts to generalizing the [-ən] suffix without

Table 2: The fraction of predicted voicing errors for each model that are /d/ → *[t] rather than /t/ → *[d]. For comparison, I show the average for children across studies.

Model	Average Fraction /d/ → *[t] Errors
Human	0.9410
AlterTP	0.9335
Neural	0.7110

applying a voicing alternation when one in fact occurs; producing *[d] for target [t] amounts to over-applying the voicing alternation to a noun where it in fact does not occur. The vast majority of AlterTP's voicing errors are /d/ → *[t], comparable to the relative error rates of children in acquisition studies.

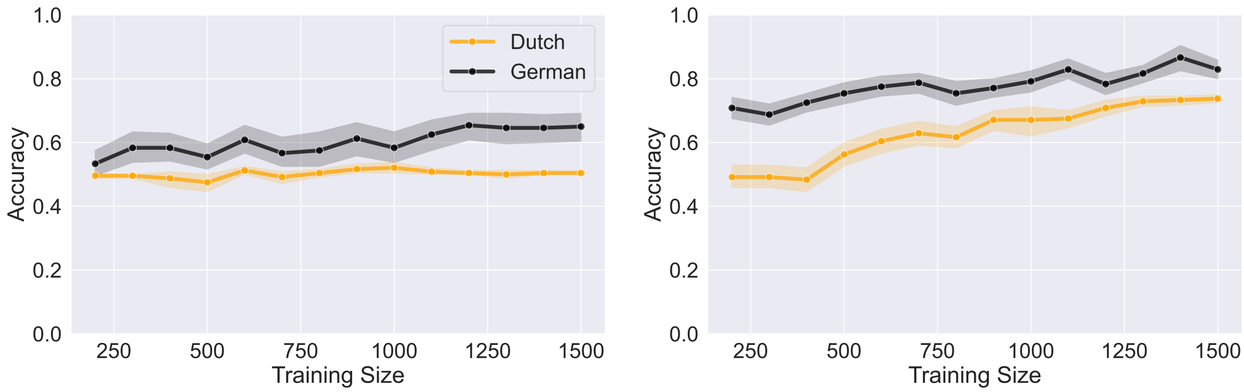
Neural also shows a discrepancy between error types, but its fraction of /d/ → *[t] errors is well below children's: it overextends the alternation too much.

To compare to Buckler & Fikkert (2016)'s children's sensitivity to mispronunciations, I had each model predict the Pl nouns used by Buckler & Fikkert (see tables 4 and 6 of their paper). Buckler & Fikkert identified these as nouns the Dutch and German children were likely to know, but it is ultimately unknown whether every child in the study indeed knew all the words. Similarly, by probing the models for predictions on these words, they are likely to be words seen during training, but it is not guaranteed. I compute the accuracy of each model at predicting [t] vs. [d] for these words in Dutch and German.

Figure 4 shows the model accuracies at predicting the stem-final stop voicing of Buckler & Fikkert's stimuli (error bands are 95% confidence intervals). The AlterTP models trained on German correctly predict voicing for a substantially higher fraction of nouns than the Dutch-learning models. This is consistent with Buckler & Fikkert (2016)'s finding that German-learning 3-year-olds show better knowledge of which nouns alternate than Dutch-learning 3-year olds. Neural also matches this result.

4.3 *Productivity of Devoicing*

Another aspect of knowledge of the voicing alternation is the productivity of devoicing. Zamuner *et al.* (2006), Kerkhoff (2007:sec.6.2), and Zamuner *et al.* (2012) performed reverse Wug tests for Dutch-learners, which involve asking participants to produce or identify the Sg form of a Pl wug, flipping the classic Sg → Pl direction (Berko, 1958). Example stimuli from Zamuner *et al.* (2006, 2012) (stimuli were comparable in Kerkhoff 2007) are in (24). Some wugs had stem-final voiced obstruents (24a), others voiceless (24b).



(a) AlterTP **(b)** Neural

Figure 4: Accuracy on Buckler & Fikkert (2016)'s Dutch and German nouns.

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

Zamuner *et al.* (2006) performed an identification reverse wug test, where age 2;7-3;7 Dutch-learning children were auditorily presented a Pl wug together with two images of an imaginary animal. The children were then presented three images—the same Pl image, a Sg version of the image, and a Sg distractor image of another animal—and asked to identify the Sg, which they heard auditorily. For example, "These are two [slɑdən]" followed by, "Can you find the [slɑt]"?

Zamuner *et al.* (2012) carried out a production version of the same experiment with children of the same age, asking them to produce the Sg. For example, "These are two [slɑdən]" followed by, "What's this?". Kerkhoff (2007:sec.6.2) performed a similar production reverse wug test on children 3;1-6;0 and adults.

Since identifying or producing the Sg wug involves morphologically decomposing the Pl wug into a stem and [-ən] suffix, which is a familiar suffix to Dutch-learning children, this task is similar to experiments testing children's ability to morphologically decompose wugs inflected with a familiar suffix. In those experiments, children robustly demonstrate an ability for such morphological decomposition (Marquis & Shi, 2012; Ladányi *et al.*, 2020) as early as 0;6 (Kim & Sundara, 2021). For non-alternating wugs, the task involves only morphological decomposition, but for alternating wugs, it also requires productively devoicing the final obstruent, which removing [-ən] strands in final position. Thus, if children have learned a productive devoicing process, they should perform comparably across alternating and non-alternating trials.

Table 3 summarizes the results from these reverse wug tests. Children appear to have difficulty with the tasks overall, but they were substantially better at the identification task than the production task (an example of the common comprehension/production asymmetry). Importantly, in every experiment and at every age, children were statistically significantly better at identifying/producing

Table 3: Summary of prior study reverse wug test results. The last two columns show the average accuracy at identifying/producing the anticipated singular for Non-Alternating and Alternating wugs. "Average" rows are column averages for each task.

Source	Task	Age	Non-Alter	Alter
Zamuner *et al.* (2006:tab.2)	Identification	2;7	0.67	0.50
Zamuner *et al.* (2006:tab.2)	Identification	3;7	0.50	0.43
Average			0.59	0.47
Zamuner *et al.* (2012:tab.4)	Production	2;4	0.22	0.14
Zamuner *et al.* (2012:tab.4)	Production	2;7	0.17	0.13
Zamuner *et al.* (2012:tab.4)	Production	3;7	0.44	0.23
Zamuner *et al.* (2012:tab.4)	Production	4;8	0.38	0.13
Kerkhoff (2007:tab.39)	Production	4;8	0.53	0.31
Average			0.35	0.19
Kerkhoff (2007:tab.42)	Production	Adults	0.99	0.93

the Sg for non-alternating than alternating wugs. This suggests, at least until age 4;8, a lack of a robustly productive devoicing process, which—if present—should have yielded comparable results for the two wug types. Since adults show much better performance in the alternating condition, it is likely that productive devoicing is eventually acquired. No reverse wug test studies are available for German.

4.3.1 Evaluating Productivity of Devoicing in Models

To evaluate the productivity of devoicing, I had each model predict the Sg form of the Pl wugs from Zamuner *et al.* (2006, 2012). I interpret model predictions in terms of the identification task, since children were overall better at it, but this choice is unimportant since the children's trend was the same in both tasks. If a model-predicted Sg did not match the Sg spoken by the experimenter, the identification was left to a random guess. I used a 0.5 probability of guessing the correct form when the predicted Sg did not match, since children had three image choices, but never chose the distractor.

AlterTP predicts the Sg as described in (3.3.5). So long as a [-ən] rule is learned, non-alternating forms like [slɑtən] are mapped to [slɑt] and are correctly identified. Alternating forms like [slɑdən] are mapped first to /slɑd/, stranding [d] in final position. If AlterTP has learned the [d] ~ [t] alternation at some node, then it will have also learned a devoicing rule, which leads AlterTP to predict [slɑt]. If AlterTP has not yet learned a devoicing rule, then the prediction will not match the expected Sg.

For Neural, when training on Dutch, I augmented the training data with a Sg → Pl pair for each training Pl → Sg pair so that the model was also trained for predicting Sg forms from Pl forms. Neural was then prompted to predict the Sg for each Pl wug; predicted forms that matched the Sg wug (i.e., had a final [t]) were treated as correct identifications.

I also trained the UCLA phonotactic learner on the surface forms for all the training nouns at each training size and each seed, each time yielding a set of phonotactic constraints. Then, for each Pl wug, I generated two candidate forms for the Sg by removing the [-ən] suffix: one with a final [d] and one with a final [t] (e.g., [slɑdən] to [slɑd] and [slɑt]). I then ran these through UCLA's phonotactic grammar to generate a harmony score for each. If the form spoken by the experimenter (the [t] form) had a higher harmony score, it was deemed that the model correctly identified the Sg.

Any model failing to predict a final [t] for an alternating wug should not be taken as a prediction that children would *produce* such forms. Indeed, in the production reverse wug tests, children never did so. This could be interpreted as evidence that they know the phonotactic restriction, but it is not evidence that they know devoicing, since we would then expect children to produce a devoiced form. Instead, children consistently repeated the plural (Kerkhoff 2007:p.224; Zamuner *et al.* 2006:p.491). This is consistent with AlterTP, which does not predict a lack of phonotactic knowledge, but rather that knowledge of phonotactics does not automatically trigger alternation learning.

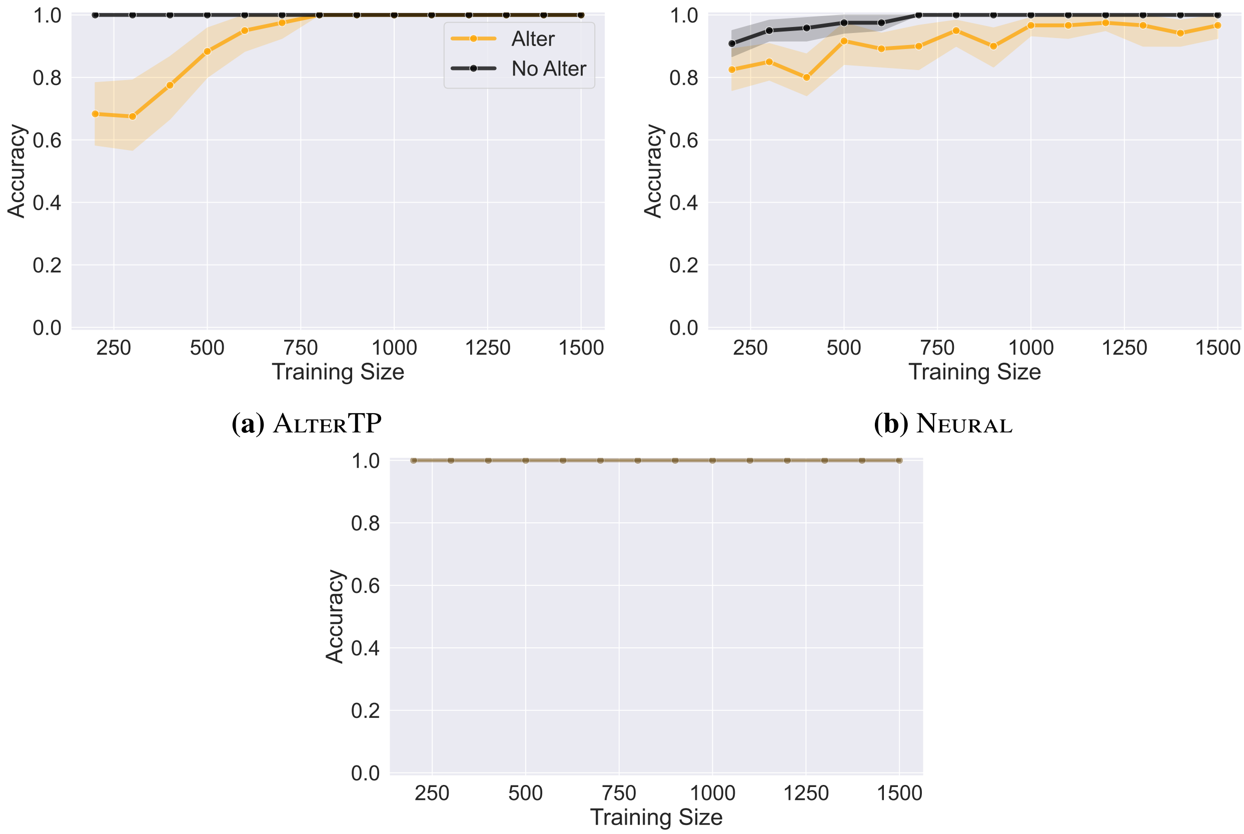
Figure 5 shows accuracies at identifying the Sg. AlterTP consistently matches children's asymmetry between non-alternating and alternating wugs until 600-700 training triples, where AlterTP comes to consistently learn devoicing. This aligns with both the fact that children show limited command of devoicing (hence an asymmetry) and that adult performance (0.93 accuracy) suggests that devoicing is eventually acquired. Indeed, Kerkhoff (2007:p. 15) argues that devoicing in loanwords also provides evidence of its eventual productivity.

Neural shows a slight asymmetry at most training sizes. UCLA's accuracies are always 1.0 for both wug types because it always learns a constraint against voiced final obstruents and consequently always assigns higher harmony scores to the [t]-final Sg than the [d]-final competitor. This failure to produce an asymmetry aligns with Zamuner *et al.* (2006:p. 495)'s conclusion, "we found little evidence that the early acquisition of phonotactics is generalized to the acquisition of morphophonological alternations."

Model accuracies are mostly higher in absolute terms than those of the children, indicating the overall difficulty of the task for children.

4.4 Productivity of Alternation

The final type of knowledge tested concerns the productivity of the voicing alternation. That is, not all nouns with obstruent-final Sg forms alternate; if presented with an obstruent-final Sg wug,



(a) AlterTP

(b) Neural

(c) UCLA

Figure 5: The fraction of Pl wugs for which each model correctly identified the Sg. Orange shows alternating wugs (e.g., [slɑdən]); black shows non-alternating wugs (e.g., [slɑtən]).

do learners expect it to alternate? This question has been studied for Dutch (Zamuner *et al.* 2006; Kerkhoff 2007:ch.5), and German van de Vijver & Baer-Henney (2014) using Sg-to-Pl wug tests.

A summary of results from prior wug tests is provided in Table 4. German learners extend the alternation to wugs at much higher rates than comparably-aged Dutch learners. Productivity of the alternation is highly restricted for Dutch learners, as children and adults rarely extend it to wugs. Even adult Dutch speakers only extend the alternation 2.3% of the time.

Zamuner *et al.* (2006) performed an identification Wug test where 2;7-3;7 year old Dutch-learning children were asked to identify a Pl image of a wug Sg: for example, “This is a [slɑt]” followed by “Can you find the [slɑdən]?”. Children at both ages were worse at identifying the plural for alternating than non-alternating wugs (0.75 alternating vs. 0.59 non-alternating accuracy at 2;7 and 0.7 alternating vs. 0.57 non-alternating accuracy at 3;7). These corroborate, from a comprehension perspective, the conclusion of limited productivity in the production Wug tests.

Similarly, Kerkhoff (2007:ch. 6) performed an acceptability experiment, in which children were presented a Sg wug and then its Pl with an alternation and asked whether the word was correct—

Table 4: Summary of wug test results from prior studies. Column "Alternation %" gives the percentage of wugs for which participants produced a voiced obstruent in the Pl.

Source	Language	Age	Alternation %
Kerkhoff (2007:tab.23)	Dutch	3;4	3.5%
Kerkhoff (2007:tab.23)	Dutch	5;2	3.5%
van de Vijver & Baer-Henney (2014:tab.4)	German	4;9	32%
Kerkhoff (2007:tab.23)	Dutch	7;2	1.9%
van de Vijver & Baer-Henney (2014:tab.4)	German	7;1	21.4%
Kerkhoff (2007:tab.35)	Dutch	Adults	2.3%
van de Vijver & Baer-Henney (2014:tab.4)	German	Adults	12%

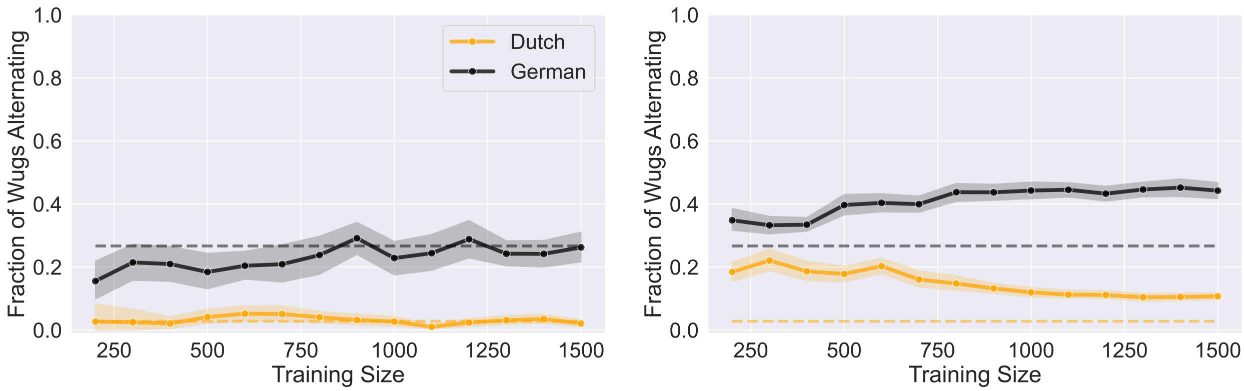
for example, "This is a [slɑt]" followed by "Now there are two, are they [slɑdən]?" The rates of acceptance were compared to a place alternation [t] ~ [p] (e.g., [slɑpən]) that is unattested in Dutch. The children (aged 3;1 to 4;0) accepted the (voice) alternating Pl forms 47% of the time, but also accepted the place alternating Pl forms 37% of the time, which was not a statistically significant difference. Again, these results are consistent with the conclusion that the alternation has very limited productivity for children.

4.4.1 *Evaluating Productivity of Alternation in Models*

To compare the models to these Wug test results, I probed each Dutch-learning model to predict the Pl for each wug from Zamuner *et al.* (2006, 2012) and from Kerkhoff (2007:p. 295), and each German-learning model to predict the Pl for each wug from van de Vijver & Baer-Henney (2014). I computed the fraction of predictions for which each model predicted a stem-final voiced obstruent in the predicted Pl.[9] German nouns have gender, but it is not clear how the participants in van de Vijver & Baer-Henney (2014)'s study interpreted the wugs' gender, or if they were cued to gender by a determiner.[10] Consequently, I had each model make predictions for the wug one time for each gender and then sampled a prediction, for AlterTP weighted by the scope of the rule that produced it, and, for Neural, uniformly.

[9] Not all the German wugs were obstruent-final. The denominator was the number of obstruent-final nouns for which the model predicted a form different from the Sg, since only those potentially alternate. Through correspondence with van de Vijver & Baer-Henney, I confirmed that this is also how they computed the fraction.

[10] German uses different determiners (*der/die/das*) for different genders. van de Vijver & Baer-Henney do not report which they used to present wugs and it is not clear that participants would use the determiner anyways, rather than using the phonology of the wug to guess gender (Zaretsky & Lange, 2015; Belth *et al.*, 2021).



(a) AlterTP **(b)** Neural

Figure 6: The fraction of wugs for which each model predicted an alternation. The dashed lines show averages of human alternation rates. AlterTP tightly matches human alternation rates in both languages.

Figure 6 shows the results for Dutch (orange) and German (black), where the y-axis shows the fraction of wugs that each model predicted to alternate. The dashed lines show the average alternation rates for children.

AlterTP predicts the difference between Dutch and German alternation rates and tightly matches the average children's alternation rates for both languages (2.8% and 26.7%, respectively). The close match to German children's alternation rates, which are well below 100%, suggests that AlterTP's learned structural conditioning for the lexical variation underlying the alternation is compatible with what children learn (see § 4.5 for an example learned tree). Moreover, there is no upward trend with training size for Dutch, consistent with the fact that Dutch-speaking adults extend the alternation at a comparable rate to children.

Neural shows a difference between Dutch and German alternation rates, but is far more likely to extend the alternation in both languages than children were.

Though there were few instances of children extending the alternation productively, Kerkhoff (2007) reported the fraction of alternating responses for each wug rhyme type. The most common context for children's alternating responses was a final [t] following a nasal. This is likewise the most common context where AlterTP sometimes learns the alternation productively. This is because alternating nouns are disproportionately common in this context in Dutch (Kerkhoff:ch. 4), and AlterTP sometimes identifies this context as predictive of the alternation. Table 5 shows, across rhyme types, the correlation between the alternation rates reported by Kerkhoff and those produced by each model. AlterTP has the highest correlation, whether measured under an assumption of a linear relationship (Pearson's r) or just a monotonic relationship (Spearman's ρ).

Table 5: Correlations, across wug rhyme types, between the fraction of alternations produced by children (Kerkhoff, 2007) and the fraction of alternations predicted by models.

Model	Pearson's r	Spearman's ρ
AlterTP	**0.747** ($p < 0.01$)	**0.674** ($p < 0.01$)
Neural	0.499 ($p = 0.04$)	0.650 ($p < 0.01$)

4.5 Example Inflection Trees and Generalization Performance

I give example inflection trees that AlterTP learned for Dutch and German, each after training on 300 triples. The Dutch tree (Fig. 7) contains two main branches, one for Dim nouns and one for non-Dim. The Dim branch contains rules for [-jə] and [-tjə] but not the other Dim allomorphs, which is consistent with the fact that [-jə] and [-tjə] are acquired earliest (Os & Harder, 1987; Van Tuijl & Coopmans, 2021); AlterTP indeed begins learning the other allomorphs at larger training sizes. As expected, the choice between [-ən] and [-s] Pl suffixes (right of tree) is determined by prosodic features of the stem. No leaf in this tree triggered the alternation to be learned—all alternating nouns are lexicalized as exceptions to one of the productive nodes. It is common, across training sizes and seeds, for most Dutch alternating nouns to be lexicalized. However, by the largest training size (1500), the alternation is sometimes learned for some restricted classes of words, in particular where a [t] is preceded by a nasal (see § 4.4.1), which causes devoicing to be consistently learned.

The German tree (Fig. 8, "*" marks unproductive nodes) is large, due to the complex conditioning factors of the various suffixes, but familiar generalizations are present, like the early split on F (feminine) and the narrow applicability of [-s] and [-ɐ] (*-er*) (Marcus *et al.*, 1995; Wiese, 1996; Yang, 2016). Unlike Dutch, the German tree contains a number of leaves where the alternation was learned (marked with "+alter" and highlighted in orange). These leaves will map stem-final obstruents to non-surface-faithful URs. The greater presence of "+alter" leaves for German than for Dutch, indicates that in German, where alternations are more prevalent (§ 2), AlterTP more often needed to recursively subdivide to separate alternating from non-alternating words in order to construct sufficient morphophonological generalizations.

I also evaluated the overall generalization ability of AlterTP. Since AlterTP demonstrated limited productivity of the alternation for Dutch, its learned generalizations will predict few alternations for test words. If generalization accuracy for Dutch is nevertheless high and productive rules cover most nouns, this supports the argument that alternations may be a small enough corner of the nominal system to be largely tolerated as exceptions.

Figure 9a shows AlterTP's accuracy predicting held-out (not seen during training) test forms for Dutch and German. The dashed lines show Neural's accuracy for comparison. On both languages, AlterTP achieves reasonably high accuracy. It is not possible to compare to children's generalization accuracy because production errors do not transparently reveal a child's grammar. However, compared to NEURAL, ALTERTP's accuracies are higher or comparable, including for Dutch, where it reaches around 0.9. NEURAL achieves similar accuracy at larger training sizes, but over-predicts the voicing alternation (Tab. 2 and Fig. 6b).

While it is difficult to interpret the reasons for NEURAL's behavior due to the opaqueness of neural models, it clearly deviates from ALTERTP and from children in its higher tendency to over-apply the voicing alternation. This was observed in § 4.2.2, where NEURAL erroneously produced *[d] for non-alternating [t] at a higher rate than ALTERTP and children did, and in the wug test in § 4.4.1, where NEURAL applied the alternation in both Dutch and German at higher rates than ALTERTP and children did. Because alternating nouns are exceptions to morphological generalizations that do not incorporate the alternation, *underapplication* of a voicing alternation is a form of *overregularization* (e.g., analagous to the English past tense) whereas *overapplication* is a form of *overirregularization*. Thus, NEURAL's extensive overapplication of the voicing alternation is reminiscent of a tendency to overirregularize. While overregularization is typically much more common than overirregularization in child development, neural network models are consistently prone to excessive overirregularization (Kodner *et al.*, 2023; Payne & Kodner, 2025). The fact that ALTERTP is similar or higher in accuracy while better-matching children's knowledge of the voicing alternating supports its proposal that the alternation is only learned to the extent needed to support morphological generalization.

Figure 9b shows the fraction of training items that are covered (i.e., not lexicalized) by ALTERTP. Dashed lines show what coverage would be if no alternating nouns were learned. German's coverage is lower overall because of the large number of suffixes and complexity of conditioning factors. However, the solid and dashed lines are nearly identical for Dutch but not German: predicting alternating forms is more important for German.

These trees and plots are consistent with the argument that Dutch voicing alternations form a comparatively small corner of the lexicon, which can largely be lexicalized without hindering generalization.

5 Conclusion

This article proposed a theory of the acquisition of Dutch and German noun morphophonology. In the proposal, learners' representations are initially concrete (Richter, 2021) and item-specific, similar to usage-based theories (Bybee, 2010). Learners' first morphological generalizations are over these representations, but resulting rules may break down in the face of alternations. If the alternations are prevalent enough, then I proposed that learners construct new representations for alternating stem-final obstruents, treating them as elements of comparatively abstract categories.

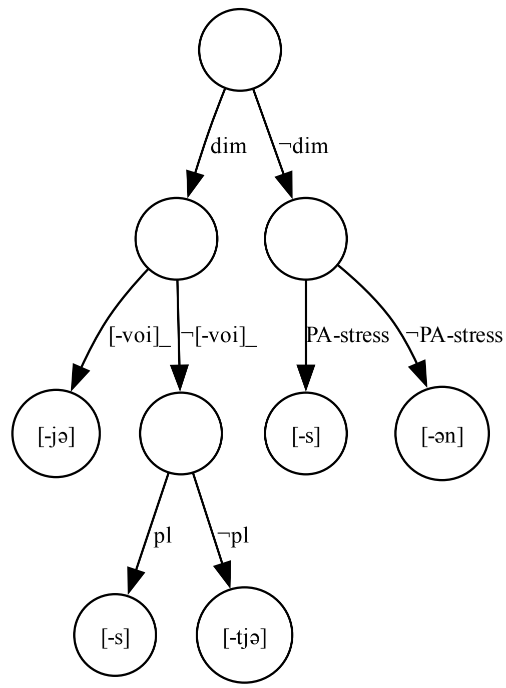


Figure 7: Example Dutch tree learned by ALTERTP.

This allows the alternation to be learned. On the other hand, if alternations are marginal, then this refinement does not occur and the representations remain concrete.

Implemented as a learning model, the proposal provides a possible explanation for the different developmental trajectories of Dutch- and German-learning children. It suggests that Dutch learners encounter the need to construct non-surface-faithful representations less rapidly and in more restricted cases than their German-learning peers, who more rapidly expand their hypothesis space to incorporate the alternation. Thus, the proposal provides a learning-based approach to the Kiparskian question of how abstract phonology is.

The proposal argues that it is the breakdown of morphological generalizations, rather than prior knowledge of phonotactics, that triggers the learner to attend to learning alternations, even when phonotactics motivate those alternations. This indeed appears to be the case: Dutch-learning children show no evidence of phonotactics mediating the acquisition of the alternation. This was a central conclusion of Zamuner *et al.* (2006:p. 495), "we found little evidence that the early acquisition of phonotactics is generalized to the acquisition of morphophonological alternations" and Kerkhoff (2007:p. 258), "even if voicing neutralisation is acquired by Dutch children (reflecting both language specific phonotactics and phonetic naturalness), there is no evidence that this knowledge is subsequently applied to the acquisition of alternations." While some artificial-language-learning studies have found a difference in learning performance between alternations that align with phonotactics vs. those that do not, (Pater & Tessier, 2005; Chong, 2021; Kuo, 2024), learn-

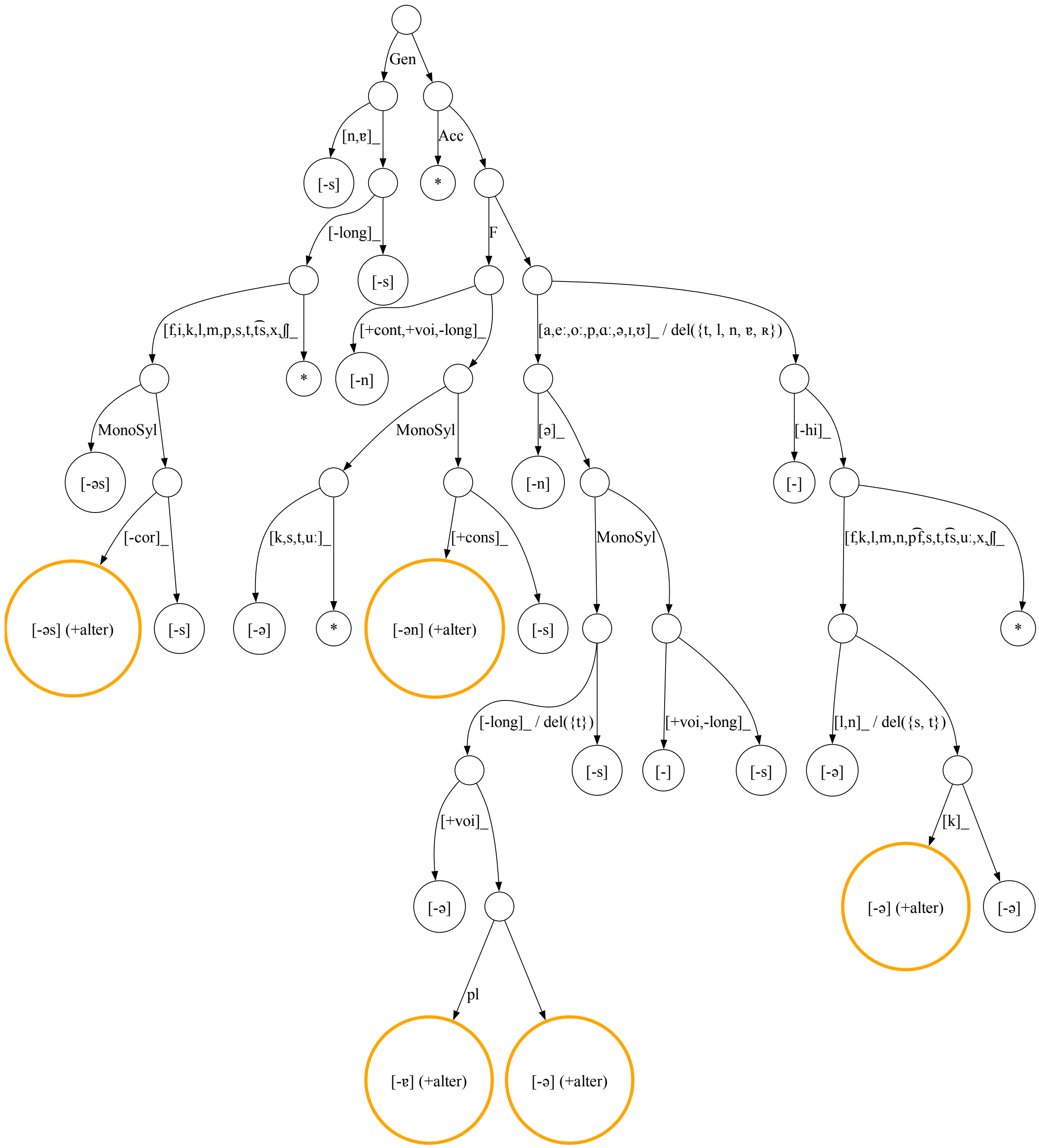
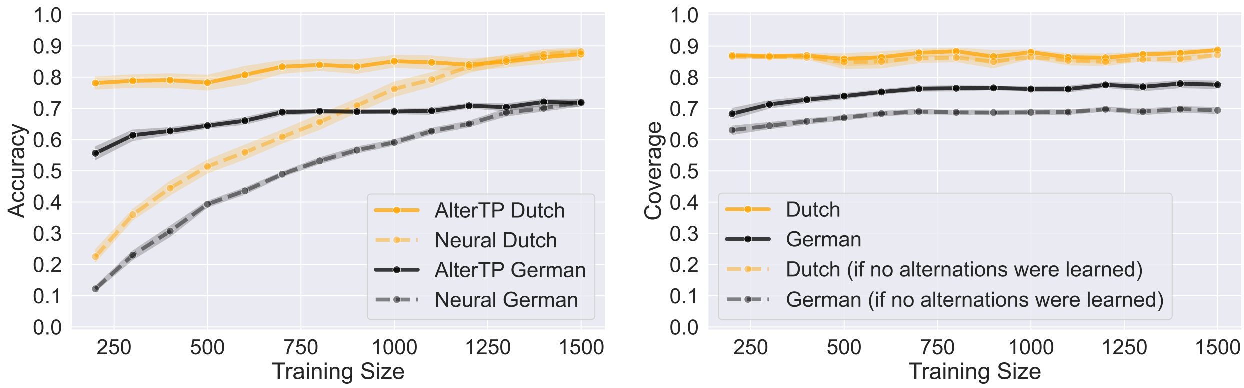


Figure 8: Example German tree learned by AlterTP.

ers in laboratory settings still do not generalize from phonotactic knowledge to alternations when the alternation was not present during training (Chong, 2021). Other studies have not found an effect of phonotactics at all (Do & Yeung, 2021). Beyond the laboratory, the present case of natural language acquisition suggests that—whatever the effect of phonotactics in laboratory settings



(a) Accuracy

(b) Coverage

Figure 9: Accuracy on held-out test words and productive coverage of training items.

may be—it is premature to conclude that phonotactic knowledge is a central mechanism behind alternation learning.

5.1 Approaches to UR Learning

A number of proposals have been made for learning morphophonological grammars in tandem with URs. Many of these models use some combination of maximum entropy learning and expectation maximization (Jarosz, 2006, 2013; Pater *et al.*, 2012; O'Hara, 2017; Nelson, 2019; Wang & Hayes, 2025). Other models use the Minimum Description Length principle (Rasin *et al.*, 2020, 2021) or Bayesian inference (Cotterell *et al.*, 2015; Ellis *et al.*, 2022). Some use constraint-based formalisms (Kager, 1999; Tesar, 2013; Nyman & Tesar, 2019). Hua *et al.* (2020) proposed a model under the assumption that the target grammar falls in the Input Strictly Local class of functions (Chandlee & Heinz, 2018). These models typically specify a hypothesis space of URs and grammars, and either formulate learning as joint inference over the hypothesis spaces or iterate between inference of URs and inference of grammars. They often make the assumption that morphological paradigms are complete, which changes the nature of the learning problem from that indicated by empirical studies of paradigm sparsity (3.1). Jarosz (2011) applied Jarosz (2006)'s learning model to Dutch voicing alternations, training the model on 6 exemplar Dutch nouns and finding that the model matched basic trends in voicing error rates for target [t]'s and [d]'s. The model was not applied to a corpus or to wugs, so it is not clear whether the model would match further developmental facts in Dutch or the asymmetries between Dutch and German. Most of these models solve small problem sets and are not expected to scale unaltered to natural-language-acquisition-scale settings, with Cotterell *et al.* (2015) being a notable exception. However, Cotterell *et al.*'s code is not available for comparison. These models differ from AlterTP because they choose URs based on an optimization problem, whereas AlterTP constructs more abstract representations only when surface-faithful-representations are not sufficient to support generalization. Moreover, these models are sometimes referred to as attempting to solve the *hidden structure* problem (Jarosz, 2019), called such because abstract representations are viewed as presenting a learning challenge. Core to the present proposal is the idea that learners construct non-surface-faithful URs precisely when they *help* generalization succeed, mirroring in learning terms the typical reason that linguists propose them: to capture productive generalizations (Hyman, 2018).

5.2 Future Directions

Ernestus & Baayen (2003) report an overall adult productive generalization rate of 25.4% for verbs (combining numbers from their tables 6-7), which is much higher than the 2.3% rate for nouns (§ 4.4) and suggests that there are important differences between nouns and verbs. Ernestus & Baayen identified a number of properties of verbs (e.g., the identity of the final obstruent and the quality of the preceding vowel) that are predictive of whether they alternate and argued that these factor into wug responses. These features are precisely the sort that AlterTP uses to partition the input into the categories over which it forms generalizations. I plan to extend the proposal to Dutch verbs, with the hypothesis that the alternation will emerge as productive for more subcategories of verbs than nouns.

It is also important to study the acquisition of more alternations, as acquisition research has disproportionately focused on phonotactics. A handful of studies on alternations do suggest that some are productive at relatively young ages. Vowel harmony alternations in Hungarian (MacWhinney, 1978) and Turkish (Altan, 2009), as well as tone sandhi in Mandarin (Tang *et al.*, 2019), show productivity before age 3. These processes are arguably more complex than the voicing alternations discussed here, since they involve non-local dependencies and/or abstract prosodic or autosegmental structure. Thus, we should continue researching the formal and language-internal properties that influence alternation acquisition.

References

Altan, Asli. (2009). Acquisition of vowel harmony in Turkish. *Pages 9–26 of: 35. yıl yazıları*. Multilingual Yabancı Dil Yayınları.

Aslin, Richard N., Saffran, Jenny R., & Newport, Elissa L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological science*, **9**(4), 321–324.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). The CELEX Lexical Database (CD-ROM).

Batsuren, Khuyagbaatar, Goldman, Omer, Khalifa, Salam, Habash, Nizar, Kieraś, Witold, Bella, Gábor, Leonard, Brian, Nicolai, Garrett, Gorman, Kyle, Ate, Yustinus Ghanggo, Ryskina, Maria, Mielke, Sabrina, Budianskaya, Elena, El-Khaissi, Charbel, Pimentel, Tiago, Gasser, Michael, Lane, William Abbott, Raj, Mohit, Coler, Matt, Samame, Jaime Rafael Montoya, Camaiteri, Delio Siticonatzi, Rojas, Esaú Zumaeta, López Francis, Didier, Oncevay, Arturo, López Bautista, Juan, Villegas, Gema Celeste Silva, Hennigen, Lucas Torroba, Ek, Adam, Guriel, David, Dirix, Peter, Bernardy, Jean-Philippe, Scherbakov, Andrey, Bayyr-ool, Aziyana, Anastasopoulos, Antonios, Zariquiey, Roberto, Sheifer, Karina, Ganieva, Sofya, Cruz, Hilaria, Karahóǧa, Ritván, Markantonatou, Stella, Pavlidis, George, Plugaryov, Matvey, Klyachko, Elena, Salehi, Ali, Angulo, Candy, Baxi, Jatayu, Krizhanovsky, Andrew, Krizhanovskaya, Natalia, Salesky, Elizabeth, Vania, Clara, Ivanova, Sardana, White, Jennifer, Maudslay, Rowan Hall, Valvoda, Josef, Zmigrod, Ran, Czarnowska, Paula, Nikkarinen, Irene, Salchak, Aelita, Bhatt, Brijesh, Straughn, Christopher, Liu, Zoey, Washington, Jonathan North, Pinter, Yuval, Ataman, Duygu, Wolinski, Marcin, Suhardijanto, Totok, Yablonskaya, Anna, Stoehr, Niklas, Dolatian, Hossep, Nuriah, Zahroh, Ratan, Shyam, Tyers, Francis M., Ponti, Edoardo M., Aiton, Grant, Arora, Aryaman, Hatcher, Richard J., Kumar, Ritesh, Young, Jeremiah, Rodionova, Daria, Yemelina, Anastasia, Andrushko, Taras, Marchenko, Igor, Mashkovtseva, Polina, Serova, Alexandra, Prud'hommeaux, Emily, Nepomniashchaya, Maria, Giunchiglia, Fausto, Chodroff, Eleanor, Hulden, Mans, Silfverberg, Miikka, McCarthy, Arya D., Yarowsky, David, Cotterell, Ryan, Tsarfaty, Reut, & Vylomova, Ekaterina. (2022). UniMorph 4.0: Universal Morphology. *Pages 840–855 of: Proceedings of the thirteenth language resources and evaluation conference*. European Language Resources Association.

Behrens, Heike. (2006). The input–output relationship in first language acquisition. *Language and cognitive processes*, **21**(1-3), 2–24.

Belth, Caleb. (2023a). A learning-based account of local phonological processes. *Phonology*, **40**, 1–33.

Belth, Caleb. (2023b). Towards a learning-based account of underlying forms: A case study in Turkish. *Pages 332–342 of:* Hunter, Tim, & Prickett, Brandon (eds), *Proceedings of the society for computation in linguistics*. Association for Computational Linguistics.

Belth, Caleb. (2024). A learning-based account of phonological tiers. *LI*, 1–37.

Belth, Caleb. (2025a). A learning based account of Turkish laryngeal alternations. *Proceedings of the annual meetings on phonology*. University of Massachusetts Amherst Libraries.

Belth, Caleb. (2025b). A theory of when and how learners construct tiers: Implications for opaque and transparent vowels. *Society for computation in linguistics*, vol. 8. University of Massachusetts Amherst Libraries.

Belth, Caleb, Payne, Sarah, Beser, Deniz, Kodner, Jordan, & Yang, Charles. (2021). The greedy and recursive search for morphological productivity. *Proceedings of the annual meeting of the cognitive science society*, vol. 43.

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

Björnsdóttir, Sigríður Mjöll. (2021). Productivity and the acquisition of gender. *Journal of child language*, **48**(6), 1209–1234.

Bol, Gerard W. (1995). Implicational scaling in child language acquisition: The order of production of Dutch verb constructions. *Pages 1–13 of: Papers from the Dutch-German colloquium on language acquisition*. Amsterdam University Press.

Booij, Geert. (1999). *The phonology of Dutch*. Oxford University Press.

Bornstein, Marc H, Cote, Linda R, Maital, Sharone, Painter, Kathleen, Park, Sung-Yun, Pascual, Liliana, Pêcheux, Marie-Germaine, Ruel, Josette, Venuti, Paola, & Vyt, Andre. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child development*, **75**(4), 1115–1139.

Buckler, Helen, & Fikkert, Paula. (2016). Dutch and German 3-year-olds' representations of voicing alternations. *Language and speech*, **59**(2), 236–265.

Burness, Phillip, & McMullin, Kevin. (2019). Efficient learning of output tier-based strictly 2-local functions. *Pages 78–90 of:* de Groote, Philippe, Drewes, Frank, & Penn, Gerald (eds), *Proceedings of the 16th meeting on the mathematics of language*. Association for Computational Linguistics.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.

Bybee, Joan. (1995). Regular morphology and the lexicon. *Language and cognitive processes*, **10**(5), 425–455.

Chan, Erwin. (2008). *Structures and distributions in morphology learning*. Ph.D. thesis, University of Pennsylvania.

Chandlee, Jane, & Heinz, Jeffrey. (2018). Strict locality and phonological maps. *LI*, **49**(1), 23–60.

Chomsky, Noam, & Halle, Morris. (1968). *The sound pattern of English*. New York: Harper & Row.

Chong, Adam J. (2021). The effect of phonotactics on alternation learning. *Lg*, **97**(2), 213–244.

Clahsen, Harald, Rothweiler, Monika, Woest, Andreas, & Marcus, Gary F. (1992). Regular and irregular inflection in the acquisition of German noun plurals. *Cognition*, **45**(3), 225–255.

Cotterell, Ryan, Peng, Nanyun, & Eisner, Jason. (2015). Modeling word forms using latent underlying morphs and phonology. *Transactions of the association for computational linguistics*, **3**, 433–447.

Dittmar, Miriam, Abbot-Smith, Kirsten, Lieven, Elena, & Tomasello, Michael. (2008). German children's comprehension of word order and case marking in causative sentences. *Child development*, **79**(4), 1152–1167.

Do, Youngah, & Yeung, Ping Hei. (2021). Evidence against a link between learning phonotactics and learning phonological alternations. *Linguistics vanguard*, **7**(1), 20200127.

Eisenbeiss, Sonja, Bartke, Susanne, & Clahsen, Harald. (2006). Structural and lexical case in child German: Evidence from language-impaired and typically developing children. *Language acquisition*, **13**(1), 3–32.

Ellis, Kevin, Albright, Adam, Solar-Lezama, Armando, Tenenbaum, Joshua B, & O'Donnell, Timothy J. (2022). Synthesizing theories of human language with bayesian program induction. *Nature communications*, **13**(1), 5024.

Elsen, Hilke. (2002). The acquisition of German plurals. *Page 117 of: Morphology 2000: Selected papers from the 9th morphology meeting*, vol. 218. John Benjamins Publishing.

Ernestus, Mirjam Theresia Constantia, & Baayen, R Harald. (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Lg*, **79**(1), 5–38.

Fikkert, Paula. (1994). *On the acquisition of prosodic structure*. Ph.D. thesis, Leiden University.

Fiser, József, & Aslin, Richard N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of experimental psychology: Learning, memory, and cognition*, **28**(3), 458–467.

Goldman, Omer, Batsuren, Khuyagbaatar, Khalifa, Salam, Arora, Aryaman, Nicolai, Garrett, Tsarfaty, Reut, & Vylomova, Ekaterina. (2023). SIGMORPHON–UniMorph 2023 shared task 0: Typologically diverse morphological inflection. *Pages 117–125 of: Proceedings of the 20th sigmorphon workshop on computational research in phonetics, phonology, and morphology*. Association for Computational Linguistics.

Gómez, Rebecca. (2002). Variability and detection of invariant structure. *Psychological science*, **13**(5), 431–436.

Gómez, Rebecca, & Maye, Jessica. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy*, **7**(2), 183–206.

Gorman, Kyle, Wiemerslage, Adam, Bartley, Travis, Ginn, Michael, & Ginn, Michael. (2024). Yoyodyne: Small-vocabulary sequence-to-sequence generation with optional feature conditioning.

Gouskova, Maria, Newlin-Łukowicz, Luiza, & Kasyanenko, Sofya. (2015). Selectional restrictions as phonotactics over sublexicons. *Lingua*, **167**, 41–81.

Hayes, Bruce. (2004). Phonological acquisition in optimality theory: The early stages. *Constraints in phonological acquisition*.

Hayes, Bruce, & Wilson, Colin. (2008). A Maximum Entropy Model of Phonotactics and Phonotactic Learning. *LI*, **39**(3), 379–440.

Henke, Ryan E. (2023). Rules and exceptions: A tolerance principle account of the possessive suffix in northern east cree. *Journal of child language*, **50**(5), 1119–1154.

Hua, Wenyue, Jardine, Adam, & Dai, Huteng. (2020). Learning underlying representations and Input-Strictly-Local functions. *WCCFL*.

Hyman, Larry M. (2018). Why underlying representations? *JL*, **54**(3), 591–610.

Itô, Junko, & Mester, Armin. (1999). The phonological lexicon. *The handbook of Japanese Linguistics*, 62–100.

Jardine, Adam, & Heinz, Jeffrey. (2016). Learning Tier-based Strictly 2-Local Languages. *Transactions of the association for computational linguistics*, **4**(04), 87–98.

Jarosz, Gaja. (2006). *Rich lexicons and restrictive grammars: maximum likelihood learning in Optimality Theory*. Ph.D. thesis, Johns Hopkins University.

Jarosz, Gaja. (2011). The roles of phonotactics and frequency in the learning of alternations. *Pages 321–333 of: Bucld 35: Proceedings of the 35th annual boston university conference on language development*.

Jarosz, Gaja. (2013). Learning with hidden structure in optimality theory and harmonic grammar: Beyond robust interpretive parsing. *Phonology*, **30**(1), 27–71.

Jarosz, Gaja. (2019). Computational modeling of phonological learning. *Annual review of linguistics*, **5**(1), 67–90.

Jusczyk, Peter W, Luce, Paul A, & Charles-Luce, Jan. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of memory and language*, **33**(5), 630–645.

Kager, René. (1999). *Optimality theory*. Cambridge University Press.

Kager, René, Van der Feest, Suzanne, Fikkert, Paula, Kerkhoff, Annemarie, & Zamuner, Tania S. (2007). *Representations of [voice]: Evidence from acquisition*. John Benjamins Publishing Company. Pages 41–80.

Kerkhoff, Annemarie Odilia. (2007). *Acquisition of morpho-phonology: The Dutch voicing alternation*. Ph.D. thesis, LOT.

Kim, Yun Jung, & Sundara, Megha. (2021). 6–month–olds are sensitive to English morphology. *Developmental science*, **24**(4), e13089.

Kiparsky, Paul. (1968). *How abstract is phonology?* Bloomington, IN: Indiana University Linguistics Club.

Kodner, Jordan. (2020). *Language acquisition in the past*. Ph.D. thesis, University of Pennsylvania.

Kodner, Jordan, & Khalifa, Salam. (2022). SIGMORPHON–UniMorph 2022 shared task 0: Modeling inflection in language acquisition. *Pages 157–175 of:* Nicolai, Garrett, & Chodroff, Eleanor (eds), *Proceedings of the 19th sigmorphon workshop on computational research in phonetics, phonology, and morphology*. Association for Computational Linguistics.

Kodner, Jordan, Khalifa, Salam, Payne, Sarah RB, & Liu, Zoey. (2023). Re-evaluating the evaluation of neural morphological inflection models. *Proceedings of the annual meeting of the cognitive science society*, vol. 45.

Koulaguina, Elena, & Shi, Rushen. (2019). Rule generalization from inconsistent input in early infancy. *Language acquisition*, **26**(4), 416–435.

Kuijpers, Cecile. (1993). *Temporal coordination in speech development: a study on voicing contrast and assimilation of voice*. Ph.D. thesis, Universiteit van Amsterdam.

Kuo, Jennifer. (2024). Types of statistical knowledge in alternation learning: insights from artificial grammar learning. *WCCFL*.

Ladányi, Enikő, Kovács, Ágnes M, & Gervain, Judit. (2020). How 15-month-old infants process morphologically complex forms in an agglutinative language? *Infancy*, **25**(2), 190–204.

Levelt, Clara. (1994). *On the acquisition of place*. Ph.D. thesis, Rijksuniversiteit Leiden.

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

MacWhinney, Brian. (2014). *The CHILDES project: Tools for analyzing talk, volume ii: The database*. London: Psychology Press.

Marcus, Gary F, Brinkmann, Ursula, Clahsen, Harald, Wiese, Richard, & Pinker, Steven. (1995). German inflection: The exception that proves the rule. *Cognitive psychology*, **29**(3), 189–256.

Marquis, Alexandra, & Shi, Rushen. (2012). Initial morphological learning in preverbal infants. *Cognition*, **122**(1), 61–66.

McCurdy, Kate, Goldwater, Sharon, & Lopez, Adam. (2020). Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for German plurals. *Pages 1745–1756 of:* Jurafsky, Dan, Chai, Joyce, Schluter, Natalie, & Tetreault, Joel (eds), *Proceedings of the 58th annual meeting of the association for computational linguistics*. Association for Computational Linguistics.

Medin, Douglas L, Wattenmaker, William D, & Hampson, Sarah E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive psychology*, **19**(2), 242–279.

Mills, Anne. (1986). *The acquisition of gender: A study of English and German*. Berlin: Springer.

Mortensen, David R., Dalmia, Siddharth, & Littell, Patrick. (2018). Epitran: Precision G2P for many languages. *Proceedings of the eleventh international conference on language resources and evaluation*. Miyazaki, Japan: European Language Resources Association.

Nelson, Deborah G Kemler. (1984). The effect of intention on what concepts are acquired. *Journal of verbal learning and verbal behavior*, **23**(6), 734–759.

Nelson, Max. (2019). Segmentation and UR acquisition with UR constraints. *Society for computation in linguistics*, vol. 2. University of Massachusetts Amherst Libraries.

Nyman, Alexandra, & Tesar, Bruce. (2019). Determining underlying presence in the learning of grammars that allow insertion and deletion. *Glossa*, **4**(1), 1–41.

Odijk, Jan. (2015). Linguistic research with PaQu. *Computational linguistics in the Netherlands journal*, **5**, 3–14.

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

Os, E den, & Harder, R. (1987). De Verwerving van de Regels voor Meervouds-en Verkleinwoordsvorming in het Nederlands L'acquisition des règles de formation du pluriel et des diminutifs en néerlandais. *Nieuwe taalgids*, **80**(3), 240–250.

Pater, Joe, & Tessier, Anne-Michelle. (2005). Phonotactics and alternations: Testing the connection with artificial language learning. *University of massachusetts occasional papers in linguisitcs*, **31**, 1–16.

Pater, Joe, Staubs, Robert, Jesney, Karen, & Smith, Brian Cantwell. (2012). Learning probabilities over underlying representations. *Pages 62–71 of: Proceedings of the twelfth meeting of the special interest group on computational morphology and phonology*.

Payne, Sarah, & Kodner, Jordan. (2025). Some innate characteristics of neural network models of morphological inflection. *Linguistics vanguard*.

Pearl, Lisa, & Sprouse, Jon. (2021). The acquisition of linking theories: A tolerance and sufficiency principle approach to deriving utah and rutah. *Language acquisition*, **28**(3), 294–325.

Pounder, Amanda. (1996). Inflection and the paradigm in German nouns. *Journal of Germanic linguistics*, **8**(2), 219–263.

Prince, Alan, & Smolensky, Paul. (1993). *Optimality Theory: Constraint interaction in generative grammar*. Tech. rept.

Rasin, Ezer, Shefi, Itamar, & Katzir, Roni. (2020). A unified approach to several learning challenges in phonology. *Pages 73–87 of: NELS*, vol. 50.

Rasin, Ezer, Berger, Iddo, Lan, Nur, Shefi, Itamar, & Katzir, Roni. (2021). Approaching explanatory adequacy in phonology using Minimum Description Length. *Journal of language modelling*, **9**(1), 17–66.

Richter, Caitlin. (2021). *Alternation-sensitive phoneme learning: Implications for children's development and language change*. Ph.D. thesis, University of Pennsylvania.

Rumelhart, David E, & McClelland, James L. (1986). On learning the past tenses of English verbs. *Psycholinguistics: Critical concepts in psychology*, **4**, 216–271.

Saffran, Jenny R, Aslin, Richard N, & Newport, Elissa L. (1996). Statistical learning by 8-month-old infants. *Science*, **274**(5294), 1926–1928.

Saffran, Jenny R., Newport, Elissa L., Aslin, Richard N., Tunick, Rachel A., & Barrueco, Sandra. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological science*, **8**(2), 101–105.

Saffran, Jenny R, Johnson, Elizabeth K, Aslin, Richard N, & Newport, Elissa L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, **70**(1), 27–52.

Santelmann, Lynn, & Jusczyk, Peter W. (1998). Sensitivity to discontinuous dependencies in language learners: Evidence for limitations in processing space. *Cognition*, **69**(2), 105–134.

Schuler, Kathryn, Yang, Charles, & Newport, Elissa. (2016). Testing the Tolerance Principle: Children form productive rules when it is more computationally efficient to do so. *Proceedings of the 38th annual meeting of the Cognitive Science Society*.

Shi, Rushen, & Emond, Emeryse. (2023). The threshold of rule productivity in infants. *Frontiers in psychology*, **14**, 1251124.

Slis, Iman H, & van Heugten, Marieke. (1989). Voiced-voiceless distinction in Dutch fricatives. *Linguistics in the Netherlands*, **6**, 123–132.

Smith, Linda B. (1979). Perceptual development and category generalization. *Child development*, 705–715.

Sundara, Megha, Zhou, ZL, Breiss, Canaan, Katsuda, Hironori, & Steffman, Jeremy. (2022). Infants' developing sensitivity to native language phonotactics: A meta-analysis. *Cognition*, **221**, 104993.

Szagun, Gisela, Steinbrink, Claudia, Franik, Melanie, & Stumper, Barbara. (2006). Development of vocabulary and grammar in young German-speaking children assessed with a German language development inventory. *First language*, **26**(3), 259–280.

Tang, Ping, Yuen, Ivan, Rattanasone, Nan Xu, Gao, Liqun, & Demuth, Katherine. (2019). The acquisition of phonological alternations: The case of the Mandarin tone sandhi process. *Applied psycholinguistics*, **40**(6), 1495–1526.

Tesar, Bruce. (2013). *Output-driven phonology: Theory and learning*. Cambridge Studies in Linguistics. Cambridge: Cambridge University Press.

Tracy, Rosemarie. (1986). The acquisition of case morphology in German. *Linguistics*, **24**, 47–78.

van de Vijver, Ruben, & Baer-Henney, Dinah. (2014). Developing biases. *Frontiers in psychology*, **5**, 634.

van der Feest, Suzanne. (2007). *Building a phonological lexicon. the acquisition of the Dutch voicing contrast in perception and production*. Ph.D. thesis, University of Nijmegen.

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

Van Tuijl, Rosita, & Coopmans, Peter. (2021). The productivity of Dutch diminutives. *Linguistics in the Netherlands*, **38**(1), 128–143.

Wang, Yang, & Hayes, Bruce. (2025). Learning phonological underlying representations: the role of abstractness. *LI*, 1–44.

Wiese, Richard. (1996). *The phonology of German*. Oxford: Clarendon.

Wu, Shijie, Cotterell, Ryan, & Hulden, Mans. (2021). Applying the transformer to character-level transduction. *Pages 1901–1907 of: Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume*. Association for Computational Linguistics.

Yang, Charles. (2016). *The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language*. Cambridge, MA: The MIT Press.

Yin, Yue. (2025). Modeling the learning of tier projection with blocking. *Program for the annual meetings on phonology*.

Zamuner, Tania S, Kerkhoff, Annemarie, & Fikkert, Paula. (2006). Acquisition of voicing neutralization and alternations in Dutch. *Pages 701–712 of: Proceedings of the 30th annual Boston University Conference on Language Development*.

Zamuner, Tania S, Kerkhoff, Annemarie, & Fikkert, Paula. (2012). Phonotactics and morphophonology in early child language: Evidence from Dutch. *Applied psycholinguistics*, **33**(3), 481–499.

Zaretsky, Eugen, & Lange, Benjamin P. (2015). No matter how hard we try: Still no default plural marker in nonce nouns in Modern High German. *Pages 153–178 of: A blend of malt: Selected contributions from the methods and linguistic theories symposium*.

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

A Constructing FormFeats

AlterTP creates an initial FormFeat containing any src-final-segment where having it as the final segment is a sufficient predictor of being in T. If the resulting FormFeat is necessary and sufficient for being in T, then it is added to the set F of features under consideration. Otherwise, the set of endings that were *not* sufficient predictors is used to form new tier representations for each src by excluding any segment in the src that is not a sufficient predictor of the tgt. The process then repeats based on the final segments of the new representation. Iteration terminates either when the resulting set of endings is a necessary and sufficient predictor or when enough segments are excluded from the representation that nothing is projected. In our example, all T triples end in $\{\mathrm{k},\ \mathrm{t},\ \mathrm{m},\ \mathrm{s}\}$ and none of the (12c) triples do, so this set of endings is necessary and sufficient for T.

If the above procedure fails to find a necessary *and sufficient* FormFeat, then a FormFeat is instead formed by adding any segment that is the final segment of more src forms in T than src forms not in T. If this feature is necessary, it is added to F.

When a FormFeat is added to F, AlterTP first tries to form a natural class for the endings in the feature by the same mechanism used in Belth (2023a) and subsequent work: distinctive features shared by all segments in the set are incrementally added to an initially-empty natural class in the order of the number of triples not in T they remove, until no further removal is possible. If the segments do not form a natural class or the natural class breaks sufficiency, the set of segments is used extensionally. In our example, the endings $\{\mathrm{k},\ \mathrm{t},\ \mathrm{m},\ \mathrm{s}\}$ are all non-laterals, and adding

[−lat] removes one (12c) triple, which ends in lateral [l]. Because [m] and [r] are both sonorants, no further features are added. However, the feature is still necessary and sufficient ($3 \leq \theta_{10} = 4.3$) and is thus added to F.

B Dataset Creation

I constructed a Dutch dataset from the Van Kampen (Van Kampen, 2009) and Groningen (Bol, 1995) CHILDES corpuses (MacWhinney, 2014), which together contain child-directed speech (CDS) for 9 Dutch-learning children, aged 1;05-6;0. I extracted all CDS words annotated as nouns on the %mor tier and computed each noun's token frequency across the two corpuses. The %mor tier marks the plurality of nouns, but not whether they are diminutives. However, the %mor tier lists lemmas (e.g., *koek* for *koekje* "cookie-Dim"), albeit with some inconsistency for diminutives. To improve the quality of the data, I followed Van Tuijl & Coopmans (2021) in extracting lemmas from the online platform PaQu (Odijk, 2015), and then used the features and lemmas from CHILDES and PaQu to infer the features Pl and Dim. Any inconsistencies between CHILDES and PaQu were discarded (2.7% of nouns). I then organized nouns by paradigm (i.e., grouping nouns with the same root) and dropped any nouns for which only a single paradigm form was present (overwhelmingly low-frequency words), since no generalization is possible from such words. From CELEX (Baayen *et al.*, 1996) I extracted IPA transcriptions and prosodic features (the placement of stress in the lemma and whether the lemma is monosyllabic). I normalized transcriptions between the lemma and inflected forms, except for the voicing alternation, since other stem changes (e.g., vowel length differences) are orthogonal to the present study. This yielded a dataset of 2727 unique nouns, falling into 1051 paradigms. These were organized into (src, feats, tgt) triples by taking, from each paradigm, every pair of words where the second word is formed by marking one feature beyond the first word (i.e., lemma → Pl, lemma → Dim, and Dim → Pl Dim). Each triple was assigned a weight equal to the probability of both the src and tgt being sampled in a frequency-weighted sample. Due to incomplete paradigms, this produced 1562 triples.

For German, I extracted all CDS nouns from the Leo (Behrens, 2006) CHILDES corpus. I retrieved noun lemmas and grammatical gender from the %mor tier, resolving any inconsistencies of gender by choosing the most-frequently-annotated gender for each lemma. I computed the token frequency of each noun lemma. To improve the data quality, I intersected the resulting lemmas with Unimorph 4.0 (Batsuren *et al.*, 2022) and, for each matching lemma, looked up the inflected form of each case (Nom, Acc, Gen, Dat) in both Sg and Pl. I extracted the Nom Sg and Pl, together with any non-Nom Sg forms that differed from the Nom;Sg and any Pl forms that differed from the Nom;Pl form. I then computed the frequency of each inflectional category in Leo and distributed the lemma frequencies proportionately across the forms extracted from Unimorph. Lastly, I converted all nouns to IPA using Epitran (Mortensen *et al.*, 2018) and, like Dutch, normalized transcriptions between the lemma and inflected forms (except for the voicing alternation) and annotated nouns with the same prosodic features as for Dutch (except stress placement, which was unavailable). This resulted in a dataset of 14340 unique noun forms belonging to 4195 paradigms. These were organized into (SRC, FEATS, TGT) triples in the same way as the Dutch dataset, producing 9390 triples.

C NEURAL Hyperparameter Tuning

For the NEURAL comparison model, I used the Yoyodyne implementation from Gorman *et al.* (2024). I tuned hyperparameters (embedding size $\in \{256, 512\}$, encoder/decoder layers $\in \{1, 2\}$, attention heads $\in \{4, 8\}$, and learning rate $\in \{0.001, 0.0001\}$) for each evaluation size and seed by forming an 80/20% train/validation split of the training triples, and then performing a grid search over the hyperparameter options by training on the 80% and evaluating generalization on the 20%. The hyperparameter combination that produced the best validation accuracy was then used to train a model on the full, recombined 80% and 20%.