Towards a Learning-Theoretic morphophonology

Vsevolod Kapatsinski

University of Oregon

Abstract

Work on acquisition of phonology has long been concerned with the nature of the mechanism that could accomplish the task. However, rather unfortunately, this work has developed in isolation from work on general learning theory. This paper argues for an integration, suggesting that basic associative learning mechanisms can provide a solid foundation on which to build a theory of phonological acquisition. While language is unique to humans, I argue that language learning is likely to be done largely using pre-existing learning mechanisms. Unlike aspects of the natural environment that force their statistics onto the organisms exposed to them, language adapts to the learning biases of its learners, allowing learners to impose their learning biases upon it. If a language is hard to learn, it is likely to change to fit the learner. This malleability of language structure makes domain-general learning mechanisms a good foundation on which to build a theory of language learning, and conversely makes language a valuable window on domain-general biases of the human learner.

Expressing the tasks of phonological learning in the common language of associationism allows for cross-domain comparisons, as well as easier integration with relevant evidence from neuroscience. At the same time, it allows for flow of information in the other direction, letting evidence from phonological learning bear on learning theory. This productive flow of information is illustrated in the present paper by applying a computational implementation of Rescorla & Wagner's (1972) associative learning theory to recalcitrant data on laboratory learning of phonology/morphology interactions. I show that the model achieves better data coverage than previous models developed within phonology, capturing both 'source-oriented' and 'product-oriented' effects. Importantly, the model is simpler than its domain-internal competitors, doing away with several ancillary mechanisms and assumptions. It also provides plausible loci for the effects having to do with the neural substrate of phonological learning: 'performance' effects of trial order, individual differences among learners, and substance-based differences among to-be-associated representations. Conversely, phonological data are shown to constrain learning theory by forcing specific representational assumptions onto the model as well as providing evidence for learning mechanisms beyond simple error-driven learning.

Introduction

The term 'learning-theoretic phonology' was introduced by Hayes & Wilson (2008) to mean "a theory whose overall architecture recapitulates the incremental process through which phonological knowledge is acquired." Phonologists have long been concerned with the acquisition of phonological knowledge. Chomsky & Halle (1965) have conceived of phonological theory as specifying the acquisition mechanism (AM) that language learners use to infer the body of knowledge that constitutes their phonological grammar from experience with language. Accounting for how phonological knowledge is acquired continues to be a primary concern of phonological theory of all stripes (e.g. rule-based phonology in Albright & Hayes 2003 and Hale & Reiss 1998; Optimality Theory in Boersma & Hayes 2001, Hayes 2004, Tesar & Smolensky 1998a, 1998b; Harmonic Grammar in Hayes & Wilson 2008; and functionalist approaches to phonology in Boersma 1998, Bybee 2001, Kapatsinski 2013, Redford 2015, Vihman & Croft 2007).

There is therefore widespread agreement in the field that a learning-theoretic phonology is desirable. However, somewhat surprisingly, there has been little work explicitly attempting to apply Learning Theory to the acquisition of phonology (though see Moreton et al. 2015 for a related effort). While Hayes & Wilson (2008) argue for a learning-theoretic phonology, no work on Learning Theory is cited in justifying the particular approach. In part, this situation may be an unfortunate side effect of Chomsky's (1959) influential review of Skinner's (1957) *Verbal behavior*, a book by one of the major early figures in Learning Theory. For a long time after Chomsky (1959), linguistic theory in general and phonological theory in particular have been concerned with designing a domain-specific learning mechanism driven solely by domain-internal considerations (Tomasello 2003, Yang 2004).

In part, lack of engagement with Learning Theory may also reflect the confrontational interaction between subsymbolic distributed connectionist models and proponents of the symbolic approaches to mental representation (Rumelhart & McClelland 1986 vs. Lachter & Bever 1988, Pinker & Prince 1988). Connectionist models can be seen as implementations of associationist learning theory (e.g. Gluck & Bower 1988). However, connectionist models and learning theory do not inherently entail a specific position on the representational question (e.g. Holyoak 1991, Smolensky 1988).

Recently the theoretical landscape has changed. The literature on generative phonology is now dominated by constraint-based approaches with their roots in connectionist modeling (Boersma 1998, Prince & Smolensky 2004, Smolensky 1988, Smolensky & Legendre 2006), which is also embraced by functionalist approaches to language (Bates & MacWhinney 1989, Bybee & McClelland 2005). Learning theory has also made major inroads into the study of first-language (Arnon & Ramscar 2012, Ramscar et al. 2010, 2013, 2014) and second-language acquisition (Ellis 2006, Ellis & Sagarra 2011) outside of phonology. Several researchers have also identified learning-theoretic explanations for psycholinguistic effects previously attributed to organization of the mental lexicon (e.g. Baayen et al. 2011, Kapatsinski 2007, Oppenheim et al. 2010). Association measures from learning theory have also been cropping up in corpus linguistics, where they have become a promising way to identify collocations (Gries 2013, Wahl 2015).

The historical development of phonological theory in the last twenty five years has seen early adoption of domain-general learning mechanisms grounded in connectionist modeling (the weighted constraints of Harmonic Grammar, Legendre et al. 1990) followed by a return to domain-specific learning with constraint reranking mechanisms of Optimality Theory (Prince & Smolensky 1993/2004, Tesar & Smolensky 1998a, 1988b) and a return to domain-general learning mechanisms with maximum-entropy models (Goldwater & Johnson 2003, Hayes & Wilson 2008, Moreton 2008b). While the pendulum may yet swing back, I believe that it is now approaching the (eventual) equilibrium point, and we will eventually settle on the belief that phonology learning is not so different from learning in other domains.

Note that by 'a learning mechanism', I do not mean a localizable brain module that does all kinds of learning (cf. Moreton's 2008a 'analysis module'). Rather, I mean the set of principles that governs modification of synaptic strength (which is a mechanism for learning active throughout the brain) and that allows the organism to pick up on the statistics of the environment; in our case, the linguistic environment. Language is somewhat unique in the world because it is one case in which the statistics of the environment are shaped largely by the biases of the learner.[1] As pointed out by Deacon (1997), linguistic change is much more rapid than genetic change. Thus, it is likely that, to a large extent, it is language structure that has adapted to the learning abilities of humans rather than vice versa (see also

[1] In iterated learning experiments where what is learned by generation N serves as input to generation N+1, learning has been shown to converge on the prior (Griffiths et al. 2008).

Tomasello 2003). This makes it quite plausible that learning theory, based as it is on findings in other domains and often with other animals, can contribute to the understanding of language learning.

In the present paper, I apply the learning theory of Rescorla & Wagner (1972), as implemented by Arppe et al. (2014), to the task of learning morphology and morphophonology. This is a good place to start because, unlike syntax, morphology provides the challenge of accounting for paradigmatic structure. Both morphology and syntax have been argued to be describable as a structured inventory of *constructions* or *schemas*, form-meaning pairings (Bybee 1985, 2001, Goldberg 2002, 2006), and one can also argue for the presence of this kind of *schematic* structure below the level of morphology, in phonaesthemes like the *gl-* in *glow* and *glisten* and sound symbolism (e.g. Bergen 2004). The existence of meaningful units like *gl-* that do not combine with other units in a concatenative manner is prima facie evidence for connections based on simple form-meaning co-occurrence (see also Baayen et al. 2011, Bergen 2004). Likewise, all of language, above and below morphology, is permeated with *syntagmatic* structure, which specified what can (or usually) does follow what in time (e.g. Misyak et al. 2010); something that even infants are capable of tracking (Aslin et al. 1998, Saffran et al. 1996). However, morphology poses the additional challenge of accounting for *paradigmatic* structure. In an associationist framework, we can say that paradigmatic structure involves connections between forms that are based on correspondence rather than contiguity.

For example, a Russian speaker learns that the Genitive Plural of an *–a*-final noun like *sobaka* 'dog', loses the *–a* suffix (*sobaka→sobak*); and that the Genitive Plural of a consonant-final noun like *kabak* 'pub' gains an *–ov* suffix (*kabak→kabakov*). When presented with a new noun ending in –a, and asked to fit it into a context demanding the Genitive Plural form (e.g. *of many* _____), a Russian speaker will delete the *–a*. When presented with a new noun ending in a consonant, they will add *-ov*. The Russian speaker thus knows not only that Genitive Plural forms should end in *–ov* or a consonant but also that the forms depend on what forms the same noun has in other contexts. The Russian speaker knows not only how Genitive Plural is expressed but also that the way you express Genitive Plural for any given noun depends on how you express Nominative Singular for the same noun. This kind of knowledge demands paradigmatic connections (here, between a word-final consonant C# in the Nominative Singular form and a word-final /a/ in the Genitive Plural form).

As argued by proponents of non-derivational approaches to syntax, there is no clear evidence that knowledge of syntax requires knowledge of paradigmatic mappings. Learners of syntax need not learn how to derive syntactic constructions from other, related constructions. For example, while the prepositional dative exemplified by *He gave the book to him* is clearly related to the double object dative exemplified by *He gave him a book*, there is no reason to assume that producing a prepositional dative involves first activating a double object dative paraphrase and then deriving the propositional dative from it, or vice versa (e.g. Goldberg 2002). While it is possible for a writer or speaker to take a prepositional dative sentence and reshape it into a double object sentence or vice versa, this kind of reshaping need not involve paradigmatic connections. Instead, the constituents of the original sentence like *he*.AGENT, *gave*.VERB, *the book*.PATIENT, and *him.*RECIPIENT can simply be slotted into the target construction. One could argue that paradigmatic structure is needed in this conversion process when a form is replaced by another form. For example, the speaker may transform *He gave the book to the fat cat on the windowsill* into *He gave it the book* (because the double object construction demands a short direct object). However, it is not clear that this kind of conversion demands paradigmatic connections between forms: the produced form could depend merely on the meaning to be expressed and the syntagmatic restrictions on the output imposed by the construction.

Paradigmatic structure is hard to learn (Brooks et al. 1993, Frigo & MacDonald 1998), and is often learned imperfectly even by native speakers (Dabrowska 2008). As a result, morphological constructions are often only semi-productive, applying to some novel words but not others, making morphological structure is a good example of *quasi*-regularity (Plaut et al. 1996, Taft & Meunier 1998): not fully describable by regular deterministic rules but demanding some mechanism beyond rote storage of experienced forms to deal with novel inputs in a way that is systematically sensitive to their characteristics. This has made it an important testing ground for competing theories of learning and representation (e.g. Albright & Hayes 2003, Kruschke 1992, van Noord & Spenader 2015, Pinker & Prince 1988, Rumelhart & McClelland 1986, Westermann & Ruh 2012).

The present paper approaches the task of learning morphology armed with the error-driven learning theory of Rescorla & Wagner (1972); henceforth *RW*. While RW is often called the 'standard theory' of learning (e.g. Dickinson 2001, Rodríguez & Alonso 2004), it does have limitations, some of them well-known. In particular, it lacks a mechanism for rapidly allocating selective attention to highly informative stimulus dimensions (Kruschke 1992, 2001; though cf. Ghirlanda 2005) and a mechanism for fusing or splitting units to generate the dimensions that would afford optimal performance in a classification task (Gluck & Bower 1988, Goldstone 2000, 2003). As standardly applied, it also does not learn to 'retrodict' effects when they are perceived before their causes, an arguably routine occurrence in human learning (Harmon & Kapatsinski In prep, Matute et al. 1996, Waldmann & Holyoak 1992, but cf. Ramscar et al. 2010). However, these are not unsolvable problems, and indeed the models of Ghirlanda (2005), Gluck & Bower (1988), Goldstone (2003), Kruschke (1992), and Westermann & Ruh (2012) can be seen as bona fide developments of the RW framework to address the issues identified. The argument here is therefore not that RW is sufficient for learning language but rather that it provides a good foundation on which to build a learning theory that would be sufficient for this task.

Paradigms and schemas

Bybee (2001) contrasts product-oriented generalizations and source-oriented generalizations. Product-oriented generalizations made by generalizing over forms with a certain meaning and are also known as constructions (Goldberg 2003) and first-order schemas (Nesset 2008). They are also much alike to the phonological templates of Vihman & Croft (2007). Source-oriented generalizations are made by generalizing over pairs of morphologically related words and specify how one form can be derived from the other, capturing paradigmatic structure. Phonological rules are a paradigm example of a source-oriented generalization.

While the need for source-oriented generalizations has been questioned (Bybee 2001), recent work has provided compelling demonstrations that they can be productive in natural languages (Gouskova & Becker 2013, Pierrehumbert 2006). For example, Pierrehumbert (2006) shows that generalization over nouns ending in *–ity* would result in N=…lɪti# being a stronger schema than N=…sɪti#: there are more *–lity*-final nouns like *sterility* than –[s]*ity* final ones like *diversity* or *telicity*. Nonetheless, English speakers turn novel [k]-final adjective like *interponic* into nouns ending in [sɪti] and not [lɪti], i.e. *interponicity* and not *interponility*. Gouskova & Becker (2013) show that Russian speaker are more likely to delete the [o] in a …CCoC noun than in a …CoCC noun in deriving a genitive, despite both deletions obeying the same potential product-oriented schema GENITIVE=…CCCa#. Furthermore, as in Pierrehumbert's (2006) data, it is unclear that one would learn that Genitives should end in CCCa by generalizing only over genitive forms: CCC clusters are not particularly common, even in Genitives.

Nesset (2008) and Kapatsinski (2013) note that source-oriented generalizations appear to be hard to learn, as evidenced by multiple laboratory studies attempting to teach arbitrary paradigmatic mappings (Brooks

et al. 1993, Frigo & McDonald 1998). They propose that these generalizations are second-order schemas, formed by associating together product-oriented first-order schemas. However, this hypothesis may not be sustainable: it is not necessarily the case that the units linked by a source-oriented generalization are overrepresented within their respective paradigm cells. For example, CCCa# does not have a particularly high type frequency in Russian genitives compared to other genitive-final tetragrams, making it unlikely it would emerge as a schema if one generalizes only over genitives. It appears possible to learn a paradigmatic association between relatively infrequent structures if they are highly predictive of each other (see also Albright 2008).

Therefore I assume that language learners can acquire arbitrary paradigmatic mappings, and that they learn them in parallel with learning product-oriented schemas or constructions. However, I retain the intuition that paradigmatic mappings are difficult to learn because of the challenges they pose to working memory (Kapatsinski 2013). It appears unreasonable to suppose that, say, whenever a plural form is encountered the learner can successfully bring to mind the corresponding singular form in order to associate aspects of the two forms together. If learners had this kind of perfect ability (and unyielding inclination) to recall morphological relatives of experienced forms and attempt to derive the experienced form from them, then we would expect the learning of arbitrary paradigms to be much quicker than it appears to be. I therefore explore the possibility that comparing morphologically related forms is a process that is available on only a minority of occasions that a particular form is encountered. The learning model should be robust enough to (eventually) learn arbitrary paradigmatic mappings even under the conditions of imperfect recall, which would slow down their acquisition relative to product-oriented schemas.

Evidence for product-oriented schemas is provided by the finding that examples of a particular paradigmatic mapping may increase productivity of other paradigmatic mappings resulting in the same product. For example, examples supporting producing a plural ending in -tʃi by simple addition of –i to tʃ (0→i/tʃ__) have been found to increase palatalization of [t] before the attached –i (t→tʃ/__i) (Kapatsinski 2012, 2013; see also Bybee 2001). As noted by Kisseberth (1970), these kinds of patterns are problematic for all approaches to morphophonology that insist on a strict separation of change and context, including rule-based approaches (e.g. Albright & Hayes 2003, Chomsky & Halle 1968) and the hybrid approach of Becker & Gouskova (2014), Gouskova et al. (2015). Under rule-based approaches, for example, examples of "0→i/p_" can help "0→i/z_" because they feature the same change in a different context, and the learner is thought to generalize over contexts to determine where particular changes occur. In contrast, examples of "0→i/tʃ__" cannot help "t→tʃ/__i" despite resulting in the same output because they do not involve the same change: the exemplified rule adds a vowel, while the other rule changes a consonant. Similarly, Gouskova et al.'s (2015) approach allows for generalizations about the kinds of words that undergo or result from a certain change (e.g., 0→i) but does not allow for generalizing about outputs that share a meaning but result from all kinds of changes. In order to capture product-oriented generalizations, we need to allow the learner to generalize across diverse changes resulting in the same product. We do this by means of direct meaning-to-form connections (product-oriented schemas or constructions), abandoning the change/context distinction in learning paradigmatic mappings and by learning syntagmatic dependencies within semantically-defined form sets (for example, learning that, in English nouns, *-ity* is often preceded by [s]).

The data to account for

To demonstrate the feasibility of the proposed approach, I apply the model to the data on human miniature artificial language learning reported in Kapatsinski (2009, 2012, 2013), Stave et al. (2013) and Smolek & Kapatsinski (Submitted). While somewhat lacking in ecological validity, miniature artificial

language data have the advantage of allowing the model to be trained on the input experienced by human learners whose behavior we hope to capture.

All of the modeled experiments examine the learning of miniature artificial languages with palatalization, comparing palatalization rates for different consonants. The languages presented to human learners by Kapatsinski (2009, 2013) are summarized in Table 1. All languages have singular and plural nouns, two plural suffixes –i and –a, where –i always attached to [k]-final singulars, and a process of velar palatalization that turns /k/ into [tʃ] before [i]. They differ in whether –i is the dominant plural suffix for singulars ending in [t] or [p] and whether singular-plural pairings in which –i attaches to a singular ending in [tʃ] are presented in training.

Table 1. Languages shown to human learners in Kapatsinski (2009, 2010, 2012, 2013)

	Tapa	Tipi	Tapachi	Tipichi
SG=…k#	PL=…tʃi#			
SG=…{t;p}#	75% PL=…{t;p}a# 25% PL=…{t;p}i#	25% PL=…{t;p}a# 75% PL=…{t;p}i#	75% PL=…{t;p}a# 25% PL=…{t;p}i#	25% PL=…{t;p}a# 75% PL=…{t;p}i#
SG=…tʃ#	Not presented		PL=…tʃi#	

Kapatsinski (2013) shows that a rule-based model of morphophonology (Albright & Hayes 2003), predicts that examples of tʃ→tʃi, added to the training in Tapachi and Tipichi, should favor addition of –i to other consonants. Rules are changes in context. The learner takes in pairs of morphologically related forms like butʃ~butʃi, splits them into a change (here, 0→i) and a context in which the change was observed (here, after [butʃ]). The learner then gradually generalizes over contexts, learning what kinds of contexts favor particular changes. Thus, butʃ→butʃi, slait→slaiti, and klop→klopi would lead to the emergence of a general rule 0→i/[-voice;-cont]_. Any additional examples of 0→i would strengthen the rule, resulting in more t→ti, p→pi and k→ki. At the same time, they would reduce the likelihood of the learner producing t→tʃi, k→tʃi, and p→tʃi, by reducing support to the competing k→tʃ change exemplified by bluk→blutʃi. However, examples of tʃ→tʃi were found to consistently lead participants to produce more t→tʃi compared to t→ti. This is unexpected under a rule-based model (since t→ti shares a change with tʃ→tʃi) and was argued to support a product-oriented model under which participants are learning that plurals often end in [tʃi].

The same problem should arise for a hybrid model recently proposed by Becker & Gouskova (2014) and Gouskova et al. (2015). Their model splits the lexicon into sublexica defined by the changes they exhibit. The model then learns about the characteristics of words that undergo each change, and the characteristics of words that result from each change. Because it generalizes over whole forms, the model can make generalizations that span the change/context boundary. However, its decision on which change to apply is still based only on characteristics of words that undergo the various changes. As in the case of Albright & Hayes (2003), experiencing tʃ→tʃi examples should help 0→i by expanding the range and number of forms that undergo this change.

Another difficulty presented by these data for source-oriented models is that singulars ending in [tʃ] take –i rather than –a even in the Tapa language, where –a is as common as –i and [tʃ]-final singulars are never experienced in training. Source-oriented models learn about what kinds of source forms take –a and what kinds of forms take –i. In Tapa, [k]-final sources are the only ones that take –i. As a result, [k]-final singulars should be much more likely to take –i than [tʃ]-final singulars while the opposite is true of the

human data. In contrast, the fact that [tʃ]-final singulars take –i even in the Tapa language can be explained by generalization over plural forms: even in Tapa, -i follows a preceding [tʃ] 100% of the time (i.e., p(i|tʃ)=100%).

At the same time, a purely product-oriented model of these data is also untenable. In particular, one needs to explain how learners come to palatalize [k] more than [t] or [p], a restriction on what kinds of sources map onto [tʃi]. In addition, stem changes resulting in [tʃi] appear to be disfavored by placing singulars and plurals sharing the stem next to each other. If generalization is purely product-oriented, the learner never needs to refer to the source form, hence its placement near the product form should not matter. Based on this finding, Kapatsinski (2013) proposed that placing stem-sharing forms without a stem change next to each other helps copying: when one encounters a word pair like butʃ-butʃi, one can notice that the singular-final consonant is retained in the plural, which then boosts retention of singular-final consonants in constructing the plural form. Kapatsinski's (2013) notion of copying is essentially equivalent to the base identity constraints of Benua (1997) and Kenstowicz (1996).

Here, I extend this proposal, showing that the modified version can account for an aspect of the data that is not captured by Kapatsinski's (2013) model, C4B. Namely, when corresponding singulars and plurals are placed next to each other, examples of tʃ→tʃi still help t→tʃi over t→ti but they help k→ki over k→tʃi (Kapatsinski 2009, chapter 3). This is unexpected because C4B restricts copying generalizations to specific experienced articulatory chunks: experiencing copying of [tʃ] helps copy [tʃ] but does nothing to help other chunks like [k] survive into the plural. We propose that copying is instead specific to a certain templatic position (e.g., the word-final consonant), but that copying can be affected (favorably or unfavorably) by the presence of various phonological units in the source form(s).

I hold the model accountable for the following results:

1) The effect of temporal order:
   Presenting corresponding singulars and plurals next to each other (in pairs) discouraged changing stops into [tʃi] (Kapatsinski 2012)
2) The effect of adding examples of tʃ→tʃi:
   a. Examples of –i simply attaching to [tʃ] (tʃ→tʃi) helped t→tʃi over t→ti. This result is taken by Kapatsinski (2009, 2012, 2013) to provide crucial support for product-oriented generalizations.
   b. However, when corresponding singulars and plurals were presented next to each other, the same examples helped k→ki over k→tʃi (Kapatsinski 2009, Chapter 3; Yin & White 2016). The result is problematic for a purely product-oriented theory (Kapatsinski 2009, Chapter 3).
   c. Examples of tʃ→tʃi had no effect on the competition between k→ki and k→tʃi when all wordforms were presented in random order (cf. Kapatsinski 2009, Chapter 4).
   d. There was no effect of tʃ→tʃi examples on p→tʃi vs. p→pi because [p] was almost never palatalized.
3) Input [tʃ] favors –i, even if never encountered:
   Singulars ending in [tʃ] heavily favored the suffix –i over –a during test even when they were not encountered in training and –a was the more common suffix (the Tapa language in Kapatsinski 2009). In particular, [tʃ]-final singulars favored -i more than [k]-final singulars did after exposure to the Tapa language, despite the fact that [k]-final singulars were always mapped onto [tʃi] in training while [tʃ]-final singulars were never presented.
4) Lack of an effect of adding examples of tʃ→tʃu:

Kapatsinski (2013) reports that adding examples of tʃ→tʃu – unlike adding examples of tʃ→tʃi – failed to significantly help t→tʃi over t→ti compared to the Tapa and Tipi languages.

5) The effect of often adding –i to {t;p}:
   Exposure to Tipi and Tipichi languages led participants to often simply add –i to [k]-final singulars, without changing the consonant, compared to Tapa and Tapachi. Kapatsinski (2009, 2010) shows that this result is predicted by the rule-based model of Albright & Hayes (2003). It is also captured by a schema-based model with the assumption that schemas start out general and gradually increase in specificity over the course of learning (Kapatsinski 2013).

Naïve Discriminative Learner

The Naïve Discriminative Learner (NDL, Arppe et al. 2014, Baayen et al. 2011, Ramscar et al. 2013) is an implementation of the RW learning theory as a fully connected two-layer connectionist network. The architecture of the network is identical to Rumelhart & McClelland's (1986) classic (and classically criticized) model of the English past tense. Each node in the network corresponds to a linguistic unit that can be associated with other units. The nodes are arranged into two layers, where one layer represents the input, and the other layer represents the output. Learning is unidirectional, in that the network learns to predict the outputs from the inputs. Each node within the input layer is a *cue*, while each node within the output layer is called an *outcome*. Learning in the network is error-driven: the network adjusts its connections if the outcomes it predicted to occur were not the outcomes that actually occurred. Cues compete with each other to predict outcomes, such that the cues that are most predictive of an outcome become strongly and positively associated with that outcome, and cues that are predictive of the absence of an outcome acquire a strong negative association with the outcome.

A learning experience for the model consists of a pairing between a set of cues and a set of outcomes. The network has a connection from every cue to every outcome. Each connection has a weight. The sign of the weight represents whether the connection is excitatory or inhibitory. Positive weights represent excitation (the cue predicts the presence of the outcome), while negative weights represent inhibition (the cue predicts the absence of the outcome). The absolute value of the weight indicates the strength of the prediction. An outcome is expected to the extent that the present cues have strong excitatory connections to the outcome. If an outcome occurs unexpectedly, the weights of the connections from the present cues to the outcome are adjusted upward (as a result, excitatory connections are strengthened, while inhibitory ones are weakened or even change sign). If an outcome is unexpectedly absent, the present cues' connections to the outcome are adjusted downward (inhibitory connections strengthen, while excitatory ones weaken or even become inhibitory). The network learns nothing about the predictions of cues that were absent during any given learning experience.

Model 1. Paradigmatic and schematic structure

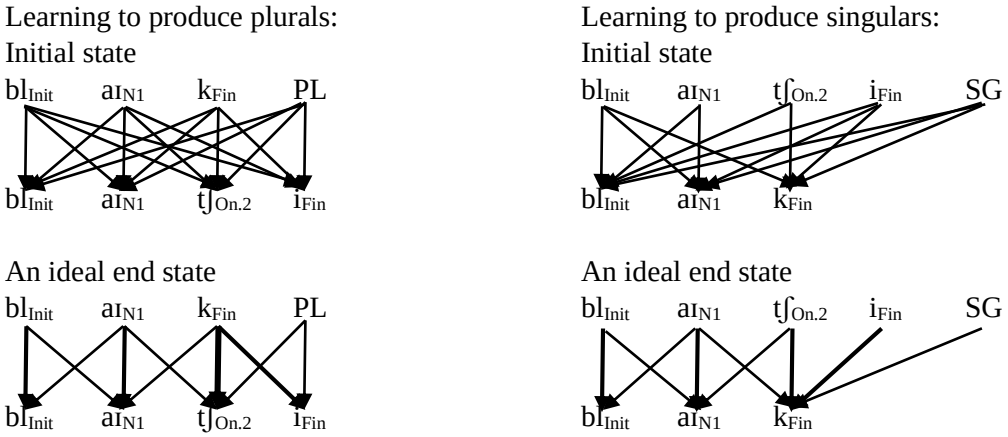
Learning task: Form prediction

The task of the present model is *form prediction*. A learning experience consists of experiencing a wordform. We assume that, when a wordform is experienced, the meaning of the wordform is activated in memory, and that – *on at least a minority of occasions* – other forms of the same word are also activated in memory. The model then attempts to predict the experienced wordform from other remembered forms

of the same word and the meaning it thinks is being expressed. In other words, the experienced wordform is a set of outcomes (i.e., an output), and the activated memories are sets of cues (i.e., input).[2]

For example, the model might experience a plural form /blaɪtʃi/, corresponding to the outcome set {$bl_{On.1}$, $aɪ_{N.1}$, $tʃ_{On.2}$, $i_{N.2}$}. This will then bring to mind the meaning of the form, crucially including the feature 'PL', and the other form of the same word, /blaɪk/, encoded as a set of cues, e.g. {$bl_{On.1}$, $aɪ_{N.1}$, $k_{Cd.1}$}. If the model works perfectly, then – as a result of many such individual experiences – it will learn the outcomes that usually occur when the plural meaning is being expressed (such as $i_{N.2}$ and $tʃ_{On.2}$). It will also learn that $k_{Cd.1}$ in the singular predicts {$tʃ_{On.2}$, $i_{N.2}$} in the plural, and that {$bl_{On.1}$} in the singular predicts {$bl_{On.1}$} in the plural while {$aɪ_{N.1}$} in the singular predicts {$aɪ_{N.1}$} in the plural. The start and end states of the network are illustrated in Figure 1 (only active cues and outcomes, about which something is learned, are shown).

Figure 1. An illustration of NDL. On the left, the network is experiencing a plural form. It activates the plural meaning and the singular form as cues (training top-to-bottom connections). On the right, the network is experiencing a singular form (training bottom-to-top connections). It activates the plural form and the singular meaning as cues. Connections from present cues to present outcomes are strengthened to the extent that the outcome was unexpected. Connections from present cues to absent outcomes are weakened to the extent that the absence of the outcome was unexpected. Connections from always absent cues or to always absent outcomes like [ki] in Table 1 are not adjusted.



Encoding scheme

An individual outcome or cue is an associable phonological unit. The repertoire of associable phonological units includes syllabic constituents like onsets and rimes (Kapatsinski 2009b), as well as segments (Newport & Aslin 2004) and subsegmental units such as gestures or features (Moreton 2008a). The network's input layer contains a cue node for each distinct phonological unit that has been experienced in inputs, and the output layer contains an outcome node for each distinct phonological unit that has been experienced in outputs. We assume that these units are position-specific. Thus, for example,

[2] Alternatively, one could say that experiencing a word generates a set of predictions about features of other forms of that word, which are then confirmed or rejected when the forms are encountered. Again, this would only have to happen on a minority occasions. The behavior of the model after a block of training is independent of directionality of prediction engaged in during the block. However, the directionality could make a difference in explaining trial-by-trial data if we had it.

predicting the occurrence of a final [i] is not the same as predicting the occurrence of a stem [i]. If a stem [i] occurs in [hita], the cues that were predicting a final [i], such as the final [t] of the singular [hit] are punished rather than rewarded.

For the languages in the present simulations, we are interested in modeling what is palatalized and when. Because of the way these languages were designed, features of the stem-final consonant are relevant for predicting the outcome, while features of other segments within the singular are not. One of the well-known limitations of RW, and therefore NDL, is lack of selective attention (see Kruschke, 1992, for an extension that does have selective attention). While we think selective attention is crucial for language learning (see also Ellis 2006), how the learner zeroes in on the important parts of the stem is not the focus of this paper. We therefore chose to present the model only with the task-relevant phonological features, namely those of the stem-final segment. The features used were sufficient to discriminate between the consonants presented in training, and to represent the articulatory similarity relations among them: [voiced], [voiceless], [lips], [tongue], [(tongue) blade] and [(tongue) body]. Thus, [p] would be [voiceless; lips], [t] -- [voiceless; tongue; blade], [k] -- [voiceless; tongue; body], and [tʃ] – [voiceless; tongue; blade; body].

Given the structure of the languages in the present study, we could have treated the body of the stem (onset+nucleus) as an undecomposable cue identifying the stem without losing any information relevant to palatalization. However, Kapatsinski (2009b)'s data suggested that bodies are not associable units in English, thus we represented each body as a combination of an onset cue and a vowel cue. We gave the model the identity of the stem vowel but treated the onsets as undecomposable cues. The onset and the vowel together serve to identify individual words, which is important for accurate performance on training data because words ending in not-to-be-palatalized consonants vary in whether they take –i or –a. The choice of the suffix for an individual word is predictable only if you know the word, which cannot be uniquely identified without knowing the stem body.

Outcome chunking

Just as Optimality Theory is a theory of constraint interactions, and not a theory of constraints (McCarthy 2008), RW and NDL do not provide an inventory of cues and outcomes. However, a good specification of the learning problem in terms of the cues available to the learner and the outcomes s/he tries to predict is crucial for modeling the observed behavior.

The Naïveté of Naïve Discriminative Learning consists of the assumption that cues are independent of each other, as are outcomes. While the model captures some dependencies between cues because cues compete with each other for predicting outcomes, dependencies between outcomes are not captured. However, the very idea of morphophonology suggests the existence of dependencies between elementary phonological units in the output. For example, the choice of the suffix is not independent of the decision to change the stem: here, the stem should be changed if the chosen suffix is –i but not if it is –a. To capture this dependency in the present paradigmatic model, we need [tʃi] to be an outcome.[3]

In general, small units like phonological features are often beneficial at the cue level since they allow one to capture the similarity structure among inputs, allowing similar inputs predicting the same outcome to provide support for the same generalization.[4] However, one could argue that small units can be detrimental to performance at the outcome level: if the outcomes are materials for constructing a word (or

[3] Later on, we will see that configural outcomes are not necessary if the present model is combined with a model engaged in syntagmatic prediction within semantically-defined sublexica, e.g. the sublexicon of plural forms.
[4] Note that it is the same *outcome*, not the same output: the output is a set of elementary outcomes.

any kind of complex, quasi-compositional behavior), then outcomes representing large chunks may allow for faster construction. This intuition is behind the preference for specific schemas in Langacker (1987) and Nesset (2008), and for morphemes that realize more morphosyntactic features in Caballero & Inkelas (2013): a more specific schema or morpheme gets you closer to the desired end product. However, since all competing schemas in the present paper are plural, and the speaker is filling out a plural phonological template, the relevant notion of specificity is at the form level rather than the meaning level.

While specificity of outcomes can be beneficial in capturing dependencies and in allowing for faster convergence on an output, specificity also has a cost. If we take specificity to an extreme, every output would be represented by a unique non-compositional outcome. With such an encoding, it would be impossible to generalize: no two combinations of cues would lead to the same outcome. The learner would therefore be lost when a novel input word is encountered.

In a full model of language acquisition, the emergence of task-relevant configural cues or configural outcomes should also be explicitly modeled (see Goldstone 2003 for a promising approach). However, for the purposes of this model, we assume that 1) chunks results from syntagmatic prediction, and 2) syntagmatic prediction is (near-) optimal. The first assumption is consistent with work suggesting that a chunk inventory provides a good description of the knowledge acquired in statistical learning tasks (e.g. French et al. 2011, Giroux & Rey 2009, Hewlett & Cohen 2011, Slone & Johnson 2015; cf. Baayen et al. 2011). Optimality of prediction suggests sensitivity to contingency relations in the environment, measured by a statistic called ΔP (Shanks 1995, Wasserman 1990). ΔP takes the conditional probability of an outcome in the presence of a cue and subtracts from it the conditional probability of the same outcome in the absence of the cue: p(Outcome|Cue)-p(Outcome|~Cue).

The use of contingency by human learners has been famously demonstrated by Gluck & Bower (1988), who devised a set of contingencies in which the predictions of ΔP differ from those of simple conditional probability of the outcome given the cue p(Outcome|Cue) and the prior probability of the outcome, p(Outcome). Gluck & Bower presented learners with two diseases, one rare (RD), and one common (CD), and a set of symptoms. In the presence of one of these symptoms (Symptom X), the probabilities of RD and CD were equal. Nonetheless, participants judged that a patient with the symptom was more likely to have the rare disease than the common disease. This is puzzling because p(RD|SymptomX)= p(CD|SymptomX) and p(RD)<p(CD). However, it is explained by the fact that p(RD|SymptomX) > p(RD|~SymptomX): while the probabilities of the two diseases are equal when Symptom X is present, RD is rarer than CD otherwise, so Symptom X is indeed diagnostic of RD. The effects of ΔP naturally fall out of RW: because cues compete with each other for predicting the outcome, the outcome becomes associated with the cues that predict it best (Gluck & Bower 1988).

Recently, ΔP has been successfully applied to the extraction of collocations in corpus linguistics (Gries 2013, Wahl 2015; see also Ellis et al. 2014 for sensitivity to contingency in syntactic processing). The present model assumes that it is also useful for identifying *sub*lexical chunks including those that cross traditional morpheme boundaries. For example, consider the languages in Table 1. Because the suffix –a attaches only to non-palatalizable stem consonants, the outcome of palatalization, [tʃ], is highly predictive of an upcoming –i, while [t] and [p] are predictive of an upcoming –a. $p(-i|t\int_{On.2}) = 100\%$, $p(-i|t_{On.2}) = p(-i|p_{On.2}) = \{25\%;75\%\}$, so $\Delta p(-i|t\int_{On.2}) = +75\%$ in Tapa and Tapachi and +25% in Tipi and Tipichi, positive in both cases. Conversely, $\Delta p(-a|t_{On.2}) = \Delta p(-a|p_{On.2}) = +75\%$ in Tapa and Tapachi and +25% in Tipi and Tipichi. Thus, we expect that, at least [pa], [ta] and [tʃi] should become chunks, at least in Tapa and Tapachi. Compare this to the situation in Tipichu, the language in which tʃ→tʃu examples were added

instead of tʃ→tʃi examples to Tipi in Kapatsinski (2013). Here, $p(-i|tʃ_{On.2})$ reduces to 50%, so [tʃi] is no longer a chunk, since $\Delta p(-i|tʃ_{On.2})$ is negative (-25%).[5]

Given the assumption that the speaker wants to construct the output form as fast as possible, large chunks have priority over smaller chunks. If one wants a given word to always be derived the same way, one could argue that [tʃi] should always be selected over [tʃ] and [i], as this is the most efficient single way to construct a word (cf. Caballero & Inkelas 2013, Nesset 2008). According to RW, evidence that a cue predicts a certain outcome does not provide any information about whether it might also predict other outcomes (cf. Miller & Matute 1998, Waldmann 2000). Because of this, one advantage of the 'always pick the biggest chunk' approach is that the behavior of the model is unaffected by whether parts of a fused outcome chunk are retained as additional separate outcomes after the formation of the chunk. Since this is a difficult question to answer empirically (e.g. Baayen et al. 2011), especially in production, this is a real advantage of this approach.[6]

On the other hand, we could also allow the parts and the whole to race for selection in parallel and let the outcome most expected in the context win (a multiple-route model). This is the dominant approach in psycholinguistics, e.g. Beekhuizen et al. (2013), Gagné & Spalding (2015), Kapatsinski (2010b), Sahel et al. (2008); cf. also Baayen et al. (1997), Kuperman et al. (2009) for perception. As emphasized by Beekhuizen et al. (2013) and Kapatsinski (2014), this means that the same exact output can be produced in many different ways by the same speaker. While requiring much redundancy and wasteful spending of long-term memory space (cf. Baayen et al. 2011, Chomsky & Halle 1965), this approach has its advantages: it is consistent with the parallel architecture of the brain, allows for robustness in the face of neural diversity, and maximizes processing speed. Somewhat surprisingly, the modeling results presented later appear to shed light on this controversy and be more consistent with the multiple-route approach.

Copy outcomes

A copy outcome is detected during training when an input unit is retained in the output (as would satisfaction of a faithfulness constraint in Optimality Theory, Prince & Smolensky 2004). We assume the existence of a copy outcome associated with each position within the template representing the inputs (cf. positional faithfulness constraints of Beckman 1998). In particular, we will need a $Copy_{Fin}$ outcome that is detected whenever the final consonant of the singular is retained in the plural output. Since this outcome is only detected when the final consonant of the singular form is retained, it will become associated with input cues that favor the retention of the stem-final consonant; in particular, [place] features of the final

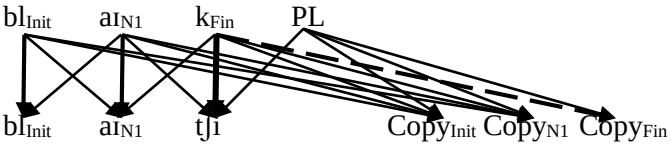
[5] To the model, units emerge out of statistics of segment co-occurrence. However, in reality prosody constrains unit formation in speech in the same way as Gestalt principles constrain unit formation in visual perception (e.g. Goldstone 2000), as evidenced by prosodic cues overcoming statistical cues to word boundaries (Johnson & Jusczyk 2001, Shukla et al. 2007). For example, consider the extraction of *–holic* out of *alcoholic* (Bybee 1985). Given that English-speaking children appear to posit word boundaries before stressed syllables (Cutler & Norris 1988), it might not be an accident that *–holic* is a stress-initial foot. An additional constraint is likely posed by the rapidity of speech and the existence of extensive co-articulation. If prediction happens in real time (rather than in memory as post-hoc analysis), it may often be impossible to predict an upcoming sound using an adjacent or near-adjacent preceding sound as a cue because both sounds are recognized simultaneously or nearly so. Simultaneous activation may in and of itself lead to fusing units together (e.g. Bybee 2002, Kapatsinski 2007) and may actually be necessary for real-time prediction.

[6] This is not true of cues, where parts of a fused "configural" cue should be retained in the model when associations of the whole may not be associations of the parts but learning associations of the whole interferes with learning associations of the parts. This pattern of data is difficult to account for if the parts are not perceived when the whole is presented (Kapatsinski 2007b, 2009b).

consonant. We will also need a $Copy_{Init}$ outcome that is detected when the initial onset of the activated input is retained in the experienced output, and a $Copy_{N1}$ output that detects the retention of the stem vowel. This set of Copy outcomes is sufficient for modeling the present data but is not exhaustive.

Let us now illustrate the operation of Copy outcomes. Suppose that a learner experiencing the plural [blaɪtʃi] activates the singular form [blaɪk]. Again, we assume that this happens at least some of the time when [blaɪtʃi] is experienced. The learner will then detect the outcomes $Copy_{Init}$ and $Copy_{N1}$ but not $Copy_{Fin}$. The learning experience will then be represented as Input = {bl, $aɪ_{N.1}$, [stop], [voiceless], [tongue], [body]}, Output = {bl, $aɪ_{N.1}$, tʃi, $Copy_{Init}$, $Copy_{N1}$}. After a few learning experiences of this kind, and other experiences in which other final consonants *are* present in the output, the learner will associate [body] with the absence of $Copy_{Fin}$, inhibiting the copying of final velars into the plural.[7] This illustrates that copying in the trained model is input-specific, i.e. conditional on meaning to be expressed and formal characteristics of activated morphologically related words, allowing the model to learn *when* and *what* to copy.[8] The erroneous output [bliki] in a language with velar palatalization, thus lacking the outcome [ki], would be predicted if $Copy_{Fin}$ has positive activation that exceeds that of [tʃi] or [$tʃ_{On.2}$].[9] The system is illustrated in Figure 2.

Figure 2. Excitatory and inhibitory links most relevant to producing the plural form of [blaɪk] when presented with the singular after training on one of the languages in Table 1. The dashed line shows an inhibitory connection. All other connections are excitatory. $Copy_{Init}$ and $Copy_{N1}$ are activated by all input cues, while $Copy_{Fin}$ is inhibited by the input consonants that are not retained in the plural.



Wordform Generation

We used the ndl package in R (Arppe et al. 2015) to derive the cue-outcome weights and outcome activations in the presence of specific cue sets. The weights in the present simulations are asymptotic: they are what the learner would arrive at after settling into a stable equilibrium according to Danks

[7] If $Copy_{Fin}$ is never obeyed, it will be absent from the set of outcomes and so will be never obeyed in any environment.

[8] By separating the position-specific faithfulness constraint from its condifitioning environment, the present approach is more flexible than positional faithfulness constraints, which are specific to both a particular position and a particular input feature (Beckman 1998). Thus, the activation of a copy outcome may vary across morphological environments and generalize across input segments, allowing the model to learn generalizations like 'the initial syllable should be preserved' (Becker et al. 2012) or 'don't retain the final /z/ of a plural form when deriving a singular'. This appears crucial for implementing *morpho*phonology, where changes affect only segments in specific positions and specific semantic contexts (such as PLURAL or SINGULAR).

[9] With a wider variety of prosodic shapes, a more complex template would be necessary. It remains to be seen whether a satisfactory template could be devised (Pinker & Prince 1988) without reference to word-internal structure. However, that goal is beyond the scope of this paper.

(2003).[10] The outcome activation is simply the sum of the weights of connections from the current set of cues to that outcome.

We assume the existence of a prosodic template associated with the to-be-expressed meaning (Kapatsinski 2013, Redford 2015, Vihman & Croft 2007). The template may grow over the course of learning, while retaining alignment with word edges by labeling positions from both the beginning and the end. We assume that the template is learned early and selected first (Shattuck-Hufnagel 2015). Our modeling effort focuses on the filling out of this template with segmental material.[11]

Some outcomes are mutually incompatible, e.g., given the input [blaɪk], [blade] (rooting for final [tʃ]) competes with $Copy_{Fin}$. Generally, outcomes are incompatible if they compete for the same position in the output template. To derive an output form, we therefore choose between the competing outcomes in each templatic position. Following Luce (1977), we assume that the choice between incompatible segmental outcomes is stochastic: the probability of the speaker choosing a segmental outcome is equal to its activation divided by the sum of activations of the full set of competing outcomes competing for the same position, including the Copy outcome associated with that position. The process stops when the selected prosodic template (e.g. OnN.OnN) associated with the to-be-expressed is filled.

Simulation results

The need for copy outcomes

In order to provide an adequate simulation of human behavior in the elicited production task, the model must construct a fully specified output form. Thus, we want the model not to merely predict the suffix and to decide whether to palatalize but rather to predict all phonological characteristics of the output (see also Albright & Hayes, 2003). Part of that is constructing the stem allomorph and, importantly, retaining the parts of the input stem allomorph that need to be retained. In this section, we show that the notion of copying is necessary for the network to retain stem parts it has just encountered.

Pinker & Prince (1988) criticized Rumelhart & McClelland's (1986) model, which is much like the present model without copy outputs, for predicting 'bizarre stem changes', such as *membled* as the past tense of the English verb *mail*. This is not a problem for NDL in the same way: even without Copy outcomes, the model performs perfectly on items that are familiar, as *mail* would be for an English speaker. In the same way that it perfectly learns which of the familiar verbs take –a, and which ones take –i, it also perfectly learns which input onsets map onto [bl] and which ones map onto [tr].

The model also performs perfectly on novel stems *as long as all stem parts during test* – i.e., cues comprising the stem allomorph – *have been experienced in training*. In Table 2, every familiar input stem onset activates the corresponding output stem onset more than it activates any other onset despite the absence of Copying. The model appears close to the perfect endstate illustrated in Figures 1-2: familiar

[10] The results from the simulated implicit learning mechanism thus indicate the limits of what the model can learn, and the model's performance woThere is also a version of NDL, ndl2, available for modeling trial-by-trial learning but this is beyond the scope of this paper since the data we are trying to fit is aggregated across participants who were presented with different random trial order.

[11] If multiple prosodic templates are required, these may be treated as distinct outcomes. The material could also be gestural, in the sense of Browman & Goldstein (1989).

input stem onsets are retained in the output. This is also true of familiar stem nuclei and not-to-be-palatalized final consonants.

Table 2. Activations of word-initial onsets by test items in Kapatsinski (2013), which shared onsets and vowels with stimuli presented during training

	bl	tr	sw	v	fl	gw	kl	b	h
bl_aɪp	**0.98**	-0.28	-0.19	-0.12	-0.02	0.25	0.17	0.17	-0.09
tr_up	-0.02	**0.72**	-0.19	-0.12	-0.02	0.25	0.17	0.17	-0.09
sw_ip	-0.02	-0.28	**0.81**	-0.12	-0.02	0.25	0.17	0.17	-0.09
v_orp	-0.02	-0.28	-0.19	**0.88**	-0.02	0.25	0.17	0.17	-0.09
fl_ot	0.02	0.2	0.07	-0.15	**1.02**	-0.27	-0.14	-0.14	0.24
gw_it	0.02	0.2	0.07	-0.15	0.02	**0.73**	-0.14	-0.14	0.24
kl_ut	0.02	0.2	0.07	-0.15	0.02	-0.27	**0.86**	-0.14	0.24
b_ut	0.02	0.2	0.07	-0.15	0.02	-0.27	-0.14	**0.86**	0.24
h_ik	-0.01	0.08	0.12	0.27	-0.01	0.02	-0.02	-0.02	**0.85**

However, consider what happens when the model encounters a novel onset. In Smolek & Kapatsinski (Submitted), all stem bodies encountered at test were novel, mostly because they had novel onsets. Table 3 shows that when a novel onset is encountered, the network without copy outcomes has no knowledge on whether that onset should be retained. While it has learned to retain familiar onsets, without copy outcomes, that knowledge is a set of arbitrary input-output associations specific to the experienced onsets. A novel onset activates all experienced onsets nearly equally, as seen with [bl] below. Other parts of the word can then decide on the onset to be produced. For example, the words [dræk], [fræp], [kwæg], [plæp] and [ʃræt] in Table 3 contain the vowel [æ], which the model has experienced in the singular-plural pairs [dæt]-[dæta] and [slæk]-[slætʃi]. Because of this, the model learned that the vowel [æ] predicts the onset [d]. With an unfamiliar input onset, the output onsets [d] and [sl] are therefore activated most highly by any word containing [æ]. This behavior is sensible but distinctly non-humanlike.

Table 3. Activations of experienced word-initial onsets by the test words in the present study

	l	gl	r	d_{On}	gw	sl	sn	p_{On}	v	g_{On}
bl_ɪt	0.06	0.05	0.05	0.06	0.05	0.06	0.05	0.05	0.06	0.05
dr_æk	0.02	0.02	0.02	0.35	0.02	0.35	0.02	0.02	0.02	0.02
fr_æp	0.02	0.02	0.02	0.35	0.02	0.35	0.02	0.02	0.02	0.02
kl_ap	0.35	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.35	0.02
kr_aɪt	0.02	0.06	0.06	0.02	0.06	0.02	0.06	0.06	0.02	0.06
kw_æg	0.02	0.02	0.02	0.35	0.02	0.35	0.02	0.02	0.02	0.02
pl_æp	0.02	0.02	0.02	0.35	0.02	0.35	0.02	0.02	0.02	0.02
pl_at	0.35	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.35	0.02
s_ɛb	0.02	0.06	0.06	0.02	0.06	0.02	0.06	0.06	0.02	0.06
ʃl_ag	0.35	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.35	0.02
ʃr_æt	0.02	0.02	0.02	0.35	0.02	0.35	0.02	0.02	0.02	0.02
ʃr_ig	0.02	0.06	0.06	0.02	0.06	0.02	0.06	0.06	0.02	0.06

When equipped with copy outcomes, the network successfully learns that the stem onset is always copied, with activation of $Copy_{Init}$ ranging between .71 and 1.0 across the test words. The copy outcomes are therefore able to outcompete the learned arbitrary associations between input vowels and codas and output onsets for every test word, successfully avoiding the 'bizarre' stem changes of Pinker & Prince (1988) that the model otherwise produces. This provides the primary motivations for Copy outcomes. We will now examine whether the model, equipped with Copy outcomes, is able to account for the results of miniature artificial language learning studies of palatalization.

Learning what (not) to change

The addition of tʃ→tʃi examples in Tapachi and Tipichi helps t→tʃi over t→ti. In the human data, when stem-sharing singulars and plurals are presented next to each other, these same examples have been observed to help k→ki over k→tʃi (Kapatsinski 2009, Chapter 3, p.63; Yin & White 2016; though cf. Kapatsinski 2013). When they are not presented next to each other, the examples have no effect on k→ki vs. k→tʃi (Kapatsinski 2009, Chapter 4, 2012). In either case, these examples are unable to help p→tʃi over p→pi because [p] is almost never turned into [tʃ] by the participants. To simplify presentation, we focus on Tapa vs. Tapachi in presenting the model results below but the results for Tipi vs. Tipichi are exactly parallel.

Note that examples of tʃ→tʃi provide evidence for both $Copy_{Fin}$ (copying of the stem-final consonant) and the generalization that plurals end in [tʃi]. Theories separating change and context (Albright & Hayes 2003, Chomsky & Halle 1968, Gouskova et al. 2015) emphasize the former fact, and therefore have trouble capturing the finding that these examples help t→tʃi over t→ti (Kapatsinski 2013). Network Theory (Bybee 1985, 2001) and Clamoring for Blends (Kapatsinski 2013) emphasize the latter fact and therefore have trouble explaining why these examples might help k→ki over k→tʃi (Kapatsinski 2009).

Without Copy outcomes, NDL behaves like Network Theory and Clamoring For Blends, predicting that tʃ→tʃi should help both t→tʃi and k→tʃi. In contrast, NDL with Copy outcomes predicts the observed difference between the effects on t→tʃi and k→tʃi. As Table 4 shows, the model correctly predicts that there should be more palatalization of [t] in Tapachi and more palatalization of [k] in Tapa. The increased palatalization of [t] in Tapachi is driven by greater activation of [tʃi] and, crucially, reduced activation of [t] in that language. In contrast to [t], [k] never occurs in plurals during training in either language, and therefore is not an outcome the network considers learning about. As a result, tʃ→tʃi examples reduce the incidence of velar palatalization, by providing support to $Copy_{Fin}$.

Table 4. Palatalization rates in NDL without copy outcomes (a), with copy outcomes for participants that always have access to the singular when perceiving the plural (b), and with copy outcomes when participants cannot recall the singular when presented with the plural half of the time in training (c).[12]

a. No copy outcomes

Singular ends in	Tapa		Tapachi
p	2%	=	1%
t	4%	<	9%
k	90%	<	100%

b. Copy outcomes

Singular ends in	Tapa		Tapachi

[12] These results hold with or without the configural/fused outcome [tʃi]. With [tʃi], the palatalization rates of [t] and [k] increase but the pattern of results is unchanged.

p	1%	=	0%
t	3%	<	6%
k	93%	>	87%

c. Copy outcomes + imperfect recall

Singular ends in	Tapa		Tapachi
p	4%	=	5%
t	5%	<	10%
k	89%	=	88%

Note that in order for a tʃ→tʃi example to provide support for $Copy_{Fin}$, the singular form must come to mind when the plural form of the same word is presented to the participant. Otherwise, repetition of the stem-final consonant across forms is not detected. As argued in Kapatsinski (2012), recalling the singular form when presented with the plural is very likely when stem-sharing singulars and plurals are presented next to each other in time, but much less likely if all wordforms are presented in random order.

In order to simulate the effect of presenting all wordforms in random order, we replaced half of the training trials with trials on which phonological characteristics of the singular are not recalled, and Copy outcomes are therefore also absent, i.e., the input consists of the PLURAL meaning, while the outcome is the fully specified plural form (Table 4c). With this modification, the rate of velar palatalization is virtually identical between the two languages and the productivity of non-velar palatalization is increased. This is exactly the pattern of results elicited when stem-sharing singulars and plurals are no longer presented next to each other in pairs, making comparison of morphologically related forms more difficult (Kapatsinski 2012).

The need for configural outcomes: tʃi vs. tʃu

Kapatsinski (2013) reports that, whereas the addition of tʃ→tʃi examples to Tapa or Tipi helped t→tʃi, the addition of tʃ→tʃu examples did not, which he takes to be evidence that a product-oriented schema needs to be aligned with a word boundary. NDL captures this effect with outcome fusion driven by syntagmatic contingency. In this version of the model, [tʃi] fuses into an outcome in all languages but Tipichu where –i is *less* expected after [tʃ] than elsewhere, whereas [ti] and [pi] fuse into outcomes in Tipichu. The outcome [tʃi] is helped by examples of tʃ→tʃi, increasing palatalization in Tapachi and Tipichi relative to Tapa and Tipi. However, it is not helped by tʃ→tʃu, which makes these examples less effective at helping palatalization, which results in palatalization rates going slightly down in Tapachu and Tipichu compared to Tapa and Tipi.

Table 5. Alveolar palatalization rates before -i in Tapa, Tipi, Tapachi, Tipichi, Tapachu and Tipichu. The rates are unaffected by the addition of tʃ→tʃu examples but increased by the addition of tʃ→tʃi examples.

	Tapa(ch{i;u})	Tipi(ch{i;u)}	
T{a;i}p(a;i}	7%	7%	=
T{a;i}p(a;i}chu	8%	6%	|
T{a;i}p(a;i}chi	14%	14%	<

Interestingly, it is crucial for parts not to be discarded after they fuse into a whole (as also argued Gluck & Bower 1988, Kapatsinski 2007, 2009b). Otherwise, no palatalization of [t] is expected in Tapachu because the outcome [tʃi] is inhibited by [t]-final singulars in this language. Palatalization can therefore

happen only if outcome [tʃ] is selected. It appears then that [tʃ] must still exist even if it fuses with –i into [tʃi] on the basis of co-occurrence.[13]

Generalizing suffix probabilities: Tapa vs. Tipi

In Kapatsinski (2009, 2010), human participants are reported to almost perfectly match the probabilities of –a vs. –i (25% -i in Tapa vs. 75% -i in Tipi) observed with [t]-final and [p]-final singulars during training (Kapatsinski 2009, chapter 3, p.54). The model successfully predicts this result, closely matching the observed probabilities. This result holds with or without copy outcomes.

However, another difference between Tapa and Tipi is that participants exposed to Tipi overgeneralize simple addition of –i to [k] (Kapatsinski 2009, p.60). The model overpredicts the use of –i with [k]-final singulars in Tapa, resulting in little to no effect of the frequency of –i use with {t;p} on the frequency of –i use with [k]. We take up this problem in the next section.

Table 6. Activation of –i vs. –a in the NDL model following training on Tapa vs. Tipi (with Copy and configural outcomes, but the results are qualitatively similar without these)

Singular ends in	Tapa Language	Tipi Language
{t;p}	24%	73%
k	83%	97%

Avoidance vs. Overgeneralization through Hebbian learning

There are two possible reasons for why participants might underuse –i with velar-final singulars: avoidance of the stem-changing –i when it would call for a stem change and overgeneralization about where –a occurs.

The first proposal, avoidance of –i, claims it may be no coincidence that –i is avoided by humans precisely when its addition would require a stem change. Rather, this is an instance of solving production difficulties by avoidance. Just as children may avoid adding suffixes that require stem changes in adult language (Do 2013), or avoid selecting words that are hard to produce (Schwarz & Leonard 1982, VIhman & Croft 2007), the experimental participants may be avoiding attaching –i when its attachment would call for a stem change. . This avoidance mechanism may operate during production at test, rather than in the course of training. In other words, having attached –i to a [k]-final singular, the speaker runs into difficulty: changing the stem is hard while a plural ending in [ki] sounds wrong. The speaker then decides to avoid this situation thereafter, fleeing a difficult choice (Albright 2003). The process can be captured by an offline feedback mechanism during the test phase: the speaker executes a choice between [ki] and [tʃi], notes that the choice was difficult and what prior choice led them to this impasse, deciding to choose differently in the future.[14] Here, choosing –i over -a as the suffix leads to the difficult choice between [ki] and [tʃi], leading participants to choose –a over –i after experiencing the difficulty in the early test trials.

[13] Note that [tʃi] does useful work in Tapachi, capturing the fact that –i is favored by [tʃ] (and vice versa), thus the alternative of claiming that [tʃi] simply does not exist in Tapachi is not tenable.

[14] Note that it is difficulty of choosing rather than goodness of the produced outcome that drives avoidance: "phew, that was hard", not "ew, that was bad", since [ka] is rated as being no better than [ki] (Kapatsinski 2009, p.79; see also Albright 2003; cf. Martin 2007).

The second proposal is that –a is overgeneralized from {t;p}-final singulars to [k]-final singulars because [t] and [p] do not constitute a natural class of stops that excludes [k]. Unlike the model, participants thus might associate –a with [stop], in addition to associating it with [lips] and [tongue blade], particularly since p(stop|a)=100%, while p(lips|a)=p(blade|a)=50%. The model does not make this connection because learning in the model is error-driven: [stop] is not predictive of the choice between –a vs. –i. For example, in Tapa, all consonants encountered in training are stops, so knowing that a consonant is a stop tells one nothing about whether –a or –i will be chosen. In contrast, knowing that the consonant is made with the tongue body makes one certain that the suffix should be –i. Thus, -i is strongly expected when the consonant is a stop made with the tongue body, i.e. [k].

Overgeneralization to natural classes containing experienced sounds associated with an outcome is expected under a Hebbian learning mechanism. Hebbian learning states that associations form on the basis of simple co-occurrence: representations of events that tend to occur at the same time are wired together (just as neurons that fire together wire together, Hebb 1949). Thus, in the present case, [stop] should become associated with –a in every language, to a language-specific extent, leading to the use of –a with stops that are not [t] or [p].

Until Kamin's (1969) demonstration of cue competition, associationist learning theory was dominated by Hebbian learning models (e.g. Bush & Mostetter 1951, Estes 1950, Hebb 1949). Kamin (1969) showed that forming an A→X association prevented formation of a B→X association if B was always encountered together with A, i.e. exposure to A→X followed by exposure to AB→X did not result in a B→X association. This *blocking effect* provided the primary motivation for RW's error-driven learning mechanism: since X is expected in the presence of A, exposure to AB (which contains A) results in no error and therefore no learning, hence nothing is learned about B.[15]

However, other explanation of the blocking effect suggest that something *is* learned about B. In particular, Miller and colleagues likewise argue that something is learned about B, and that B is actually associated with X after AB→X trials (Matzel et al. 1985, Miller & Matzel 1988). Several studies provide support for this hypothesis by showing that exposure to A alone following AB→X trials reduces or eliminates blocking, so that B alone comes to reliably activate X despite no additional training on B→X (e.g. Blaisdell et al. 1999). Based on this kind of data, Miller and colleagues have suggested that learning is actually based on simple Hebbian contiguity and all co-occurring stimuli are associated together in a non-competitive manner: B is associated with X in AB→X trials but A→X interferes with retrieval of this association. [16]

Human learning, unlike animal learning, has most often been studied within a 'causal learning' or 'contingency learning' paradigm. In a typical version of this task, participants are presented with diseases and their symptoms and learn to either predict diseases from symptoms or symptoms from diseases (e.g. Gluck & Bower 1988). Their acquired knowledge is then explicitly tested by asking them questions like "Is B a symptom of X?" or "Is X a disease caused by B?". Blocking can be observed in this paradigm when participants deny that B causes X after AB→X trials. Matute et al. (1996) showed that blocking is observed with some test questions but not others, suggesting that either different questions tap into

[15] Bayesian network models (e.g. Holyoak & Cheng 2011, Lu et al. 2008) take a similar approach, assuming that A→X 'explains away' the presence of X on AB→X trials.

[16] In addition, Mackintosh and colleagues (e.g. Mackintosh 1975) has proposed that AB→X trials teach participants not to attend to B because it makes no difference in the outcome. In support of this hypothesis, Mackintosh & Turner (1971) and Kruschke & Blair (2000) report that B is harder to associate with outcomes other than X following AB→X training, suggesting that something *is* learned about B during AB→X trials.

different stores of associations, or the same memory store is flexibly used in different ways depending on the question (e.g. whether the question is about which cue is *most* predictive of an outcome or about which cues are predictive of an outcome *at all*). Other recent work has suggested that blocking can also be reduced or even reversed in human causal learning, by putting participants under time pressure or having them perform a demanding secondary task during training (e.g. DeHower & Beckers 2003, Sternberg & McClelland 2009, Vadillo & Matute 2010).

This work can be interpreted as suggesting that both Hebbian learning and error-driven learning are used to learn co-occurrences, with Hebbian learning being less resource-demanding than error-driven learning. This interpretation is supported by neuroscientific findings that there is a distinct, localizable neural signature for prediction error (Corlett et al. 2004, den Ouden et al. 2010) and that amnesics appear to learn in a more Hebbian fashion than non-amnesic individuals (McClelland 2006). In particular, producing erroneous responses appears to reinforce these responses in amnesics even when a clear error signal is provided.

The case for overgeneralization

The languages studied in Kapatsinski (2009, 2013) do not allow us to distinguish between avoidance of the palatalizing suffix with to-be-palatalized consonants and Hebbian overgeneralization of the non-palatalizing suffix beyond the not-to-be-palatalized consonants, since the palatalizing suffix is the only suffix that should be attached to the to-be-palatalized consonants.

However, the languages studied by Stave et al. (2013), described in Table 7, do provide an opportunity to distinguish the two explanations. In these languages, palatalization is supposed to be applied before –a but not before –i. Despite the phonetic unnaturalness of the pattern, it is learned by the participants in that participants almost never palatalize before -i. As shown in Figure 3, palatalization rates before –a are little affected by the identity of the stem vowel. The *non*-palatalizing suffix –i is to be used only when the to-be-palatalized consonant is not changed, and is therefore available as a way to avoid choosing between palatalizing and not palatalizing. The target rate of –i use with [a] stems ending in the to-be-palatalized consonant is 50%, affording the opportunity to see overuse of –i (avoidance) as well as overuse of –a (overgeneralization). Here, we focus on the Velar palatalization group, for comparability with Kapatsinski (2009), 2013).

Table 7. Labial, Alveolar and Velar palatalization patterns presented to participants in Stave et al. (2013).

	Labial Palatalization	Alveolar Palatalization	Velar Palatalization
Singular	Plural	Plural	Plural
…ap	…a{tʃa;pi}	…apa	…apa
…ip	…itʃa	…ipa	…ipa
…at	…ata	…a{tʃa;ti}	…ata
…it	…ita	…itʃa	…ita
…ak	…aka	…aka	…a{tʃa;ki}
…ik	…ika	…ika	…itʃa

Figure 3. Whether the stem consonant is retained (not palatalized) before –a depending on whether the consonant is [k], which should be palatalized (ToBePal="yes"), and the identity of the stem vowel.

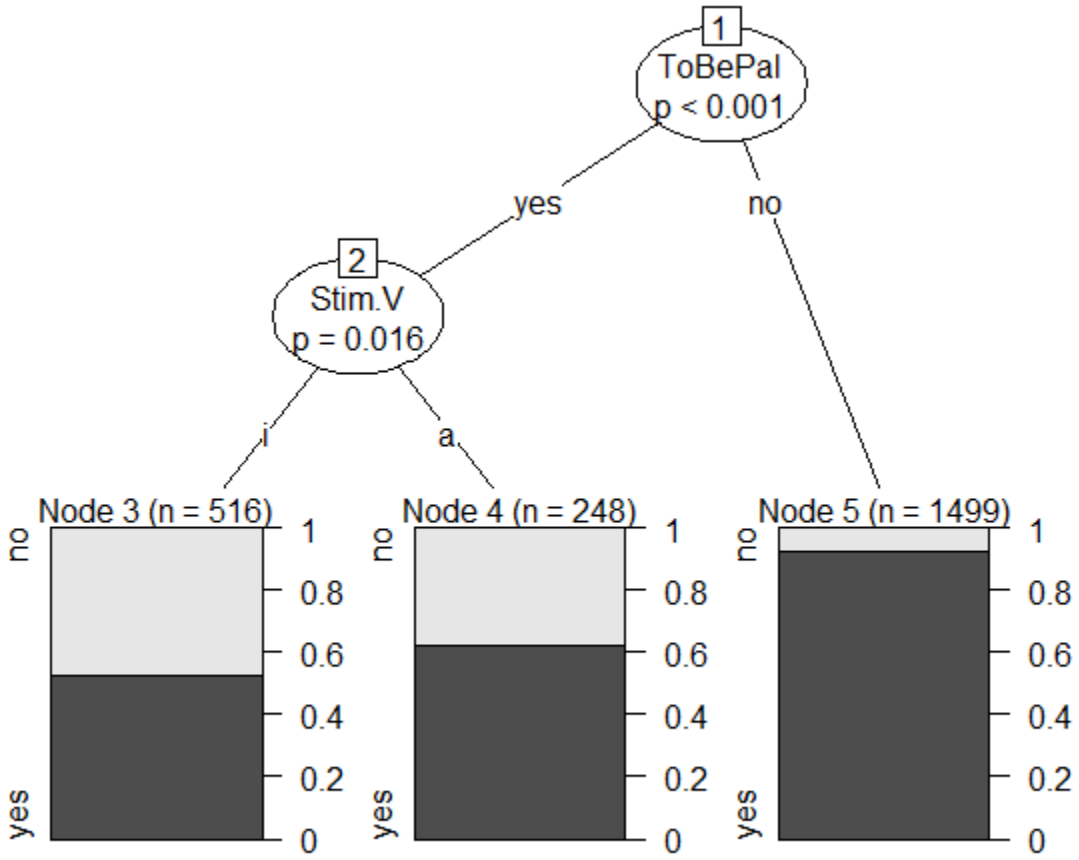
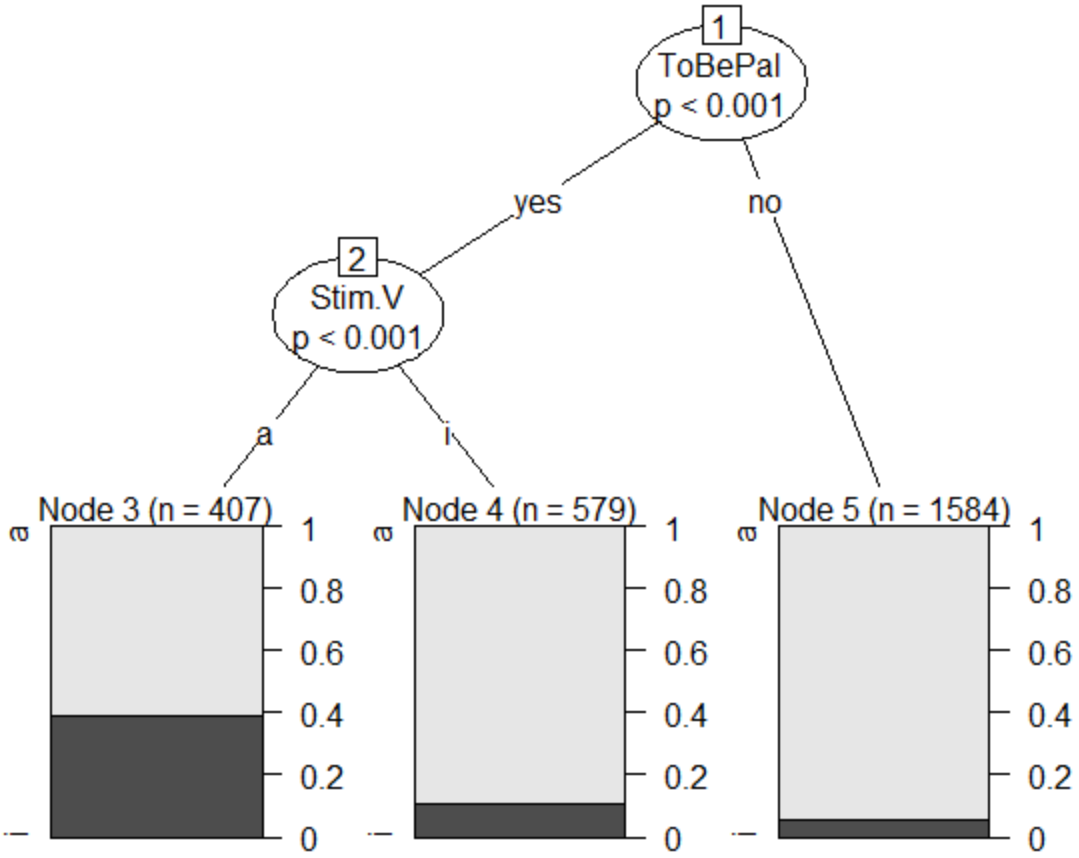


Figure 4 shows that participants were quite accurate in where they used –i vs. –a. As shown in Figure 3, palatalization was approximately equally difficult before –a regardless of the preceding stem vowel. Thus, avoidance of –a should be expected regardless of stem vowel. We clearly do not see this: while there is some overuse of –i with to-be-palatalized stems containing [i], there is also underuse of –i with to-be-palatalized stems containing [a]. Participants exposed to velar palatalization by Stave et al. (2013) did not show avoidance of the palatalizing suffix (-a) with velars, suggesting that the underuse of –i with [k]-final stems in Kapatsinski (2009, 2013) is not due to avoidance. Rather, it appears to be due to overgeneralization of –a beyond the context in which it occurs. In the case of Stave et al. (2013), we see a bit of overgeneralization for both –a and –i with no systematic avoidance of the palatalizing -a. The amount of overgeneralization is smaller as the environment for –i is well defined. Nonetheless, -i expands beyond its original environment (after [ak]) to other [k]-final stems and (less so) all stems.[17]

[17] The same pattern of overgeneralization for –i is seen for participants trained on alveolar or labial palatalization, with less underuse of –i after [a] followed by to-be-palatalized consonant. For those trained to palatalize [t] before –a, -i is used 47% after [at], 19% after [it] and 5% elsewhere. For those trained to palatalize labials before –a, -i is used 51% after [ap], 18% after [ip], 8% after [a{t;k}] and 2% after [i{t;k}].

Figure 4. Use of –a vs. –i by participants in Stave et al. (2013) depending on whether the consonant was to be palatalized (ToBePal) and whether the stimulus stem vowel (Stim.V) was [i] or [a]. An ideal learner would use –i only after to-be-palatalized consonants, and only when the stem vowel is [a], and would use –i 50% of the time in this environment.



The amount of suffix overgeneralization in Stave et al. (2013) appears reduced compared to Kapatsinski (2009, 2013). This may have to do with how well-defined the environments conditioning suffix choice are: in Kapatsinski (2013), the environment favoring –a cannot be described as a conjunction of features, while in Stave et al. (2013) the environments favoring –a and –i can (see also Moreton et al. 2015). As noted by configural association theory (Gluck & Bower 1988, Rescorla 1973), compound-cue AB→X trials do not necessarily result in strong A→X and B→X associations. A well-defined environment means that the suffix is conditioned by a consistent compound of features, so the individual features may not become associated with the suffix, reducing overgeneralization (Gluck & Bower 1988, Kurtz et al. 2013, Moreton et al. 2015).

Perseveration at test

The greater use of –i in Tipi predicted in Table 6 does not necessarily mean predicting increased production of k→ki, because k→tʃi also involves activating –i. Unlike humans (Stave et al. 2013, Smolek & Kapatsinski, Submitted)), the model feels no difficulty changing [k] into [tʃ]. Therefore, [k] is *always* expected to change into [tʃ] in both languages. This is clearly not the case for humans (e.g. 67% k→tʃ before –i in Tapa and 38% in Tipi in Kapatsinski 2009, p.56). The overprediction of stem change happens

even if the model is equipped with Copy outcomes because the model successfully learns that velar stops are not copied, acquiring an inhibitory association from [body] to $Copy_{Fin}$. The human participants either do not acquire this association (to the same extent), or are prevented from changing the stem at test by an extragrammatical tendency to perseverate during production. Given that changing the stem is judged by the participants to be more acceptable than not changing the stem (Kapatsinski 2012), we believe the association is in fact acquired during training, and participants are prevented from manifesting this knowledge fully in production by production-internal perseveration that boosts copy outcomes (Smolek & Kapatsinski, Submitted; see also Zuraw 2000 for evidence that stem changes may be preferred to no change yet seldom produced in a natural language).

If copy outcomes are boosted by a constant amount in both languages, as one would expect from a constant perseveratory tendency, the greater incidence of k→ki in Tipi is predicted. This is not because [tʃ] or [tʃi] is weakened in Tipi, nor because $Copy_{Fin}$ is strengthened: these outcomes are unaffected by the increased use of –i. Thus the proportion of trials featuring palatalization remains unchanged between the two languages. However, [i] is strengthened relative to [a]. Thus, when one looks only at the trials where –i was attached, a lower proportion of these trials will feature palatalization: even when palatalization loses to no change, -i is still likely to win over –a. Thus, for example, adding .5 to Copy outcome activations across languages predicts that, when –i is attached to a [k]-final singular, the [k] will be retained only 38% of the time in Tapa but 58% of the time in Tipi. As in humans (Kapatsinski 2009, 2010), the difference in palatalization rates before -i comes from greater use of k→ki in Tipi rather than from reduced use of k→tʃi.

The problem of tʃ-final singulars and the need for syntagmatic structure

The main problem faced by the present model is presented by [tʃ]-final singulars in Tapa and Tipi. In these languages, [tʃ]-final singulars are never presented during training. As a result, the model has no choice but to treat final [tʃ] as an unfamiliar consonant. There are good reasons to treat [tʃ] as having the place features of both coronals and dorsals: both the tongue body and the tongue blade are involved in its articulation (Yun 2006), and alternations between [t] and [tʃ] are about as common as those between [k] and [tʃ] across languages (Bateman 2007, Kochetov 2011). Having place features intermediate between the dorsal [k] and the coronal [t], [tʃ] should therefore show intermediate behavior.

In all languages except Tipichu, but especially in Tapa, [k] favors the attachment of –i while [t] favors the attachment of –a; [tʃ] is expected to be in between. In contrast, human learners attach –i to [tʃ] more than they attach it to anything else, with an 80% -i attachment rate even in Tapa where the frequencies of –i and –a are equal in training. As noted above, this may be alleviated by the additional Hebbian learning mechanism. However, it appears that not only is use of –i with [k] overpredicted but the use of –i with [tʃ] is underpredicted (if mildly, e.g. 83% -i after [k] vs. 71% after [tʃ] in Table 8b).

More problematically, the input feature [coronal] becomes associated with the output [t], predicting that a singular-final [tʃ] will often map onto [t] in the plural. Copy outcomes barely alleviate this problem, as [tʃ]'s [dorsal] feature is associated with lack of final-consonant copying. Thus, [tʃ] is expected to map onto [t] 50% of the time without copy outcomes and 40% of the time with copy outcomes. Configural outcomes do not help much: given a [tʃ]-final singular, [tʃi] is activated only slightly more than [ta] (.45 vs. .37 in Table 8c), resulting in the network turning [tʃ] into [t] 43% of the time with both copy and configural outcomes present.

Unfaithful mapping of [tʃ] onto [t] does happen in the experiment. However, it is quite rare (<10%). While it is possible to reduce the rate of tʃ→t by boosting Copy outcomes, implementing perseveration at test, a boost to $Copy_{Fin}$ that reduces the incidence of tʃ→t also reduces the incidence of t→tʃ, which is incorrectly expected to be lower than the incidence of tʃ→t even without boosting Copy outcomes (5% vs. 40% in Table 8b).[18]

Table 8. Final consonant and suffix activations in Tapa (especially problematic cells are highlighted by bolding)

a. Without copy outcomes

	a	**tʃ**	i	p	t
p	0.67	0.03	0.25	0.87	0.03
t	0.67	0.07	0.3	0.03	0.87
k	0.07	0.87	0.9	0.03	0.07
tʃ	0.38	**0.59**	0.71	-0.08	**0.59**
PL	0.34	0.13	0.24	0.33	0.13

b. With copy outcomes

	a	**tʃ**	$Copy_{Fin}$	i	p	**t**
p	0.58	0.05	0.79	0.25	0.74	0.05
t	0.6	0.09	0.8	0.29	0.04	0.75
k	0.15	0.78	0.1	0.74	0.04	0.07
tʃ	0.18	**0.45**	0.18	0.45	-0.24	**0.42**
PL	0.19	0.08	0.25	0.14	0.18	0.07

c. With configural outcomes

	a	**tʃ**	**tʃi**	**$Copy_{Fin}$**	i	p	pa	**t**	**ta**
p	0.58	0.05	0.05	0.79	0.25	0.74	0.58	0.05	0
t	0.6	0.09	0.09	0.8	0.29	0.04	0.02	0.75	0.58
k	0.15	0.78	0.78	0.1	0.74	0.04	0.04	0.07	0.11
tʃ	0.18	**0.45**	**0.45**	**0.18**	0.45	-0.24	-0.19	**0.42**	**0.37**
PL	0.19	0.08	0.08	0.25	0.14	0.18	0.14	0.07	0.05

I would suggest that a [tʃ]-final singular should not be treated as either novel or intermediate between a [k]-final singular and a [t]-final singular. After experiencing k→tʃi, learners know what to do with a [tʃ]-final singular (namely, that it should take –i, becoming [tʃi]): k→tʃi is taken to imply tʃ→tʃi.

While there is agreement on this point in the field (cf. Stave et al. 2013, White 2014), there is disagreement on the underlying reason. On the one hand, this can be seen as a special case of a proposed bias against saltation. According to this proposal, if X→Y and there is a sound Z that is in between X and Y, then X→Y is taken to imply Z→Y (Boersma 1998, Hayes & White 2015, Moreton et al. 2015, White

[18] This comparison also underscores the fact that the problem cannot be solved by assuming that the model maximizes accuracy in choosing the output, always picking the most activated outcome from the set of competitors. While this would eliminate tʃ→t, it would also eliminate t→tʃ as well as preventing the model from matching the probabilities of –i vs. –a after {p;t}.

2014, White & Sundara 2014). According to White and colleagues, this is a generalization about *changes* that relies on a knowledge base about change magnitudes: a bigger change implies a smaller change resulting in (avoidance of) the same output (see also Boersma 1998, Steriade 2008, Zuraw 2007). Empirical support for this position is provided by White's (2014) finding that participants trained on p→v would also change [b] into [v] while those trained on b→v would not change [p] into [v]. However, we should note that participants in this study were trained to criterion, and (given that p→v is harder to learn) received more training on changing the stem than participants trained on b→v.

Smolek & Kapatsinski (Submitted) did not train participants to criterion, instead using a constant number of training trials (as in Kapatsinski 2009, 2013). We examined learnability of labial (p→tʃ) vs. lingual (k→tʃ or t→tʃ) palatalization. As shown in Table 9, [t] and [k] are in between [p] and [tʃ], thus it may be expected that p→tʃ would be taken to imply t→tʃ and k→tʃ. In contrast, [t] is not between [k] and [tʃ], so k→tʃ need not imply t→tʃ. Similarly, t→tʃ need not imply k→tʃ. Thus, there should be more [t] palatalization after training on p→tʃ than on k→tʃ, and more velar palatalization after training on p→tʃ than on t→tʃ. However, there were no significant differences, contrasting with the strong preference for tʃ→tʃi after training on k→tʃi in Kapatsinski (2009, 2013).[19] Thus, tʃ→tʃi appears to be implied by X→tʃi more strongly than any other change resulting in [tʃi]. It is not simply a case of a smaller change resulting in the same output.

Table 9. Featural descriptions of English stops and [tʃ].

+lips *-tongue blade* *-tongue body*	-lips +tongue blade *-tongue body*	-lips *-tongue blade* +tongue body	-lips +tongue blade +tongue body
p	t	k	tʃ

Based on this work, we suggest that the knowledge acquired by the participants is a syntagmatic product-oriented generalization: exposure to X→tʃi increases the probability of [tʃ] preceding [i] and the probability of –i following [tʃ]. These kinds of between-segment transitional probabilities have been shown to be automatically learned even by young infants (Aslin et al. 1998, French et al. 2011, Pelucchi et al. 2008, Perruchet & Desaulty 2008) and uncontroversially play a role in language acquisition (despite not being sufficient, e.g. Gambell & Yang 2003, Johnson & Tyler 2010). The next section presents an NDL model designed to learn these kinds of generalizations, showing that it recovers the transitional probabilities between segments within plurals (see also Baayen et al. 2011) and shows how this kind of knowledge could be brought to bear in producing a novel form of a known word.

Model 2: Syntagmatic prediction

The task: Syntagmatic prediction

Unlike in the schema/paradigm model, the task for the present model is syntagmatic prediction, i.e., predicting what follows what. While it might be argued that prediction of the future given the present is better motivated than retrodiction of the past given the present (e.g. Ramscar et al. 2010), there is evidence that human language learners are sensitive to transitional probabilities in both directions. In particular, Pelucchi et al. (2008) and Perruchet & Desaulty (2008) demonstrate sensitivity to backward transitional probability in the lab. Multiple studies have also shown that backwards transitional

[19] Unfortunately, we did not test whether [tʃ]-final singulars (not presented in training) take –i at rates similar to the to-be-palatalized consonant within the same study. This remains for future research.

probability conditions reduction, suggesting that words that are highly probable given the *following* word are treated as predictable (Barth 2015, Barth & Kapatsinski 2014, Gregory et al. 1999, Jurafsky et al. 2001). This is not too surprising given that preceding words may often need to be guessed on the basis of following words in a noisy environment, and following words may sometimes be planned before words that precede them. Word sequences high in backwards transitional probability also cohere together, being unlikely to be interrupted in repetition and replacement disfluencies (Harmon & Kapatsinski 2015, Kapatsinski 2005b). Sensitivity to backward transitional probability has also been argued to be a major advantage of the TRACX model of chunking (French et al. 2011). In keeping with these results, we consider that the learner predicts following segments from preceding segments and vice versa. However, given that these probabilities are to be used when one of the segments is chosen, and the other has yet to be picked, we require the transitional probabilities (and the small units surrounding the transition) to be kept, rather than used for forming a larger chunk but discarded once the larger chunk is created.

I propose that every time the learner encounters a word, s/he tries to predict all segments of that word from all other segments. That is, exposure to a word can be modeled as a set of cue-outcome pairings, where each pairing is an attempt to predict a segment from all other segments comprising the word, e.g. blutʃi = {bl, u, i}→tʃ; {bl, u, tʃ}→i; {bl, tʃ, i}→u; {u, tʃ, i}→bl. For the sake of simplicity, we focus on just the final CV here.

I assume that the learned syntagmatic probabilities are cell-specific. Thus, a separate set of probabilities is kept for plurals and for singulars. As a result, the model's syntagmatic predictions are meaning-specific, i.e., the model is better described as learning syntagmatic relationships within a schema than as learning phonotactics. The crucial prediction of this proposal is that augmenting training with homophones of product forms should not help generate the product forms. For example, adding examples like $\text{blutʃi}_{\text{SG}}$→$\text{blutʃi}_{\text{PL}}$ should not be helping –i call forth a preceding [tʃ].

I have recently tested this hypothesis with languages in which CVCVCV sources mapped onto CVCVC products (Kapatsinski 2016). Participants were then tested on CVCV sources to see whether they add a consonant, generating the familiar product (CVCVC) or delete one, applying the familiar change. In one language, the final C of CVCVC could be any stop or fricative with equal probability. After exposure to this language, participants almost never added a consonant in forming a product. In other languages, CVCVk examples were added to training, with either the source meaning or the product meaning, potentially providing the participants with something to add. When CVCVk was paired with the source meaning, the incidence of [k]-adding did not increase. However, when CVCVk was paired with the product meaning, [k]-adding became as common as vowel deletion. Thus, [k] was only added when it was unexpectedly frequent *in the product meaning*.[20]

Simulation results

As shown in Table 10, the model learns the conditional probabilities in both directions. In Tapa, there is a 75% probability of -a following [p] or [t], 100% probability of -i following [tʃ]; 50% [p] or [t] before –a and 67% [tʃ] before –i. Now we turn to how these learned syntagmatic probabilities within products can be harnessed for constructing a product.

[20] Meaning-specificity does not appear to be true of templates: extra examples of CVCVk in the *source* meaning increased the likelihood of participants adding a consonant to CVCV (suggesting the CVCVk examples in the source supported the CVCVC template). However, they did not increase the likelihood of the added consonant being [k] (Kapatsinski 2016). The template appears to be more meaning-independent than segmental material, which may have to do with templates being more meaning-independent in first-language experience.

Table 10. Activations of outcomes (columns) from the cues (rows) that syntagmatically co-occur with them in Tapa plurals

	p	t	tʃ	a	i
p	0	0	0	0.75	0.25
t	0	0	0	0.75	0.25
tʃ	0	0	0	0	1
a	0.5	0.5	0	0	0
i	0.17	0.17	0.67	0	0

Combining syntagmatic and paradigmatic knowledge

There are two ways in which syntagmatic probabilities within products could be brought to bear on production. One possibility is that some parts of the product are constructed first, which allows them to be used as cues in constructing the rest of the product (a la rule-based theories of language, e.g. Chomsky & Halle 1968). In order for the attached –i to influence the choice of the stem-final consonant, the –i has to attach first. In order for the product [tʃ] to influence the selection of the suffix, it has to be chosen first. The alternative is to construct (or pre-activate) multiple possible products and let them compete with each other on their merits (as in Optimality Theory / Harmonic Grammar; Prince & Smolensky 2004). Thus, an input [tʃ] would lead to the construction of products ending in [tʃi] and [ta] (among others), which could then be evaluated on the basis of product-oriented syntagmatic generalizations.

The basic problem to be solved is that [tʃ] is all too frequently mapped onto [ta], rather than [tʃi], in Tapa. Note that a source [tʃ] favors product –i over –a, and product –i favors [tʃ] over [t]. Thus, one path towards decreasing the incidence of tʃ→ta and increasing the incidence of tʃ→tʃi would be to choose the suffix first, and choose what to do with the stem-final consonant later.

Why would the suffix become available first? One possibility is that suffix and stem consonant selection actually proceed in parallel but the choice of the suffix completes earlier because it is easier than the choice of the stem-final consonant given a source [tʃ].[21] The choice of the product stem-final consonant given a source [tʃ] is more uncertain than the choice of the suffix even in Tapa without copy outcomes: the entropy of the suffix choice for a source [tʃ] is 0.65 bits in Table 8a, while the entropy of final consonant choice is 1 bit. If Copy outcomes are indeed independent outcomes (i.e. copying is a choice distinct from applying an arbitrary source-product mapping that results in the same product), they do not greatly reduce the uncertainty of final consonant choice for a [tʃ]-final input: entropy of final consonant choice reduces to only 0.86 bits in Table 8b. Thus, the relative uncertainty of consonant choice compared to final vowel choice is a viable explanation for why the suffix would be chosen first, with or without copy outcomes.

Interestingly, in the presence of syntagmatic knowledge about products, the network no longer needs configural outcomes to account for the difference between Tapachi and Tapachu. The addition of tʃ→tʃi examples boosts t→tʃi in part because it increases the probability of [tʃ] before –i. The addition of tʃ→tʃu examples does not exert this influence. Furthermore, tʃ→tʃu examples make suffix choice more uncertain, making it less likely that the suffix is chosen before the final consonant of the stem is already selected.

[21] Choice difficulty is routinely assumed to translate into longer decision time (e.g. Usher & McClelland 2001).

This hurts {k;t;p}→tʃ before –i because –i favors these mappings and the absence of –i in the input makes it more likely that copying will outcompete the stem change.

Lexical cues and ignoring the stem body

In the paradigmatic model above, the identity of the stem was captured by a set of cues jointly identifying the stem body. The identity of the stem is important because there is a unique plural for every stem that, in the case of [p]-final and [t]-final stems, is not fully predictable from the identity of the final consonant: some stems take –a, while others take –i. During training, the cues comprising the stem body are therefore very informative about the choice of the final suffix when the final consonant is [p] or [t]: they fully determine suffix choice, which could otherwise be either –i or –a. However, at test, participants appear to largely rely on the identity of the final consonant to choose between –i and –a, resulting in probability matching for test stimuli ending in [p] or [t]. Thus, participants exposed to Tapa choose –a ~75% of the time while those exposed to Tipi choose –i ~75% of the time.

This would not be the case if participants were relying on the features defining the stem body, as most of the test stimuli in Kapatsinski (2009, 2013) were minimal pairs of training stimuli calling for a different outcome. These test stimuli shared the stem bodies with their experienced neighbors but differed in the identity of the stem-final consonant. For example, the test stimuli [blaɪt] and [blaɪp] were minimal pair neighbors of the training stimulus [blaɪk] corresponding to the trained plural [blaɪtʃi]. Nonetheless, participants would attach –a, rather than –i, to [blaɪp] and [blaɪt].

The model correctly learns to map particular final consonants onto [tʃ]. However, if provided with cues describing the stem body at test, it largely uses these cues to decide on the choice of the suffix. As shown in Table 11, the model makes exactly the wrong predictions regarding suffix choice if provided with the stem body as a set of cues. In Tapa, [p]-final and [t]-final singulars take –i (despite the input being 75% -a), while [k]-final singulars always take –a if sharing the stem body with a training stimulus that took –a. The only case where a [k]-final singular takes –i is observed for [buk], whose experienced neighbor [bup] also takes –i. Again, this result is inevitable in NDL, given that the cues that jointly comprise the stem body are, jointly, extremely informative during training about the choice of the suffix, unless the final consonant is [k].

Table 11. Activations of outcomes in the paradigmatic model exposed to Tapa. Even without copy outcomes the final consonants are mapped onto the right pre-suffix consonants. However, with or without copy outcomes, the choice of the suffix is strongly affected by the stem body.

	a	ch	**i**	p	t	Training neighbor(s)	Training product(s)
bl_ai_lab_pl	**0.37**	0.11	**0.54**	0.79	0.01	blaɪk	blaɪtʃ**i**
tr_u_lab_pl	**0.37**	0.11	**0.54**	0.79	0.01	truk	trutʃ**i**
sw_i_lab_pl	**0.37**	0.11	**0.54**	0.79	0.01	swik	switʃ**i**
v_or_lab_pl	**0.37**	0.11	**0.54**	0.79	0.01	vork	vortʃ**i**
bl_ai_tongue_blade_pl	**0.44**	0.13	**0.56**	0.01	0.86	blaɪk	blaɪtʃ**i**
tr_u_tongue_blade_pl	**0.44**	0.13	**0.56**	0.01	0.86	truk	trutʃ**i**
sw_i_tongue_blade_pl	**0.44**	0.13	**0.56**	0.01	0.86	swik	switʃ**i**
v_or_tongue_blade_pl	**0.44**	0.13	**0.56**	0.01	0.86	vork	vortʃ**i**
b_u_tongue_blade_pl	**0.07**	0.02	**1.01**	0.22	0.85	bup	bup**i**
fl_o_tongue_blade_pl	**1.07**	0.02	**0.01**	0.22	0.85	flop	flop**a**
gw_i_tongue_blade_pl	**1.07**	0.02	**0.01**	0.22	0.85	gwip	gwip**a**

kl_u_tongue_blade_pl	**1.07**	0.02	**0.01**	0.22	0.85	klup	klup**a**

The syntagmatic model does not help here as it suffers from the same malaise: plural-final [i] and [a] are predicted following [t] or [p] only on the basis of the stem body cues.

Table 12. Activations of outcomes in the syntagmatic model exposed to Tapa. The choice of the suffix vowel is driven by the stem body.

	a	i
bl_ai_p	0.42	**0.62**
tr_u_p	0.42	**0.62**
sw_i.s_p	0.42	**0.62**
v_or_p	0.42	**0.62**
bl_ai_t	0.43	**0.55**
tr_u_t	0.43	**0.55**
sw_i.s_t	0.43	**0.55**
v_or_t	0.43	**0.55**
b_u_ch	-0.42	**1.38**
fl_o_ch	**0.58**	0.38
gw_i.s_ch	**0.58**	0.38
kl_u_ch	**0.58**	0.38

The addition of the word identity as an additional cue, separate from the cues describing the word at the sublexical level, helps but does not help enough. This is most easily seen with the paradigmatic model. As shown in Table 13, the additional cue appropriately reduces the use of –i with [t]-final and [p]-final singulars and the use of –a with [k]-final singulars.[22] Nonetheless, Table 14 shows that the predicted proportions of –a with [t]- or [p]-final singulars are still too low to match the probabilities in the Tapa input, even if the model is augmented with lexical cues. Lexical cues capture some of the variance in suffix choice (which improves the model) but they do not capture *enough* of the variance. In fact, *all* of the variance attributed to the influence of stem body cues by the model should be attributed to the lexical cues if the model is to engage in matching the suffix probabilities given *only* the probability of the suffix given the final consonant.[23]

[22] Importantly, the additional lexical cues do no damage to the paradigmatic model's ability to choose the stem-final final consonant for the plural form, increasing the incidence of palatalization within realistic limits (to ~20% for [t]). The choice of the final consonant remains (appropriately) driven largely by the final consonant of the singular.

[23] One possible reason for why participants might ignore the stem body cues so much is that the test stage presents the participants with multiple minimal pairs differing only in the identity of the final consonant. If the participants expect such stimuli to call for different responses, they should zero in on the final consonant cues as the most informative ones.

Table 13. Changes in activations with the addition of lexical cues. In the paradigmatic model on the left, the attachment of –i to labial- and coronal-final singulars sharing bodies with palatal-final ones decreases (appropriately), and the attachment of –i to velar-final singulars sharing bodies with labial-final singulars increases. In the syntagmatic model on the right, all activation levels become less extreme but with the end effect that the incorrect differences in activation levels between –a and –i decrease.

	a	i		a	i
bl_ai_lab_pl_stop	0	-0.04	bl_ai_p	-0.17	-0.33
tr_u_lab_stop_pl	0.04	-0.04	tr_u_p	-0.16	-0.34
sw_i.s_lab_stop_pl	0.13	-0.15	sw_i.s_p	-0.12	-0.38
v_or_lab_stop_pl	0.03	-0.04	v_or_p	-0.17	-0.34
bl_ai_tongue_blade_pl_stop	0.02	-0.03	bl_ai_t	-0.18	-0.28
tr_u_tongue_blade_stop_pl	0.06	-0.04	tr_u_t	-0.17	-0.29
sw_i.s_tongue_blade_stop_pl	0.15	-0.14	sw_i.s_t	-0.13	-0.33
v_or_tongue_blade_stop_pl	0.05	-0.03	v_or_t	-0.18	-0.3
b_u_tongue_body_pl_stop	0.25	-0.3	b_u_ch	0.44	-0.86
fl_o_tongue_body_stop_pl	-0.07	0.05	fl_o_ch	-0.33	-0.14
gw_i.s_tongue_body_stop_pl	-0.15	0.09	gw_i.s_ch	-0.34	-0.09
kl_u_tongue_body_stop_pl	-0.25	0.2	kl_u_ch	-0.38	-0.06

Table 14. Activations of –a vs. –i in the paradigmatic and syntagmatic models with lexical cues added

	a	i		a	i
bl_ai_lab_pl	0.42	**0.46**	bl_ai_p	0.25	**0.29**
tr_u_lab_pl	0.46	0.46	tr_u_p	0.26	**0.28**
sw_i_lab_pl	**0.55**	0.35	sw_i.s_p	**0.3**	0.24
v_or_lab_pl	0.45	**0.46**	v_or_p	0.25	**0.28**
bl_ai_tongue_blade_pl	0.46	**0.53**	bl_ai_t	0.25	**0.27**
tr_u_tongue_blade_pl	0.5	**0.52**	tr_u_t	0.26	0.26
sw_i_tongue_blade_pl	**0.59**	0.42	sw_i.s_t	**0.3**	0.22
v_or_tongue_blade_pl	0.49	**0.53**	v_or_t	0.25	0.25
b_u_tongue_body_pl	-0.17	**1.2**	b_u_ch	0.02	**0.52**
fl_o_tongue_body_pl	0.51	**0.55**	fl_o_ch	**0.25**	0.24
gw_i_tongue_body_pl	0.43	**0.59**	gw_i.s_ch	0.24	**0.29**
kl_u_tongue_body_pl	0.33	**0.7**	kl_u_ch	0.2	**0.32**

It is important to note that insensitivity to the cues comprising the stem body as a determinant of suffix choice makes lexical cues the only source of lexical sensitivity: if a participant can learn that [bup] takes –a while [flup] takes –i, without associating [b] or [bu] with –a and [fl] or [flu] with –i, then they must be choosing the suffixes for known words based on lexical identity, not sublexical cues. Existence of lexical knowledge in the absence of sublexical associations would thus provide evidence for lexical storage and retrieval (cf. Baayen et al. 2011). We do not yet have such evidence in the present experimental paradigm: participants in Kapatsinski (2013) and Kapatsinski (2009, Chapter 4) were not presented with familiar singulars at test, while those in Kapatsinski (2009, Chapter 3) were trained on a large lexicon and

displayed no lexical knowledge. This remains an important avenue for future research.[24] Nonetheless, I am optimistic that such evidence will be found if participants are tested on old vs. new words after extensive training on a small lexicon, for two reasons.

First, arbitrary long-distance syntagmatic associations are notoriously hard to learn and appear to have little effect on suffix choice (Albright & Hayes 2003, Kapatsinski 2005). The evidence for existence of non-local sublexical syntagmatic associations in known natural language cases of lexically-specific morphophonology is somewhat ambivalent but it appears clear that such associations (if they exist are weak). Albright & Hayes (2003) suggest that their analogical model if the English past tense is misled by sensitivity to non-local associations, and that adding any sensitivity to non-local associations reduces the performance of their rule-based model. Some evidence *for* non-local associations between stem bodies and vowel suffixes in Russian verbs is provided by a wug test reported in Kapatsinski (2005). However, that study reports on a rather small data sample, the magnitude of the body identity effect is small, and no random effects are included in the analysis. In both studies, the non-local associations (if significant) appear much too weak to account for the difference between deterministic pattern selection with known words and pattern probability matching with their unknown neighbors (cf. also Smith & Minda 2000 for similar results with visual patterns, where minimal deviations from familiar patterns are treated as only 17% similar to the familiar pattern). More generally, lexical cues appears necessary to explain why phonological alternations lose productivity with age, coming to be largely restricted to words they occur in (Bybee 2001, Kapatsinski 2010).

Whether or not participants in the present studies acquire lexical cues, they do not appear to learn (or at least use) arbitrary long-distance syntagmatic dependencies between phonological units. We should expect the dependency between the stem onset and the suffix should be particularly difficult to learn, as the two are separated by two segments (Newport & Aslin 2004), and the dependency does not connect values of the same feature (Moreton 2008a, 2012). There are also as many distinct stem bodies as there are stems, which leaves the sublexical solution without even its usual advantage of simplicity (Moreton 2008b). The relationship between stem and suffix vowels may me more learnable, since both are vowels. However, unlike in previous studies of learning dependencies between vowels, which focused on harmony and disharmony, the relationship is arbitrary (cf. Finley 2008, Moreton 2008a, Peperkamp 2015, Peperkamp et al. 2006. Pycha et al. 2003, Stave et al. 2013). Furthermore, there are many distinct vowels, again depriving the 'grammatical' sublexical solution of its usual simplicity advantage over the use of whole-word cues.

Without the onset available as a cue, the model does much better in matching the suffix probabilities after [t] or [p] than either with both onset and nucleus cues, or with only nucleus cues, as shown in Table 15. However, as noted earlier, both stem onset and stem nucleus cues should be unavailable or ignored to achieve suffix probability matching given the final consonant.

[24] For example, lexical specificity in the absence of non-local sublexical associations would be demonstrated if the familiar [blaɪk] was palatalized significantly more than the unfamiliar neighbors [blɪk], [baɪk] and [laɪk] with no difference between the neighbors of [blaɪk] and non-neighbors like [zuk].

Table 15. Suffix activations in the lexical paradigmatic model without access to either onset cues (left) or nucleus cues (right)

	a	i	a	i
bl_ai_lab_pl_stop	**0.72**	0.2	0.33	**0.57**
tr_u_lab_stop_pl	**0.54**	0.41	**0.5**	0.4
sw_i.s_lab_stop_pl	**0.79**	0.15	**0.5**	0.4
v_or_lab_stop_pl	**0.52**	0.45	**0.53**	0.42
bl_ai_tongue_blade_pl_stop	**0.74**	0.25	0.34	**0.67**
tr_u_tongue_blade_stop_pl	**0.57**	0.47	**0.51**	0.5
sw_i.s_tongue_blade_stop_pl	**0.81**	0.21	**0.51**	0.5
v_or_tongue_blade_stop_pl	**0.54**	0.51	**0.54**	0.52
b_u_tongue_body_pl_stop	-0.03	**1.02**	-0.12	**1.13**
fl_o_tongue_body_stop_pl	0.23	**0.74**	0.46	**0.6**
gw_i.s_tongue_body_stop_pl	0.22	**0.76**	0.17	**0.77**
kl_u_tongue_body_stop_pl	-0.03	**1.02**	0.38	**0.63**

Labial vs. Lingual palatalization: A bias against large changes

Stave et al. (2013) and Smolek & Kapatsinski (Submitted) focused on demonstrating a learning bias against labial palatalization by comparing the learning of languages in Tables 7 and 16. In both studies, participants palatalized labials following training on labial palatalization less than they palatalized non-labials following training on alveolar or velar palatalization. This result is argued to follow from a bias against large changes rather than against unnatural rules (changes in context), since it is observed both before –i (a natural context for palatalization) and between [a]'s, an unnatural context (see also Skoruppa et al. 2011). The bias does not appear to be due to preferences against certain stop-vowel sequences: faithful illegal outcomes were disliked equally, whether they contained a labial, a coronal or a velar. For example, participants exposed to the Labial palatalization language in Table 7 came to dislike plurals ending in [pa] as much as those exposed to Alveolar palatalization language came to dislike plurals ending in [ta].

Table 16. Labial, Alveolar and Velar palatalization patterns presented to participants in Smolek & Kapatsinski (Submitted).

	Labial Palatalization	Alveolar Palatalization	Velar Palataization
Singular	Plural	Plural	Plural
…p	…tʃi	…{pi;pa}	…{pi;pa}
…b	…dʒi	…{bi;ba}	…{bi;ba}
…t	…{ti;ta}	…tʃi	…{ti;ta}
…d	…{di;da}	…dʒi	…{di;da}
…k	…{ki;ka}	…{ki;ka}	…tʃi
…g	…{gi;ga}	…{gi;ga}	…dʒi

Smolek & Kapatsinski (Submitted) attributed the findings to a bias against associating dissimilar units. A bias against associating dissimilar stimuli has been documented in the associative learning literature by Rescorla & Furrow (1977) and Rescorla (1986). Rescorla & Furrow studied what they call 'second-order

conditioning', in which a conditioned stimulus (like a light or a tone) is associated with another conditioned stimulus that had previously been associated with an unconditioned stimulus (like food or electric shock). Similarity between the conditioned stimuli resulted in faster learning. The bias against changing the input appears to be stronger in production than in judgment (see also Zuraw 2000). Particularly, labial palatalization is judged as acceptable but seldom produced, while velar and coronal palatalization are judged acceptable about as often as they are produced. Based on this finding, we argue that it is articulatory similarity, rather than perceptual similarity that affects associability in the present case; particularly, overlap in the active articulators involved in producing the to-be-associated sounds. There is a good neurological motivation for such a bias in that more synaptic modification is required to associate together representations that are relatively far from each other in the brain (see also Kapatsinski 2011, Moreton 2008a, Warker & Dell 2006). The motor cortex is largely segregated by articulator, hence gestures produced with different articulators should have rather dissimilar/neurally distant representations within the production system.

Our model, even when equipped with copy outcomes, does not have this bias and should therefore be equally good at learning labial and lingual palatalization. This was indeed the case: with the stimuli used by Smolek & Lapatsinski (Submitted), the model predicted that to-be-palatalized should always be palatalized, while the not-to-be-palatalized should never be palatalized in all conditions. This failure of the model supports the existence of a *bias* against changing labials into alveopalatals. Fortunately, the bias can be implemented in the model very directly, by weakening p→tʃ connections compared to t→tʃ and k→tʃ connections or decreasing the learning rate on p→tʃ connections.[25] With this addition, the model successfully captures the main finding of Stave et al. (2013) and Smolek & Kapatsinski (Submitted), namely that labials are changed into alveopalatals less than non-labials are.

One might be concerned that the unbiased model fails to show a difference between conditions because its performance is so good (i.e., because of a ceiling effect). We therefore considered a number of alternative ways of bringing the model's performance off the ceiling. First, the model's performance was unaffected by replacing some training experiences by 'product-oriented' training experiences in which the input phonological features and copy outcomes were missed by the learner (unless 100% of experiences with some input type were replaced). It was also unaffected by milder versions of this impairment, in which some input phonological features were missed on some trials, e.g. [k] and [t] were sometimes perceived as simply [lingual], becoming identical on those trials, or [p] was perceived as merely a stop of unknown [place].

Second, we assumed that on some training trials a plural ending in [tʃi] is perceived to correspond to a singular ending in [tʃ]. For example, hearing the plural [butʃi], the learner might erroneously recall having heard the singular [butʃ]. Some support for this hypothesis is presented by Kapatsinski (2009, pp.51-52), who asked participants to repeat back singular-plural pairs during training. If responding after hearing the plural, the participants would often repeat pairs like [buk]-[butʃi] as [butʃ]-[butʃi] (hear [buk] → hear [butʃi] → repeat [butʃ butʃi]). The error was clearly a recall error because participants almost never repeated [buk] as [butʃ] with immediate repetition (hear [buk] → repeat [buk] → hear [butʃi] → repeat

[25] In the version of NDL used here, which models the behavior of the learner after it reaches equilibrium, only the former option would produce a bias. However, in reality the human learners have not learned all they can learn from the data. Since they are not at equilibrium, the model should not be either, and learning rate should matter. The learning rate implementation of the bias would mean that the bias is soft, which appears more realistic because labial palatalization can be productive in natural languages (Bennett & Braver 2015): given enough data, the learner should therefore be able to learn labial palatalization as well as they can learn lingual palatalization.

[butʃi]). It is possible that an input [p] would be particularly unlikely to be recalled when [tʃi] is perceived if the input activation is driven in part by similarity to the output.

Recall errors had very little effect on the paradigmatic model's use of p→tʃ vs. p→p unless all training instances of p→tʃ were misremembered as tʃ→tʃ. The reason is that the misperceived tʃ→tʃi examples do not result in knowledge that is applied to [p]-final singulars during test: at test, the singular is correctly perceived as ending in a labial, hence the p→tʃi examples misperceived as tʃ→tʃi do not impact their palatalization rates. While it is possible to change this by assuming that a [p] is treated as simply a stop with unknown place some proportion of the time during test, this would only produce an asymmetry between labial and non-labial palatalization if labiality were missed much more often than coronality or velarity during test. However, Smolek & Kapatsinski (Submitted) report that participants repeated singulars correctly at test. It also does not appear feasible to claim that [labial] is ignored at test because it is perceived to be irrelevant: without a bias against changing labials, the use of [labial] is necessary for the model to account for the low rates of labial palatalization in Kapatsinski (2009, 2013) and in the non-labial training conditions of Smolek & Kapatsinski (Submitted) and Stave et al. (2013). Without extensive misperception at test, or inattention to [labial], input recall errors during training have little effect on the paradigmatic model. The syntagmatic model is unaffected by misremembering p→tʃi as tʃ→tʃi since the product is the same in both cases.

One type of training error did have an effect on performance but not in a way that would allow the model to capture the finding that a to-be-palatalized [p] is changed into [tʃ] less than a to-be-palatalized [k] or [t] is. First, if we assume that [ki] and [ti] may often be misperceived, or misremembered as [tʃi] (Guion 1998), labial palatalization is predicted to overgeneralize to linguals but lingual palatalization is predicted not to generalize to labials. This does bring the paradigmatic model closer to accounting for the human data. However, it does not capture the effect of the place of articulation on palatalization rates of *to-be-palatalized* consonants. The main difference between labial and non-labial training in both papers was in the palatalization rates of to-be-palatalized consonants, not in the amount of overgeneralization (Smolek & Kapatsinski, Submitted). This is not expected by the paradigmatic model with misperception because misperception would large affect palatalization rates of not-to-be-palatalized consonants. Even more problematically, the syntagmatic model would expect more palatalization following training on labial palatalization than following training on lingual palatalization, contrary to the data: labial palatalization would feature [ti] and [ki] misperceived as [tʃi], while lingual training would feature correctly perceived [pi]. Thus p(tʃi) would be higher after labial training. Accounting for the reduced rate of palatalization in production following training on p→tʃi compared to t→tʃi or k→tʃi thus appears to require a bias against p→tʃ. Fortunately, the architecture of our model provides us with a natural locus for this bias.

A comparison to Clamoring for Blends

Kapatsinski (2013) used much of the data discussed here to argue for a theory of morphophonology that combines aspects of Bybee's (1985, 2001) Network Theory and Legendre & Smolensky's Harmonic Grammar (Legendre et al. 1990, Smolensky & Legendre 2006). Network Theory donated product-oriented schemas, while Harmonic Grammar donated the architecture along with faithfulness constraints.

Schemas are extracted from the experienced lexicon by means of conditional inference trees like the ones in Figures 2-3. The trees were tasked with detecting sets of features overattested in a particular paradigm cell. Any such set of features is a schema. Specifically, the trees aim to predict the occurrence of every form in a paradigm cell from the features of that form. Schemas are paths through the tree going down from the root, where a path is a set of informative feature values. In a conditional inference tree, the most

informative feature is placed at the top, and therefore at least one of its values is included in every schema. Other features are added to the tree (resulting in additional schemas) if they help predict form occurrence. All downward paths through the conditional inference tree predicting occurrence of a form in the cell that are attested in at least one word are said to be schemas, which are weighted by type frequency (the number of forms that support them).

Faithfulness constraints of the form "keep a certain chunk" are associated with specific chunks (articulatory units) comprising the input/base form. The weight of a constraint increases whenever it is observed to be obeyed and decreases whenever it is observed to be disobeyed. A stem change is produced whenever a schema demanding the outcome of the change is incompatible with the input, e.g. it demands that the plural end in [tʃi] while the input singular ends in [p], and is stronger than the relevant faithfulness constraint (here, "keep [p]").

As in Harmonic Grammar and Optimality Theory, the speaker constructs multiple candidate outputs, which are blends of the input and a schema associated with the to-be-expressed meaning. The faithfulness constraints and schemas obeyed by a candidate blend then vote ("clamor") for it to be produced. Because of this process, the overall model was named "Clamoring for Blends" or C4B. Candidate blends are chosen for production with a probability proportional to the amount of support they have from their schematic and chunky supporters.

As used here, NDL has counterparts to both schemas and faithfulness constraints. Source-oriented schemas are represented by connections linking form inputs to segmental outcomes. Product-oriented schemas are comprised of meaning-form connections and syntagmatic form-form connections within semantically-defined subsets of the lexicon. However, NDL does away with the C4B assumption that forming a paradigmatic mapping requires one to first extract two schemas to map to each other. First-order and second-order schemas are learned in parallel and are both simply predictive dependencies, albeit involving different predictors. The difference in how easy the various connection types are to form is reduced to how likely the different types of input are to come to mind when an output is perceived. For example, the meaning is presumably available almost every time, while other forms of the same word may be brought to mind only rarely. Faithfulness is represented by connections to Copy outcomes, which are again formally identical to other connections. The strengths of these connections are adjusted in exactly the same way as the strengths of other connections, and they compete for predictive power with these other connections. Their strengths are therefore easy to compare: they are on the same scale.

One of the primary motivations behind work in NDL is the desire to "do without" unnecessary mechanisms (Baayen 2015). NDL is in many ways a much simpler model than C4B. The rather convoluted process of schema extraction is replaced by the much simpler process of fusion between co-occurring units that is of proven value in extracting collocations (Gries 2013) and is supported by work on human statistical learning (Aslin et al. 1998, Pelucchi et al. 2008, Perruchet & Desaulty 2008). The fusion process in NDL is truly product-oriented: it looks only on the existing forms that share a meaning. In contrast, the schema extraction process in C4B requires either construction of phonotactically possible forms that happen not to exist in a particular paradigm cells to compare with existing forms, or comparisons between existing forms across cells. Otherwise, the conditional inference tree has nothing to predict: within the set of existing forms within a paradigm cell, the value of the dependent variable 'occurrence in the cell' is always 1. Finally, NDL does not necessarily require the construction of multiple candidate outputs and subsequent competition among these outputs: only one output is ever generated.

Both C4B and NDL are able to capture much of the empirical data in Kapatsinski (2009, 2012, 2013). However, only NDL is able to predict that tʃ→tʃi examples can help t→tʃi over t→ti while helping k→ki

over k→tʃi (when stem-sharing forms are presented next to each other), as in Kapatsinski (2009, Chapter 3; Yin & White 2016). According to C4B, these examples help k→tʃi less than they help t→tʃi. The [k]-final input leads to competition between a [tʃi]-final candidate and a [ki]-final candidate, the first of which is vastly stronger than the second even without tʃ→tʃi examples. Thus additional help the [tʃi]-final candidate gets from a stronger "plurals must end in [tʃi]" schema makes little difference to the competition. However, C4B's version of faithfulness is specific to gestures and not prosodic positions. Examples of tʃ→tʃi provide no support for "keep [k]" and therefore are not expected to support k→ki.

The bias against large stem changes documented by Skoruppa et al. (2011), Stave et al. (2013), and Smolek & Kapatsinski (Submitted) finds a much more natural home in NDL than in C4B. In NDL, this bias has to do with the learnability of a paradigmatic association linking the alternating segments. It has been known for a long time that similarity influences associability (Rescorla & Furrow 1977, Rescorla 1986; see also Kapatsinski 2011, Moreton 2008a, Warker & Dell 2006), so it is not too surprising to see this bias emerge here as well. C4B does not have paradigmatic associations, so the only home for the bias is in the relative weights of "keep [p]" vs. "keep [t]" or "keep [k]". However, it is not clear that labials are *overall* more changeable than other consonants, regardless of what they are changed into. For example, White (2013: 145) reports that participants more readily changes labial stops into labiodental fricatives than they changes coronal stops into coronal fricatives. It seems that the bias is about the size of the change, and not about changeability of the input gestural chunk.

Limitations and Future Directions

First, the model as formulated here has been applied to only a few datasets and a single test task. It is important to also apply the model to other datasets, including morphophonological patterns from a variety of natural languages. Baayen et al.'s (2011) results on applying NDL to learning large lexical databases suggest that this kind of scaling up should be possible.

The model also provides intuitive ways of capturing individual differences in grammatical knowledge and grammatical performance. For example, some speakers (including children and individuals high on the autism spectrum) may perseverate more than others (Dell et al. 1997, Rehfeldt & Chambers 2003, Smith et al. 1999), which would result in a lower ability to change the stem by increasing the strengths of copy outcomes (for children, see Kerkhoff 2007). Weak central coherence, also associated with autism, may result in a reduced tendency for using configural cues and outcomes (Plaisted et al. 2003). According to the paradigmatic model, this should increasing the likelihood of probability matching rather than exaggerating likelihoods of outcomes in the presence of cues that favor them (as if they apply the representativeness heuristic, cf. Morsanyi et al. 2009). Low working memory or executive control would disrupt the ability to acquire arbitrary paradigmatic associations potentially leading to collapsing gender classes in languages with multiple genders.

The current application of NDL examined the behavior of the model at equilibrium (Danks 2003). It is important to see how well the model captures trial-by-trial learning; particularly, within the rather extended test sessions. Dependencies between trials suggesting between-trial perseveration have been noted anecdotally in the literature (Bickel et al. 2007, Caballero 2010, Lobben 1991). If substantiated by a more formal investigation, they may need to be incorporated into the model by treating preceding test trials as additional inputs.

The balance of power between schematic, paradigmatic and syntagmatic structure deserves more work. Free parameters can be associated with weights of meaning→form, paradigmatic form→form and syntagmatic form→form connections, allowing us to determine what kind of structure is used more by

what kind of learner. Similarly, the balance of power between Hebbian and error-driven learning remains an open issue (McClelland 2006).

Order of selection is an area about which little is known. For example, the structure of the model allows for biases in favor of selecting the suffix first or the deciding on the shape of the stem first. It is not clear whether there is a general bias affecting order of selection (e.g. Chomsky & Halle 1965, Caballero & Inkelas 2013), or whether selection is driven solely by choice difficulty, and indeed whether the assumption that only a single output is constructed can be maintained.

The model relies on prosodic templates to represent inputs and outputs (Vihman & Croft 2007). As part of trial-by-trial learning and scaling up to natural-language datasets, it will be important to grow the template as additional words with novel prosodic structure are encountered, since it is unrealistic to imagine that the template is set a priori. The primary challenge in growing the template is to do so without losing generalizations that are made about the paradigmatic mappings and likelihood of copying in various templatic positions. It will also be necessary to elucidate the relationship between templates and schemas. Preliminary work (Kapatsinski 2016) suggests that prosodic templates may be more meaning-independent, becoming stronger when experienced in any meaning, than the schemas that are used to fill them (which grow stronger only when experienced in the same meaning).

The model lacks a mechanism for selective attention to informative features, something that is crucial for language learning (as argued above; see also Kapatsinski 2014). Future work should explore mechanisms for endowing the learner with selective attention, e.g. by learning cue weights prioritizing cues, or even *dimensions* – sets of mutually exclusive cues – that have been more informative in recent experience (Kruschke 1992, 2006).

The broader context: The nature of learning

Most importantly, learning theorists are not unified behind RW. Many possible difficulties for the theory have been identified since 1972, one of which (lack of selective attention) we have also encountered in the present dataset (see also Kruschke 1992). On the broadest level, there is an ongoing debate between associationist approaches like RW and cognitive approaches, including causal theory (Waldmann 2000, Waldmann & Holyoak 1992), the closely related Bayesian network approach (Courville et al. 2006, Gopnik et al. 2004, Kruschke 2008, Shanks 2006), and even the proposal that associative learning is propositional in nature (Mitchell et al. 2009). My choice of the associationist approach to model phonology is largely motivated by 1) the relatively effortless way in which the "brain-style" computations of this approach (Rumelhart 1990) allow for integrating description at the level of behavior and the level of neural activity, and 2) the ability of this approach to account for learning across species in a unified fashion, suggesting that the proposed principles are basic to all neural systems.

As we saw with the bias against large changes, the fact that associationism allows for description of the mind with the same basic vocabulary as description of the brain allows for explaining the biases of the system in a mechanistic way. The to-be-associated representations map onto places in the brain and our ability to learn a correlation between the represented events, stimuli or actions should depend directly on how plastic the relevant neural connections are and how much the brain needs to change to match the environmental statistics.

While it is possible (and even likely) that there are multiple complementary learning systems affecting human behavior in any learning task (e.g. McClelland et al. 1995, van Osellaer et al. 2004), it is clear that associative learning is involved, as we can see it operating in simple organisms like snails and mollusks whose neural architecture is well understood and can be examined directly (Castro & Waaserman 2009,

Hall 2009, Hawkins & Kandel 1984, Prados et al. 2013). It is rather implausible to assume that a mollusk is engaged in forming propositions (Mitchell et al. 2009) or doing sophisticated sampling from a vast hypothesis space as suggested by the Bayesian approach understood mechanistically (e.g. Courville et al. 2006).

Learning must be near-Bayesian to be adaptive: it is not of much use for the organism to learn if learning will not pick up on environmental statistics (e.g. Anderson 1990, Shanks 1995). Because of this, Bayesian analyses of environmental statistics provide a valuable source of explanation for characteristics of behavior. However, they do not provide a mechanism for acquiring the behavior (e.g. Shanks 1995) nor do they provide an explanation for the biases of the system, ascribing them all to the prior. Language is rather unique in that it is created by its learners, hence cross-linguistic statistics are strongly reflective of the learners' biases (see Griffiths et al. 2008 for argumentation that iterated learning converges on the learners' priors). As linguists, we are interested in why languages are the way they are (e.g. Chomsky & Halle 1965). The answer 'they are this way because of the prior' is as unsatisfactory as the answer 'they are this way because of Universal Grammar' (and is indeed very nearly the same answer). What we need to do is explain where the prior comes from.[26] Mechanistic accounts provided by neural networks and associationist learning theory help us start addressing this question because they are explicit about the parts of the system that give rise to the various biases affecting the modeled behavior and its acquisition (e.g. Bays 2015, McClelland et al. 1995).

What makes us human

As best we can tell, language is unique to humans. One might, therefore, be tempted to dismiss species-general learning mechanisms as a way to acquire language (e.g. Chomsky & Halle 1965). There must be some abilities to language acquisition that non-human animals do not have. An endorsement of associationist learning theory does not amount to saying that human learning, and human language learning, is the same as learning of other behaviors by other animals. It merely makes all kinds of learning comparable by expressing it using the same vocabulary of nodes and associations.

The task that the present model is trying to accomplish is of course specifically human. It is indeed difficult to imagine possible analogs to learned paradigmatic structure outside of language. One possible, though imperfect, analogy might be to conceive of words as tools used to accomplish particular goals (Zipf 1949), and learning to produce novel forms of known words as learning to modify tools in specific ways for specific purposes. According to the current model, the learner comes to know 1) what parts the tool should have to accomplish the desired goal, and 2) what parts of other tools should be reshaped into the goal-appropriate parts.

RW claims that learning is predictive (and therefore error-driven). Given this assumption, one source of differences between animal and human learning lies in what is being predicted from what. In part, these differences come from the species-specific goals, attentional biases and representations. In addition, prediction in non-human animals appears to be tightly time-bound: future events are predicted from current events (e.g. Ramscar et al. 2010). On the other hand, human learning may be able to partially transcend time. For example, in language learning, the learner is working towards being both a perceiver and a producer. They want to be able to move in both directions, producing form from meaning and recovering meaning from form. There is some evidence that parents usually name objects the child is already looking at (Yu & Smith 2012). Thus the child perceives the object before perceiving its name.

[26] Here, Moreton's 2008 distinction between channel bias and analytic bias is on the right track but too coarse-grained.

Under the traditional interpretation of RW, where cues are antecedents of outcomes, the NDL model exposed to objects before names would learn to predict names from features of objects but not vice versa (Ramscar et al. 2010). However, Harmon & Kapatsinski (2015) show that language learners seem to learn predictive dependencies in both directions. For example, they learn to both predict that small objects predict a word bearing the suffix *–nem* and that a word with the suffix *–nem* should be mapped onto a small object. The data are well captured by NDL, as long as it assumed that learners are both predicting forms from meanings and retrodicting meanings from forms. Acquisition of associations accurately reflecting environmental probabilities in both directions is also documented in human causal learning (Matute et al. 1996).

According to ideomotor theory, bidirectional association learning has been argued to be crucial to learn the relationship between an action and its sensory outcomes for motor imitation (Elsner & Hommel 2001, Shin et al. 2010), which is of crucial importance for learning to produce speech (see also Boersma 1998). In learning to produce speech, the learner perceives the sensory consequences of another person's motor action. These sensory consequences serve as the perceptual target for speech production, as evidenced by the finding that speakers compensate for perturbation of perceived acoustics (Perkell 2012, Purcell & Munhall 2005, VIllacorta et al. 2007). To produce the target, the speaker needs to be able to activate the correct action(s) when the target is activated. This appears to require comparing perceptual consequences of produced actions to the perceptual goal in order to train the association going from these perceptual consequences to the relevant motor representations (see also Redford 2015), connections going back in time during learning.

Another way in which human associative learning is likely to be more powerful than non-human associative learning may come down to processing resources available for predictive learning. In particular, acquisition of paradigmatic dependencies may depend on being able to make predictions about other forms of experienced words and then waiting patiently to experience these other forms while maintaining the predictions in memory. This would obviously require rather extensive working memory resources. Alternatively, paradigmatic associative learning may be accomplished by recalling other forms of a word when a word is perceived and attempting to predict the perceived word from these other wordforms. This mechanism would require rather sophisticated cognitive control abilities: as there is little immediate motivation to attempt to predict a word that is actually being perceived, something needs to direct the learner to engage in this kind of prediction (the frontal lobe being an obvious candidate). In either case, paradigmatic association learning may require resources that are well beyond the capacity of non-human animals (and maybe even young children) despite being accomplished by the species-general mechanisms of Hebbian and error-driven association learning.

Conclusion

It has long been a major goal of linguistic theory to specify an acquisition mechanism for deriving a grammar from linguistic experience. I have argued that this goal of linguistic theory is fruitfully approached by building on the foundation of domain-general learning mechanisms that are inherent to any neural system. Here, we have focused on a specific part of the grammar, namely productive morphophonology, which allows speakers to generate novel forms of known words, even words they hear for the first time.

I have argued that speakers accomplish this task by activating the meaning to be expressed as well as known forms of the word. To derive the novel form, they make use of word-sized meaningful prosodic templates (a la Vihman & Croft 2007) and other form-meaning associations (Bybee 1985, 2001, Goldberg 2003), meaning-specific syntagmatic dependencies, phonologized perseveration (knowing what and when

to copy from the activated forms of the word, a la the positional faithfulness constraints of Beckman 1999), and arbitrary paradigmatic associations between phonological units (a la the rules of Chomsky & Halle 1968 without the need to distinguish between change and context). The novel form of a known word is derived by chunks copied from the input and chunks activated by the meaning and/or by the input racing in parallel to fill out a prosodic template. Once part of a template is filled, it may become available to bias the filling of the remaining parts of the template through syntagmatic co-occurrence relations.

With these basic building blocks in place, the acquisition mechanism for productive morphophonology is well approximated by classic theories of associative learning (e.g. Rescorla & Wagner 1972). The present modeling effort is able to capture both product-oriented rule conspiracies (Bybee 2001, Kapatsinski 2013, Kenstowicz 1970), and arbitrary paradigmatic mappings (Becker & Gouskova 2014, Booij 2010, Nesset 2008, Pierrehumbert 2006). It also (arguably) provides the simplest account of the data so far. Compared to product-oriented models, it does away with construction and evaluation of multiple candidate outputs, separate learning mechanisms for schema extraction and paradigm learning, and separate learning stages for extracting first- and second-order schemas. Compared to source-oriented models, it does away with the need to split words into changes and the contexts in which they occur, and relaxes the tacit assumption that learning phonology always proceeds by comparing two morphologically related wordforms.

More importantly, by expressing the acquisition of phonology in the common vocabulary of associative learning, Learning Theoretic phonology opens up avenues for comparisons across domains, and even species, allowing work on acquisition of phonology to inform general learning theory and vice versa (cf. Moreton et al. 2015 for a related research program connecting phonological theory to work on concept learning). For example, while showing that much of morphophonology can be acquired using the simple learning mechanism of Rescorla & Wagner (1972), the present modeling effort also exposes its insufficiency. Future work should explore how error-driven discriminative learning interacts with other types of learning necessary for language acquisition, including Hebbian learning (Hebb 1949), learned (in)attention (Kruschke 1992), and distributional learning in continuous space (Olejarczuk & Kapatsinski 2016) that creates the very categories to be associated or attended to.

References:

Albright, A. (2008). Explaining universal tendencies and language particulars in analogical change. In J. Good (Ed.), *Linguistic universals and language change*, 144-181. OUP.

Albright, A. (2003). A quantitative study of Spanish paradigm gaps. In *West coast conference on formal linguistics 22 proceedings* (pp. 1-14).

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, *90*(2), 119-161.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409-429.

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, *122*(3), 292-305.

Arppe, A., Hendrix, P., Milin, P., Baayen, R. H., & Shaoul, C. (2014). ndl: Naive Discriminative Learning. R package version 0.2.16. http://CRAN.R-project.org/package=ndl

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321-324.

Baayen, R. H. (2015). doing without. Paper presented at American International Morphology Meeting, Amherst, MA.

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language*, *37*(1), 94-117.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438-481.

Barth, D. G. (2015). *To have and to be: Function word reduction in child speech, child directed speech and inter-adult speech* (Doctoral dissertation, University of Oregon).

Barth, D., & Kapatsinski, V. (2014). A multimodel inference approach to categorical variant choice: construction, priming and frequency effects on the choice between full and contracted forms of am, are and is. *Corpus Linguistics & Linguistic Theory*.

Bateman, N. (2007). *A crosslinguistic investigation of palatalization* (Doctoral dissertation, UCSD).

Bates, E., & MacWhinney, B. (1989). Functionalism and the competition model. In MacWhinney, B., & Bates, E. (Eds.) *The crosslinguistic study of sentence processing*, 3-73. CUP.

Bays, P. M. (2015). Spikes not slots: noise in neural populations limits working memory. *Trends in Cognitive Sciences*, *19*(8), 431-438.

Becker, M., & Gouskova, M. (2012). Source-oriented generalizations as grammar inference in Russian vowel deletion. Ms. Indiana University and NYU. Available on LingBuzz at http://ling. auf. net/lingbuzz/001622.

Becker, M., Nevins, A., & Levine, J. (2012). Asymmetries in generalizing alternations to and from initial syllables. *Language*, *88*(2), 231-268.

Beckman, J. N. (1999). *Positional faithfulness: an optimality theoretic treatment of phonological asymmetries*. Routledge.

Beekhuizen, B., Bod, R., & Zuidema, W. (2013). Three design principles of language: The search for parsimony in redundancy. *Language and speech*, *56*(3), 265-290.

Bennett, W. G., & Braver, A. (2015). Phonology or morphology: Inter-speaker differences in Xhosa labial palatalization. Paper presented at the Annual Meeting of the Linguistic Society of America, Portland, OR.

Benua, L. (1997). *Transderivational identity: Phonological relations between words*. (Doctoral Dissertation, University of Massachusetts, Amherst).

Berg, T. (1998). *Linguistic structure and change: An explanation from language processing*. Oxford University Press.

Bergen, B. K. (2004). The psychological reality of phonaesthemes. *Language*, 80(2), 290-311.

Bickel, B., Banjade, G., Gaenszle, M., Lieven, E., Paudyal, N. P., Rai, I. P., ... & Stoll, S. (2007). Free prefix ordering in Chintang. *Language*, 43-73.

Blaisdell, A. P., Gunther, L. M., & Miller, R. R. (1999). Recovery from blocking achieved by extinguishing the blocking CS. *Animal Learning & Behavior*, *27*(1), 63-76.

Boersma, P. (1998). *Functional phonology*. (Doctoral dissertation, University of Amsterdam).

Boersma, P., & Hayes, B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic inquiry*, *32*(1), 45-86.

Booij, G. (2010). Construction morphology. *Language and Linguistics Compass*, *4*(7), 543-555.

Brooks, P. J., Braine, M. D., Catalano, L., Brody, R. E., & Sudhalter, V. (1993). Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. *Journal of memory and language*, *32*(1), 76-95.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, *6*(2), 201-251.

Bush, R. R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review*, *58*(5), 313-323.

Bybee, J. (2002). Sequentiality as the basis of constituent structure. In Givón, T., & Malle, B. F. (Eds.). *The evolution of language out of pre-language*, 109-134. John Benjamins.

Bybee, J. (2001). *Phonology and language use*. Cambridge University Press.

Bybee, J. (1985). *Morphology: A study of the relation between meaning and form*. John Benjamins.

Bybee, J., & McClelland, J. L. (2005). Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review*, *22*(2-4), 381-410.

Caballero, G. (2010). Scope, phonology and morphology in an agglutinating language: Choguita Rarámuri (Tarahumara) variable suffix ordering. *Morphology*, *20*(1), 165-204.
Caballero, G., & Inkelas, S. (2013). Word construction: Tracing an optimal path through the lexicon. *Morphology*, *23*(2), 103-143.
Castro, L., & Wasserman, E. A. (2009). Rats and infants as propositional reasoners: A plausible possibility? *Behavioral & Brain Sciences*, *32*, 203-204.
Chomsky, N. (1959). A review of BF Skinner's Verbal Behavior. *Language*, *35*(1), 26-58.
Chomsky, N. & Halle, M. (1968). *The sound pattern of English*. Harper & Row.
Chomsky, N., & Halle, M. (1965). Some controversial questions in phonological theory. *Journal of Linguistics*, *1*(02), 97-138.
Corlett, P. R., Aitken, M. R., Dickinson, A., Shanks, D. R., Honey, G. D., Honey, R. A., ... & Fletcher, P. C. (2004). Prediction error during retrospective revaluation of causal associations in humans: fMRI evidence in favor of an associative model of learning. *Neuron*, *44*(5), 877-888.
Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in cognitive sciences*, *10*(7), 294-300.
Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human perception and performance*, *14*(1), 113-121.
Dąbrowska, E. (2008). The later development of an early-emerging system: The curious case of the Polish genitive. *Linguistics*, 46, 629-650.
Danks, D. (2003). Equilibria of the Rescorla–Wagner model. *Journal of Mathematical Psychology*, *47*(2), 109-121.
De Houwer, J., & Beckers, T. (2003). Secondary task difficulty modulates forward blocking in human contingency learning. *Quarterly Journal of Experimental Psychology Section B*, *56*(4), 345-357.
Dell, G. S., Burger, L. K., & Svec, W. R. (1997). Language production and serial order: A functional analysis and a model. *Psychological review*, *104*(1), 123-147.
den Ouden, H. E., Daunizeau, J., Roiser, J., Friston, K. J., & Stephan, K. E. (2010). Striatal prediction error modulates cortical coupling. *The Journal ofNneuroscience*, *30*(9), 3210-3219.
Dickinson, A. (2001). Causal learning: An associative analysis. *The Quarterly Journal of Experimental Psychology: Section B*, *54*(1), 3-25.
Do, Y. A. (2013). *Biased learning of phonological alternations*. (Doctoral Dissertation, MIT).
Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, *27*(2), 164-194.
Ellis, N. C., O'Donnell, M. B., & Römer, U. (2014). Second language verb-argument constructions are sensitive to form, function, frequency, contingency, and prototypicality. *Linguistic Approaches to Bilingualism*, *4*(4), 405-431.
Ellis, N. C., & Sagarra, N. (2011). Learned attention in adult language acquisition. *Studies in Second Language Acquisition*, *33*(04), 589-624.
Elsner, B., & Hommel, B. (2001). Effect anticipation and action control. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(1), 229-240.
Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review*, *57*(2), 94-107.
Finley, S. (2008). *Formal and cognitive restrictions on vowel harmony*. (Doctoral Dissertation, Johns Hopkins University).
French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: a recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological review*, *118*(4), 614-636.
Frigo, L., & McDonald, J. L. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language*, *39*(2), 218-245.
Gagné, C. L., & Spalding, T. L. (2015). Written production of English compounds: effects of morphology and semantic transparency. *Morphology*, OnlineFirst, 1-23.

Gambell, T., & Yang, C. D. (2003). Scope and limits of statistical learning in word segmentation. In *Proceedings of the 34th Northeastern Linguistic Society Meeting (NELS)* (pp. 29-30). Stony Brook University New York.

Ghirlanda, Stefano. (2005). Retrospective revaluation as simple associative learning. *Journal of Experimental Psychology: Animal Behavior Processes* 31(1): 107-111.

Giroux, I., & Rey, A. (2009). Lexical and sublexical units in speech perception. *Cognitive Science*, *33*(2), 260-272.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General*, *117*(3), 227-247.

Goldberg, A. E. (2003). Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, *7*(5), 219-224.

Goldberg, A. E. (2002). Surface generalizations: An alternative to alternations. *Cognitive Linguistics*, *13*(4), 327-356.

Goldberg, A. (1995). *Constructions, A Construction Grammar approach to argument structure*. Chicago University Press.

Goldstone, R. L. (2003). Learning to perceive while perceiving to learn. In R. Kimchi, M. Behrmann & C. R. Olson (Eds.), *Perceptual organization in vision: Behavioral and neural perspectives*, 233-278. Lawrence Erlbaum.

Goldstone, R. L. (2000). Unitization during category learning. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(1), 86-112.

Goldwater, S., & Johnson, M. (2003, April). Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm workshop on variation within Optimality Theory* (pp. 111-120).

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychological review*, *111*(1), 3-32.

Gouskova, M., & Becker, M. (2013). Nonce words show that Russian yer alternations are governed by the grammar. *Natural Language & Linguistic Theory*, *31*(3), 735-765.

Gregory, M. L., Raymond, W. D., Bell, A., Fosler-Lussier, E., & Jurafsky, D. (1999). The effects of collocational strength and contextual predictability in lexical production. In *Chicago Linguistic Society* (Vol. 35, pp. 151-166).

Gries, S. T. (2013). 50-something years of work on collocations: what is or should be next…. *International Journal of Corpus Linguistics*, *18*(1), 137-166.

Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science*, *32*(1), 68-107.

Guion, S. G. (1998). The role of perception in the sound change of velar palatalization. *Phonetica*, *55*(1-2), 18-52.

Hale, M., & Reiss, C. (1998). Formal and empirical arguments concerning phonological acquisition. *Linguistic inquiry*, *29*(4), 656-683.

Hall, G. (2009). Learning in simple systems. *Behavioral & Brain Sciences*, *32*, 210-211.

Harmon, Z., & Kapatsinski, V. (2015). Determinants of Lengths of Repetition Disfluencies: Probabilistic syntactic constituency in speech production. *Chicago Linguistic Society*.

Hawkins, R. D., & Kandel, E. R. (1984). Is there a cell-biological alphabet for simple forms of learning?. *Psychological Review*, *91*(3), 375-391.

Hayes, B. (2004). Phonological acquisition in Optimality Theory: The early stages. In R. Kager, J. Pater & P. Zonneveld (Eds.), *Constraints in phonological acquisition*, 158-203. CUP.

Hayes, B. & White, J. (2015). Saltation and the P-map. *Phonology*, 32(2), 1–36.

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

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.

Hewlett, D., & Cohen, P. (2011). Word segmentation as general chunking. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning* (pp. 39-47). Association for Computational Linguistics.

Holyoak, K. J. (1991). Symbolic connectionism: toward third-generation theories of expertise. In Anders, E. K., & Smith, J. (Eds). *Toward a general theory of expertise: Prospects and limits*, 301-335. Cambridge University Press.

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual review of psychology*, *62*, 135-163.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*(4), 548-567.

Johnson, E. K., & Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. *Developmental Science*, *13*(2), 339-345.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In Bybee, J. L., & Hopper, P. J. (Eds.). *Frequency and the emergence of linguistic structure* (pp.229-245). John Benjamins.

Kamin, L.J. (1969). Predictability, surprise, attention and conditioning. In B.A. Campbell & R.M. Church (eds.), *Punishment and aversive behavior*, 279–96, New York: Appleton-Century-Crofts

Kapatsinski, V. (2016). Emerging conspiracies of addition and subtraction. Linguistic Society of America Annual Meeting, Washington, DC, January 7-10.

Kapatsinski, V. (2013). Conspiring to mean: Experimental and computational evidence for a usage-based harmonic approach to morphophonology. *Language*, *89*(1), 110-148.

Kapatsinski, V. (2012). What statistics do learners track? Rules, constraints and schemas in (artificial) grammar learning. In S. Th Gries & D. Divjak (Eds.), *Frequency effects in language learning and processing*, 53-73. Berlin: Mouton de Gruyter.

Kapatsinski, V. (2011). Modularity in the channel: The link between separability of features and learnability of dependencies between them. In *Proceedings of the XVIIth International Congress of Phonetic Sciences* (pp. 1022-1025).

Kapatsinski, V. (2010). Velar palatalization in Russian and artificial grammar: Constraints on models of morphophonology. *Laboratory Phonology*, *1*(2), 361-393.

Kapatsinski, V. (2009). *The architecture of grammar in artificial grammar learning: Formal biases in the acquisition of morphophonology and the nature of the learning task* (Doctoral dissertation, Indiana University).

Kapatsinski, V. (2009b). Testing theories of linguistic constituency with configural learning: The case of the English syllable. *Language*, *85*(2), 248-277.

Kapatsinski, V. (2007). Frequency, neighborhood density, age-of-acquisition, lexicon size, neighborhood density and speed of processing: Towards a domain-general, single-mechanism account. In S. Buescher, K. Holley, E. Ashworth, C. Beckner, B. Jones, and C. Shank (Eds.). *Proceedings of the 6th Annual High Desert Linguistics Society Conference*, 121-40. Albuquerque, NM: High Desert Linguistics Society.

Kapatsinski, V. (2007b). Implementing and testing theories of linguistic constituency I: English syllable structure. *Research on Spoken Language Processing Progress Report No*, *28*, 241-76.

Kapatsinski, V. (2005). Constituents can exhibit partial overlap: Experimental evidence for an exemplar approach to the mental lexicon. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society* (Vol. 41, No. 2, pp. 227-242). Chicago Linguistic Society.

Kapatsinski, V. (2005b). Measuring the relationship of structure to use: Determinants of the extent of recycle in repetition repair. In *Annual Meeting of the Berkeley Linguistics Society* (Vol. 30, No. 1).

Kenstowicz, M. (1996). Base identity and uniform exponence: Alternatives to cyclicity. In J. Durand & B. Laks (Eds.) *Current trends in phonology: Models and methods*, vol. 1 (pp. 363-393). University of Salford.

Kerkhoff, A. O. (2007). *Acquisition of morpho-phonology: The Dutch voicing alternation* (Doctoral dissertation, University of Nijmegen).

Kisseberth, C. W. (1970). On the functional unity of phonological rules. *Linguistic inquiry*, *1*(3), 291-306.
Kochetov, A. (2011). Palatalisation. In C. Ewen, B. Hume, M. van Oostendorp, & K. Rice (Eds.) *Blackwell Companion to Phonology*, 1666-1690. Wiley-Blackwell.
Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior*, *36*(3), 210-226.
Kruschke, J. K. (2003). Attention in learning. *Current Directions in Psychological Science*, *12*(5), 171-175.
Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*, *45*(6), 812-863.
Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological review*, *99*(1), 22-44.
Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review*, *7*(4), 636-645.
Kuperman, V., Schreuder, R., Bertram, R., & Baayen, R.H. (2009).Reading polymorphemic Dutch compounds: Toward a multiple route model of lexical processing. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 876-895.
Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (2013). Human learning of elemental category structures: Revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 552-572.
Lachter, J., & Bever, T. G. (1988). The relation between linguistic structure and associative theories of language learning—A constructive critique of some connectionist learning models. *Cognition*, *28*(1), 195-247.
Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites* (Vol. 1). Stanford University Press.
Legendre, G., Miyata, Y., & Smolensky, P. (1990). *Can Connectionism Contribute to Syntax?: Harmonic Grammar, with an Application*. University of Colorado, Boulder, Department of Computer Science.
Lobben, M. (1991). *Pluralization of Hausa nouns, viewed from psycholinguistic experiments and child language data* (MA Thesis, University of Oslo).
Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*(4), 955-984.
Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology*, *15*(3), 215-233.
Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*(4), 276-298.
Mackintosh, N. J., & Turner, C. (1971). Blocking as a function of novelty of CS and predictability of UCS. *The Quarterly Journal of Experimental Psychology*, *23*(4), 359-366.
Martin, A. T. (2007). *The evolving lexicon*. (Doctoral dissertation, UCLA).
Matute, H., Arcediano, F., & Miller, R. R. (1996). Test question modulates cue competition between causes and between effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 182-196.
Matzel, L. D., Schachtman, T. R., & Miller, R. R. (1985). Recovery of an overshadowed association achieved by extinction of the overshadowing stimulus. *Learning and Motivation*, *16*(4), 398-412.
McCarthy, J. J. (2008). *Doing Optimality Theory: Applying theory to data*. Wiley.
McClelland, J. L. (2006). How far can you go with Hebbian learning, and when does it lead you astray. *Processes of change in brain and cognitive development: Attention and performance XXI*, *21*, 33-69.
McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419-457.
Miller, R. R., & Matute, H. (1998). Competition between outcomes. *Psychological Science*, *9*(2), 146-149.

Miller, R. R., & Matzel, L. D. (1988). The comparator hypothesis: A response rule for the expression of associations. *The Psychology of Learning and Motivation*, *22*, 51-92.

Misyak, J. B., Christiansen, M. H., & B. Tomblin, J. (2010). Sequential expectations: The role of prediction-based learning in language. *Topics in Cognitive Science*, *2*(1), 138-153.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, *32*(2), 183-198.

Moreton, E. (2012). Inter-and intra-dimensional dependencies in implicit phonotactic learning. *Journal of Memory and Language*, *67*(1), 165-183.

Moreton, E. (2008a). Analytic bias and phonological typology. *Phonology*, *25*(1), 83-127.

Moreton, E. (2008b). Modelling modularity bias in phonological pattern learning. In *Proceedings of WCCFL* (Vol. 27, pp. 1-16).

Moreton, E., Pater, J., & Pertsova, K. (2015). Phonological concept learning. Ms. Chapel Hill: University of North Carolina, and Amherst: University of Massachusetts Amherst.

Morsanyi, K., Handley, S. J., & Evans, J. S. (2009). Heuristics and biases in autism: Less biased but not more logical. *CogSci Proceedings*, 75-80.

Nesset, T. (2008). *Abstract phonology in a concrete model: Cognitive linguistics and the morphology-phonology interface*. Berlin: Mouton de Gruyter.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*(2), 127-162.

Oppenheim, G. M., Dell, G. S., & Schwartz, M. F. (2010). The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition*, *114*(2), 227-252.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, *113*(2), 244-247.

Peperkamp, S., Skoruppa, K., & Dupoux, E. (2006). The role of phonetic naturalness in phonological rule acquisition. In *Proceedings of the 30th annual Boston University Conference on Language Development*. Cascadilla Press.

Perkell, J. S. (2012). Movement goals and feedback and feedforward control mechanisms in speech production. *Journal of Neurolinguistics*, *25*(5), 382-407.

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation?. *Memory & Cognition*, *36*(7), 1299-1305.

Pierrehumbert, J. B. (2006). The statistical basis of an unnatural alternation. In L. Goldstein, D. H. Whalen & C. Best (Eds.), *Laboratory phonology 8* (pp.81-107). Mouton de Gruyter.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1), 73-193.

Plaisted, K., Saksida, L., Alcántara, J., & Weisblatt, E. (2003). Towards an understanding of the mechanisms of weak central coherence effects: Experiments in visual configural learning and auditory perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *358*(1430), 375-386.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review*, *103*(1), 56-115.

Prados, J., Alvarez, B., Acebes, F., Loy, I., Sansa, J., & Moreno-Fernández, M. M. (2013). Blocking in rats, humans and snails using a within-subjects design. *Behavioural Processes*, *100*, 23-31.

Prince, A., & Smolensky, P. (2004). *Optimality Theory: Constraint interaction in generative grammar*. Wiley.

Purcell, D. W., & Munhall, K. G. (2006). Compensation following real-time manipulation of formants in isolated vowels. *The Journal of the Acoustical Society of America*, *119*(4), 2288-2297.

Pycha, A., Nowak, P., Shin, E., & Shosted, R. (2003). Phonological rule-learning and its implications for a theory of vowel harmony. In *Proceedings of the 22nd west coast conference on formal linguistics* (Vol. 22, pp. 101-114). Somerville, MA: Cascadilla Press.

Ramscar, M., Dye, M., & Klein, J. (2013). Children value informativity over logic in word learning. *Psychological Science*, *24*(6), 1017-1023.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science*, *6*(1), 5-42.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The feature-label-order effect in symbolic learning. *Cognitive Science*, *34*(7), 5-42.

Redford, M. A. (2015). Unifying speech and language in a developmentally sensitive model of production. *Journal of Phonetics*, *53*, 141-152.

Rehfeldt, R. A., & Chambers, M. R. (2003). Functional analysis and treatment of verbal perseverations displayed by an adult with autism. *Journal of Applied Behavior Analysis*, *36*(2), 259-261.

Rescorla, R. A. (1986). Two perceptual variables in within-event learning. *Animal Learning & Behavior*, *14*(4), 387-392.

Rescorla, R. A. (1973). Evidence for" unique stimulus" account of configural conditioning. *Journal of Comparative and Physiological Psychology*, *85*(2), 331-338.

Rescorla, R. A., & Furrow, D. R. (1977). Stimulus similarity as a determinant of Pavlovian conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, *3*(3), 203-215.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H., & Prokasy, W. F. (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Rodríguez, G., & Alonso, G. (2004). Perceptual learning in flavor-aversion learning: Alternating and blocked exposure to a compound of flavors and to an element of that compound. *Learning and Motivation*, *35*(3), 208-220.

Rumelhart, D. E. (1990). Brain style computation: Learning and generalization. In Zornetzer, S. F., Davis, J. L., & Lau, C. (Eds). *An introduction to neural and electronic networks* (pp.405-420). San Diego, CA: Academic Press.

Rumelhart, D. E., & McClelland, J. L. (1986). *On learning the past tenses of English verbs*. In: Rumelhart, D. E., McClelland, J. L., & The PDP Research Group. *Parallel distributed processing, Vol. 2*. Cambridge, MA: The MIT Press.

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

Sahel, S., Nottbusch, G., Grimm, A., & Weingarten, R. (2008). Written production of German compounds: Effects of lexical frequency and semantic transparency. *Written Language & Literacy*, *11*(2), 211-227.

Schwartz, R. G., & Leonard, L. B. (1982). Do children pick and choose? An examination of phonological selection and avoidance in early lexical acquisition. *Journal of child language*, *9*(02), 319-336.

Shanks, D. R. (2006). Bayesian associative learning. *Trends in cognitive sciences*, *10*(11), 477-478.

Shanks, D. R. (1995). Is human learning rational?. *The Quarterly Journal of Experimental Psychology*, *48*(2), 257-279.

Shattuck-Hufnagel, S. (2015). Prosodic frames in speech production. In M. A. Redford (Ed.), *The handbook of speech production* (pp.419-444). Wiley.

Shin, Y. K., Proctor, R. W., & Capaldi, E. J. (2010). A review of contemporary ideomotor theory. *Psychological Bulletin*, *136*(6), 943-974.

Shukla, M., Nespor, M., & Mehler, J. (2007). An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology*, *54*(1), 1-32.

Skinner, B. F. (1957). *Verbal behavior*. Prentice-Hall.

Slone, L. K., & Johnson, S. P. (2015). Statistical and chunking processes in adults' visual sequence learning. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society. Pasadena, CA: Cognitive Science Society*.

Smith, L. B., Thelen, E., Titzer, R., & McLin, D. (1999). Knowing in the context of acting: the task dynamics of the A-not-B error. *Psychological Review*, 106(2), 235-260.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and brain sciences*, *11*(01), 1-23.

Smolensky, P., & Legendre, G. (2006). *The harmonic mind: From neural computation to optimality-theoretic grammar*. MIT Press.

Sternberg, D., & McClelland, J. L. (2009). When should we expect indirect effects in human contingency learning. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 206-211).

Taft, M., & Meunier, F. (1998). Lexical representation of gender: A quasiregular domain. *Journal of Psycholinguistic research*, *27*(1), 23-45.

Tesar, B., & Smolensky, P. (1998a). Learnability in optimality theory. *Linguistic Inquiry*, *29*(2), 229-268.

Tesar, B. B., & Smolensky, P. (1998b). Learning optimality-theoretic grammars. *Lingua*, *106*(1), 161-196.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review*, *108*(3), 550-592.

Vadillo, M. A., & Matute, H. (2007). Predictions and causal estimations are not supported by the same associative structure. *The Quarterly Journal of Experimental Psychology*, *60*(3), 433-447.

van Noord, R., & Spenader, J. K. (2015). Modeling the learning of the English past tense with memory-based learning. *Computational Linguistics in the Netherlands (CLIN), Antwerp*, *6*.

Van Osselaer, S. M., Janiszewski, C., & Cunha Jr, M. (2004). Stimulus generalization in two associative learning processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(3), 626-638.

Vihman, M., & Croft, W. (2007). Phonological development: Toward a "radical" templatic phonology. *Linguistics*, *45*(4), 683-725.

Villacorta, V. M., Perkell, J. S., & Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *The Journal of the Acoustical Society of America*, *122*(4), 2306-2319.

Wahl, A. (2015). Intonation unit boundaries and the storage of bigrams: Evidence from bidirectional and directional association measures. *Review of Cognitive Linguistics*, *13*(1), 191-219.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 53-76.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, *121*(2), 222-236.

Warker, J. A., & Dell, G. S. (2006). Speech errors reflect newly learned phonotactic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(2), 387-296.

Wasserman, E. A. (1990). Detecting response-outcome relations: Toward an understanding of the causal texture of the environment. *The Psychology of Learning and Motivation*, *26*, 27-82.

Westermann, G., & Ruh, N. (2012). A neuroconstructivist model of past tense development and processing. *Psychological Review*, *119*(3), 649-667.

White, J. (2014). Evidence for a learning bias against saltatory phonological alternations. *Cognition*, *130*(1), 96-115.

White, J. (2013). *Bias in phonological learning: Evidence from saltation*. (Doctoral Dissertation, UCLA).

White, J., & Sundara, M. (2014). Biased generalization of newly learned phonological alternations by 12-month-old infants. *Cognition*, *133*(1), 85-90.

Williams, D. A., Sagness, K. E., & McPhee, J. E. (1994). Configural and elemental strategies in predictive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(3), 694-709.

Yang, C. D. (2004). Universal Grammar, statistics or both? *Trends in Cognitive Sciences*, *8*(10), 451-456.

Yin, H., & White. J. (2016). Neutralization avoidance and naturalness in artificial language learning. Linguistic Society of America Annual Meeting, Washington, DC, January 7-10.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition, 125*(2), 244-262.

Yun, G. H. (2006). *The interaction between palatalization and coarticulation in Korean and English*. (Doctoral dissertation, University of Arizona).

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Oxford: Addison-Wesley.

Zuraw, K. (2007). The role of phonetic knowledge in phonological patterning: corpus and survey evidence from Tagalog infixation. *Language, 83*(2), 277-316.

Zuraw, K. (2000). *Patterned exceptions in phonology* (Doctoral Dissertation, UCLA).