Why not to change (the stem): A production-internal account of paradigm uniformity
Amy Smolek & Vsevolod Kapatsinski
University of Oregon

Abstract
Languages tend to avoid stem changes, and especially large stem changes. This bias is known as paradigm uniformity. This paper proposes that the bias arises within the production system. First, when generating a new form of a known word (or having difficulty retrieving the form from memory), the speaker activates known forms of the same word. Parts of those words are then available for being perseverated on. Because the target form is unknown or inaccessible, perseveration in this situation is in fact functional, reducing error, but has the side effect of leveling stem changes. The bias against large changes is argued to arise from the fact that dissimilar representations are hard to associate. We argue that learning to produce a stem change involves learning an association from the input to the change to its output. When the input and the output are very difficult, acquisition of the association is difficult. This likely has a neural basis: The more distant the to-be-associated representations, the more synaptic change needs to be effected to learn the association. We present experimental support for the proposed theory from a miniature artificial language learning experiment, contrasting predictions of the theory with existing proposed motivations for paradigm uniformity, including concatenative nature of morphology, avoidance of perceptible changes, and the preference for unidimensional classical categories.

Keywords: phonology, substantive bias, alternations, perseveration, similarity, associative learning

Introduction
Paradigm uniformity refers to the force militating against stem changes, or, in other words, against having multiple forms of the same stem. Especially abhorrent to paradigm uniformity are cases in which the allomorphs of the stem are phonologically dissimilar. In many languages, we can observe paradigm uniformity in action when productive phonological processes seem to fail to apply whenever their application would violate paradigm uniformity or seem to overapply in order to increase paradigm uniformity. For example, in Canadian Raising, exemplified in (1), /aɪ/ raises to [ʌɪ] before voiceless consonants, hence [bʌɪt] but [bɑɪd]. However, Canadian Raising overapplies before voiced flaps that correspond to voiceless stops in other forms of the same stem (Joos 1942). As a result of this overapplication, the paradigm is more uniform: all forms of the same stem share vowel height.

(1) bʌɪt ‘bite’
    bɑɪd ‘bide’
    ɓʌɪɾɪŋ ‘biting’
    ɓɑɪɾɪŋ ‘biding’

In order to account for patterns like this, work within Optimality Theory (Benua 1997, Burzio 1996, Flemming 1995, Kenstowicz 1996, McCarthy 1998) has proposed base identity constraints militating in favor of preserving particular features of the derivational or inflectional base. In the case of Canadian
Raising, the feature [low] in biting would be preserved by a high ranking of an Ident\(\alpha\)-[low] constraint, which would force biting to retain the raised [\(\alpha\)] of its base bite.\(^1\) Base identity constraints are thought to be universal and to sit at the top of the constraint hierarchy unless demoted by learning (Hayes 2004, McCarthy 1998). This initial high ranking of the PU constraints is consistent with observations that children often level stem changes that adults in the same community reliably produce. For example, Kerkhoff (2007) documents that Dutch voicing alternation (final devoicing and undoing thereof) is less productive for Dutch children than for Dutch adults, despite a high degree of regularity and phonetic naturalness. Do (2013) documents that Korean children avoid producing stem changes using a variety of repair strategies. However, explaining these findings by a high ranking of PU constraints begs the question: Why are there PU constraints, and why should they be highly ranked in children?

The present paper argues that PU constraints are grounded in paradigmatic motor perseveration. As children are observed to show more motor perseveration across domains (Dell et al. 1997, Smith et al. 1999), this proposal explains why children are more likely to preserve the stem compared to adults.\(^2\) By grounding paradigm uniformity in production, we also explain why speakers may prefer stem changes in judgment tasks while dispreferring them in production (Kapatsinski 2012, Zuraw 2000). As we will see shortly, these predictions differ from those of alternative groundings of paradigm uniformity in perception (Hayes & White 2015; Kenstowicz 1996; Steriade 2000, 2009), categorization (Moreton & Pater 2012a), or storage economy (Kenstowicz 1998). Before addressing alternatives, however, we will provide the motivation for the present proposal.

Paradigmatic perseveration

Imagine a participant in an elicited production wug test (Berko 1958). The participant is given a form of a word and asked to produce another form, using the given form as a base. It would indeed be puzzling if the participant did not perseverate on aspects of the form given to them, copying them into the output she constructs. Perseveration on aspects of recently produced forms not sharing the stem with the target has been independently noted in wug tests by several researchers (Bickel et al. 2007, Caballero 2010, Lobben 1991). In particular, Lobben (1991) found that Hausa plurals produced in seeming violation of the rules of the language in a wug test could be explained if one examined the plural forms produced immediately by the same participant immediately before the anomalous form, as these rule-obeying forms were usually formally similar to the anomalous form produced immediately afterwards. Caballero (2010) noted that certain morphemes were produced in non-scopal orders if they had just been produced in the same order in an environment where that order was scopal. Importantly, while we notice perseveration when it produces errors, perseveration on aspects of the base is usually beneficial, as the to-be-produced form always retains at least most aspects of the base (Pinker & Prince 1988): full suppletion is relatively rare. Indeed, morphology would not exist if it were otherwise.

\(^1\) In other approaches (e.g. McCarthy 2008), paradigm uniformity is enforced by constraints evaluating entire paradigms of forms. We avoid postulating paradigm evaluation here, as the aim of the present paper is to examine whether we can derive paradigm uniformity from processes that operate during everyday language processing (see also Boersma 1998). We expect that effects attributed to evaluation of paradigms can emerge from the fact that more than a single base form is usually simultaneously activated when a word is being produced.

\(^2\) In addition, children face lexical retrieval problems more often. This may cause them to activate other forms of the same word before the target form is accessed or constructed, leading to perseveration on various parts of those activated words.
Note that we do not believe the proposed motor perseveration tendency to be a methodological artifact of the elicited production procedure. While elicited production tests encourage motor perseveration by providing the speaker with a base form, we consider the same tendency to play a role at any time that the speaker produces a novel form (or a novel derivative) of a known word. In any such situation, known forms of the word are activated, allowing the speaker to perseverate on the gestures comprising them. The main difference from elicited production in the lab is that multiple forms of a word, rather than a single base, are activated during normal production (pace Albright 2008).³

Kapatsinski (2010) presents empirical evidence for the notion that co-activation of morphological relatives can generate interference at the level of surface form, resulting in errors where parts of co-activated morphological relatives erroneously surface in the produced target (see also Spalek & Schriefers 2005). These paradigmatic perseveration errors, akin to the syntagmatic perseveration error in (2) where [u] is erroneously replaced by [ou] and the errors induced by interference from unrelated but temporally adjacent forms in wug tests, substantiate the proposed mechanism of paradigmatic perseveration.

(2) Was there, like, an expl[ou]sion of p[ou]p?

As illustrated in (3), Russian bez-initial adjectives always have corresponding prepositional phrases but frequently lack corresponding bez-less adjectives. This makes prepositional phrases likely derivational bases for the bez-initial adjectives: the forms of bez-initial adjectives are predictable under the assumption they are – or have been – constructed from prepositional phrases via an adjectivizing schema like [[[N]nyj]ADJ.MASC.SG.NOM but not under the assumption that they are constructed from bez-less adjectives by attachment of bez.

(3) bezkr'lyj   bez kryl'jev   *kr'lyj
‘wingless’ ‘without wings’ ‘winged’
bezmyslennyj   beg smysla   *myslennyj
‘pointless’ ‘without a point’ ‘having a point’
/z/ [s] <s>   /z/ [s] <z>

Both the prefix bez- and the preposition bez underlingly end in /z/, which is pronounced as such before vowels and voiced consonants but devoices to [s] before voiceless consonants. The preposition has a constant spelling across contexts, corresponding to the underlying form, with the Cyrillic equivalent of <z>. However, the prefix, like all /z/-final prefixes in Russian, is supposed to be spelled the way it is pronounced, with <s> before voiceless consonants and <z> elsewhere.⁴

³ Rich-get-richer dynamics favor re-use of previously used forms and paradigmatic mappings between forms (e.g. Martin 2007, Zipf 1949). As a consequence, a single wordform may often have a dominant influence on the constructed output: it is more frequent than competitors and/or is most predictive of characteristics of the output (Albright 2008). However, this should not obscure the fact that activation spreads to multiple words in parallel during lexical retrieval (Dell 1986, Roelofs 1992), giving multiple words the opportunity to influence the output.

⁴ In Cyrillic, where <z> is <з> and <s> is <с>.
Kapatsinski (2010) shows that low-frequency Russian adjectives with the prefix bez- (e.g. bezkreditnyj ‘creditless’ ← bez kredita ‘without credit’), are often erroneously spelled like the corresponding prepositional phrases (with <z> instead of <s> before voiceless consonants, where the prefix is pronounced [bes]). The errors are not observed with other /z/-final prefixes, which are supposed to obey the same spelling rule as bez- but do not have corresponding phrases with a differently-spelled preposition, i.e. iz-, roz-, and voz-. The verbal prefix iz- ‘out’ is especially instructive here. Like bez-, it has a synonymous preposition, iz ‘out of’, yet the constant spelling of the preposition does not appear to interfere with the spelling of the prefix. Spelling errors for iz- are an order of magnitude lower than for bez-. The explanation can be found in the fact that iz-initial verbs do not have corresponding prepositional phrases. It is not the preposition bez that interferes with the prefix bez-. It is the prepositional phrase that interferes with the adjective. The errors in bez- seem to come from basing production on an activated orthographic form of the prepositional phrase constituting the base of a bez-initial adjective when the bez-initial adjective under construction is rare / hard to retrieve from memory. The iz-initial verbs, lacking corresponding prepositional phrases, do not suffer from interference.

When perseveration can drive production

We suggest that these kinds of errors illustrate the mechanism behind paradigm uniformity: a morphologically-related form (here, a prepositional phrase with bez) is activated more or earlier than the target (here, a rare bez-initial adjective), resulting in aspects of its form being perseverated on in producing the target. The proposed mechanism is therefore not restricted to operate only during novel word production: perseveration on aspects of morphologically related forms is always there in the background, racing against lexical retrieval of the target form (cf. Baayen et al. 1997). However, for known words, it is usually outpaced by lexical retrieval of the target form, making its influence most useful (and most easily observable) when a novel form of a known word is to be produced.

In children, retrieval or planning of the target is slower. As a result, children would often partially activate a morphological relative long before fully accessing the target, allowing aspects of the relative’s form to interfere with production of the target, thus leveling stem changes (Do 2013, Kerkhoff 2007). Children also show more perseveration than adults (Dell et al. 1997, Smith et al. 1999). As a result, they should be especially unlikely to change the stem when the stem is presented to them in an elicited production test.

Perseveration (i.e. allowing a recently activated form to drive production) may help children learn language by encouraging imitation. Children have been observed to often repeat words recently produced by an interlocutor, which sometimes manifests in pronoun confusion errors e.g. using you instead of I in response to Do you wanna eat? (Clark 1974, 1977, Rubino & Pine 1998). However, aside from the relatively rare pronoun confusions and leveling stem changes in wug tests, such imitative repetition is functional in allowing the child to re-produce correct structures recently produced by the interlocutor that would be too complex for the child to produce compositionally on their own (Ambridge & Lieven 2011, pp.165-166, Clark 1977, Farrar 1992, Rubino & Pine 1998). For example, Rubino & Pine (1998) observed that children acquiring Brazilian Portuguese correctly produced verbs with the right inflectional suffixes only immediately after hearing them from their caregiver. Like perseverating on aspects of the given form or a recently produced form in a wug test (e.g. Bickel et al. 2007, Caballero
2010, Lobben 1991), imitation of the interlocutor involves repeating aspects of a given form when the target is difficult to plan or retrieve. We thus suggest that the same perseveratory mechanism is involved in this kind of immediate repetition.

**Elicited Production vs. Judgment**

The Perseveration Hypothesis attributes paradigm uniformity to paradigmatic perseveration in production. As a result, it predicts that paradigm uniformity should have an especially strong influence on behavior in the elicited production task. Conversely, its influence should be attenuated when knowledge of phonology is tested using judgments.

As noted above, in the production task, the participant is provided with the input / base form and asked to generate another form of the same word. Thus, the form on which the participant may potentially perseverate is the only form provided. The surface form of the to-be-changed stem is effectively primed while the changed surface form is not. In a judgment task, the (possible) output of production is provided along with the input.

According to the Perseveration Hypothesis, changing the stem is hard, so an output’s production may be prevented by paradigm uniformity because it involves a (major) stem change. Nonetheless, the output may be judged as acceptable because it contains all the right cues to the meaning being expressed. Thus any production-internal paradigm uniformity biases may be attenuated in the judgment task, to the extent that the judge fails to base their judgment on an accurate simulation of the production process and the difficulties of changing the stem.

The pattern of results predicted by the Perseveration Hypothesis has been previously observed in a small-scale study comparing elicited production and judgment in natural language. Zuraw (2000) examined the productivity of Tagalog nasal substitution in (4) vs. nasal assimilation in (5) with elicited production and judgment tests. The participants seldom produced nasal substitution, which requires a stem change, but judged it more acceptable than nasal assimilation, which does not require a stem change. It is this pattern of results that we take to constitute evidence for a production locus of a bias against stem changes.

(4) maN + bigaj → mamigaj
   ‘give’ → ‘distribute’
(5) paN + kaliskis → paŋkaliskis
   ‘scales’ → ‘tool for removing scales’

**Phonologization of perseveration**

To summarize the argument so far, we argue that the universal pressure towards paradigm uniformity is the extra-grammatical tendency of the production system to perseverate on activated units. This extragrammatical perseveratory pressure is responsible for 1) avoidance of stem changes in children (Kerkhoff 2007, Do 2013), and 2) production/judgment dissociations, where a stem change may be judged as better than no change but rarely produced (Zuraw 2000).
Like other processing pressures, perseveration can become conventionalized (i.e. phonologized), coming to be part of the grammar of a language. We call conventionalized perseveration *copying*, in that a certain part of the input is conventionally copied into the output being constructed. Following the basic insight behind faithfulness constraints (Prince & Smolensky 2004), we propose that the pressure to copy can vary in strength across submorphological structures, with these strengths being acquired from language experience. However, for us, the targeted structures are positions within a prosodic template rather than substantively defined units like gestures or features. Thus, for example, CopyFin cares about copying the final element of the input. This proposal is motivated by 1) the fact that changes do not target all instances of a unit within the base, but rather only instances in certain positions, and 2) the fact that learners need to be able to generalize from experienced elements in a certain position to novel elements in the same position. For example, a learner that has encountered a number of examples in which the word-initial onset is preserved in the output should conclude not only that the experienced onsets are copied but also that novel onsets would be copied as well.

While copying is position-specific and not unit-specific, it is conditioned by the meaning that the speaker wants to express (e.g. PLURAL), and phonological characteristics of the input (e.g. a final [p]). Thus, in one language CopyFin might be activated when the speaker wants to express a plural meaning and the input ends in [p], while in another language it might be inhibited, allowing the speaker to replace the [p] with something else. In particular, speakers of Southern Bantu languages, where labial palatalization (p→tʃw) is productive (Braver & Bennett 2015, Ohala 1978) must have learned an inhibitory association from [p] (in final position) to CopyFin, presumably on the basis of observing many cases in which a [p] is not preserved when a form with a certain meaning is produced from a morphologically related form.

A simple language with labial palatalization is illustrated in Figure 1. The figure shows only input nodes representing the word [bлаp] and output nodes that are either inhibited or activated by these input nodes when the network is asked to express the PLURAL meaning. However, the full network includes nodes for all phonological units of the language and all meanings corresponding to cells in a morphological paradigm. The full network also tries to learn mappings between every pair of forms in a paradigm, e.g. both from singular to plural and from plural to singular. Kapatsinski (Submitted) shows that the state of knowledge depicted in Figure 1 can be achieved by the network on the basis of experiencing singular and plural forms, and – on a minority of occasions – recalling the singular form when the plural form is experienced, trying to predict the experienced plural form from the recalled singular and punishing connections that lead to incorrect predictions.

In the limit, exposure to unfaithful mappings between two cells in a morphological paradigm can cause the learner to come to prefer not to retain the units that need to be frequently replaced or deleted during production of a novel wordform from a known one. The ability to learn a preference for non-retention (anti-faithfulness) is useful for learning stem changes, subtractive patterns (including run-of-the-mill affix stripping), and ‘morphological toggles’ (Alderete 2001, Horwood 2001, Kurisu 2001). On
this account, part of learning a language is learning when and what to copy. This learning is responsible for language-specific preferences in favor of or against preserving specific production units in specific morphophonological environments.

**Figure 1.** A language with phonologized labial palatalization. Width of the line shows connection strength. The dashed line shows an inhibitory connection. Word-initial onsets are always copied into the output, as are the vowel nuclei that follow them. However, a final [p] is not copied into the plural, being replaced with [tʃi]. Every feature of the input excites CopyInit and CopyN1 because initial onsets and following vowels are always copied between paradigm cells in this language. A final [p] strongly activates output [tʃi] and inhibits copying of the consonant in the final position (itself). The plural meaning is strongly associated with the plural suffix –i (present whenever the plural meaning is present and not otherwise) and is more weakly associated with a preceding [tʃ] (which is overrepresented in plurals). Copying of onsets and first-syllable nuclei is associated with plurality as it is with other input features, while copying of the final consonant is associated more weakly since it does not always happen in the plural.

![Diagram](image)

*Alternative explanations of Paradigm Uniformity*

In order to further establish that a perseveratory basis of paradigm uniformity might be worth pursuing, we will now compare the present proposal to previously proposed motivations for paradigm uniformity. The aim is not to deny that other motivations for paradigm uniformity may play a role in its emergence but to argue that the perseveratory motivation is at least competitively plausible.

In their influential critique of connectionist modeling of language, Pinker & Prince (1988) criticized Rumelhart & McClelland (1986), a model much like the one outlined in Figure 1 sans copy outputs, for predicting *mumbled* as the past tense of *mail*. As they write, “because the relation between stem and past tense is portrayed as a transduction from one low-level featural representation to another, literally replacing every feature in the input, it becomes an inexplicable accident that the regular formation rule preserves the stem unaltered.” The tendency to preserve too little of the given form has remained a persistent difficulty for connectionist models of behavior in the wug test, including the model in Figure 1 if Copy outcomes are omitted.

Pinker & Prince (1988) argue that stem preservation is an argument for an item-and-arrangement model of morphology featuring default concatenative morphological rules (“a tendency toward preservation of

---

6 This templatic coding scheme is sufficient for the languages presented to learners in the present paper, where stems are monosyllabic but would have to be extended for real languages.

7 For example, if the model has no CopyInit, it predicts that a novel onset will always be replaced by some previously encountered onset. Which one it is depends on which onsets have co-occurred with the rest of the word and the to-be-expressed meaning in the model’s experience. Kapatsinski (Submitted) shows that Copy outcomes solve this problem.
stem identity, a typical linguistic phenomenon, is an immediate consequence of the existence of morphology as a level of description: if the rule is Word = Stem + Affix, then ceteris paribus the stem comes through\(^\)\(^\text{a}\).

However, morphology as a level of description does not require novel forms of known words to be generated by combining stems and affixes and could instead operate on the basis of whole-word representations organized into networks of paradigmatic relations (e.g. Booij 2010, Bybee 1985, Hockett 1954, Matthews 1965, 1972, Robins 1959; see J. P. Blevins 2013 for a review). The ubiquity (and even the very existence) of morphological rules of the type envisioned by Pinker & Prince (1988) has been questioned in the morphological literature based on several sources of evidence, two of which are particularly relevant for the present purposes. First, there are complex morphological systems in which derivation of forms from stems and affixes is impossible because one needs to know more than one complete form of a word to derive other forms, i.e. the paradigm has multiple principal parts (see Ackerman et al. 2009, Ackerman & Malouf 2013, J. P. Blevins 2006, Stump 2001). Based on this kind of evidence, word-and-paradigm and constructionist approaches to morphology argue that morphological structure is better described as a network of word-sized units.\(^\text{b}\) Second, parts of the stem can fuse with affixes they co-occur with and the meaning in whose context they co-occur, forming a construction or schema whose boundaries do not align with erstwhile morpheme boundaries. For example, \([\_\_\_\_\_\_\_\text{holic}]\) was extracted from alcoholic and generalized to mean ‘addicted to’, as in workaholic, which is difficult to explain if alcoholic is regularly derived from alcohol+ic (Bybee 1985). The formation of cross-boundary units can happen even in the earliest stages of acquiring a miniature morphological system. For example, Kapatsinski (2012, 2013) shows that encountering examples like the singular form [blutʃ] mapping onto plural [blutʃi] leads learners to think that [tʃ]-final singulars should map onto [tʃi]-final plurals as well (slitʃ~slitʃi) rather than mapping onto [ti]-final plurals (slit~sliti). Kapatsinski (2012, 2013) argues that learners take examples of tʃ~tʃi as supporting the generalization that [tʃi] means ‘plural’, favoring non-compositional [slit]-[slitʃi], over the generalization that the plural is Stem+i, which would favor [slit]-[sliti]. As shown in Kapatsinski (Submitted), this finding is captured by our model but is problematic for models that assume that morphologically related forms are separated into a change and a context/stem (Albright & Hayes 2003, Gouskova et al. 2015). In general, there is little evidence that language learners are particularly respectful of the stem-affix boundary. This makes it likely that the apparent tendency to preserve the stem is not due to the stem being a basic unit from which words are derived by morpheme concatenation.\(^\text{c}\)

Another problem for explaining paradigm uniformity as a consequence of concatenative morphology is that the bias for stem preservation appears to be much stronger than the bias to maintain a constant shape for an affix (Beckman 1998, Benua 1997, McCarthy & Prince 1995). For example, vowel harmony tends to spread features of the stem vowels onto affix vowels rather than affix vowels onto stem

---

\(^{8}\) In construction morphology (Booij 2010), some of these word-sized units are partially underspecified, e.g. [chair [N]i] but they are still word-sized. So, for example, that the chair in chairperson is not the word chair, as in ‘I am sitting on a chair’ but rather an instance of the [chair [N]i] construction as evidenced by the difference in meaning between the two. For example, the chaircat in ‘I was approached by the chaircat of the feline food evaluation committee’ is not a cat that likes chairs.

\(^{9}\) The cross-linguistic ubiquity of concatenative morphology is, on this account, a consequence of the fact that affixes are former words that have grammaticalized in situ (Bybee 1985: 41, Lehmann 1992).
vowels, a typological generalization that is accounted for by a tendency to preserve the stem in Bakovic (2003). Finley (2015) even documents a learning bias in favor of vowel harmony (and therefore against PU) in affixes. This stem/affix asymmetry is unexpected if stem preservation and affix preservation both come from a word=stem+affix rule. In contrast, as in Correspondence Theory (McCarthy & Prince 1995), the affix is usually absent from the input in our model.\(^\text{10}\) The affix form is activated from the meaning to be expressed and arbitrary paradigmatic associations. Not being part of the input (surface form of an activated morphological relative), it cannot be perseverated on.

Most fundamentally, the morphological explanation for paradigm uniformity is undermined by the fact that some aspects of the stem are more likely to be preserved than others. For example, in the Canadian Raising example (1), the height of the vowel is preserved at the phonological level but the duration of the following stop closure and absence of vocal fold vibration are not: both voiced and voiceless stops become flaps. To account for this, any analysis needs to target specific submorphological units for preservation, rather than entire stems and affixes (cf. Prince & Smolensky 2004). These include what appear to be phonological units like the vowel height feature but can also include subphonemic characteristics of sounds (Steriade 2000). The existence of phonological and subphonemic preservation effects motivates our proposal that the root of stem preservation effects lies in motor perseveration rather than in the allegedly concatenative nature of morphology.

Another proposed motivation for paradigm uniformity falls into the long linguistic tradition of assuming the necessity of storage economy, sometimes derogatorily called the “Head Filling Up Problem” (Johnson 1997). As Kenstowicz (1998) put it, paradigm uniformity “finds plausible psychological motivation on the assumption that words are stored in memory in their surface phonetic form. To the extent that two instances of a given lexical item share the same phonological structure, the amount of space required to store the words in memory is minimized.” However, there still exists no psychological or neuroscientific evidence that long-term memory storage space is limited and needs to be saved (see also Householder 1966, Johnson 1997). Indeed, humans appear to be really good at storing vast amounts of information, down to individual episodes. For example, Shepard (1967), Standing et al. (1970) and Standing (1973) found that thousands of novel pictures were automatically memorized based on a single exposure trial and retained over several days, which is beyond the time span attributable to short-term or working memory. More recently, Brady et al. (2008) and Konkle et al. (2010) showed that these memories are not only numerous but also fairly detailed. For words, Palmeri et al. (1993) show that voice-specific memories are stored in an old/new recognition task, facilitating recognition of words when they are repeated by the same speaker compared to when the same words are repeated by a novel speaker. Interestingly, the same amount of voice-specific facilitation was obtained regardless of the number of distinct voices in the experiment (up to the tested maximum of 12). Based on this evidence, Palmeri et al. (1993) argue that long-term voice-specific memories of words are formed even if 12 distinct voice-specific memories per word would be required. In light of these impressive abilities, it is not clear that one extra form representation per paradigm would impose a significant load on long-term memory.\(^\text{11}\)

\(^{10}\) Unless it is part of the surface form of an activated morphological relative, in which case it is still activated less than the stem, which is part of all activated morphological relatives.

\(^{11}\) There is some evidence that formation of voice-specific word memories is not fully automatic, as it is more likely when those memories are specifically attended to because they are useful for performing the task; for example,
One limitation of the above accounts of PU is that they fail to account for the finding that PU favors small changes over large changes as well as favoring no change over change. In particular, Skoruppa et al. (2011) and White (2013, 2014) find that a change in two features is harder to learn than a change in one feature. It is not clear how this follows from storage economy: in both cases, a non-uniform paradigm would require more storage space. A preference against large changes provides motivation for accounts of PU based on avoidance of perceptual dissimilarity between allomorphs. There are two varieties of such accounts, one of which attributes PU to its benefits for word recognition (Kenstowicz 1996) while the other attributes it to a desire to get away with changing the language without anyone noticing (Steriade 2009).

Kenstowicz (1996) suggests that PU “serves to improve the transparency of morphological relationships between words and thus may enhance lexical access.” One way in which difficulties with lexical access could give rise to PU is by affecting word choice (cf. Berg 1998, Martin 2007). Namely, lack of PU can cause the listener to fail to interpret an unfamiliar word that could have been understood if only its morphological relatives were activated or the stem were parsed out. This may give rise to paradigm uniformity without implicating a bias against large changes: The word, having not been understood, will then be unlikely to be re-used by the listener when they speak, causing it to fall out of the language.

Alternatively, the speaker may actively avoid large changes for the benefit of the listener (i.e., because a large change would be hard for the listener to undo). A necessary caveat is that we do not know whether a large change, say one from [p] to [tʃ], is harder for a listener to undo than a smaller change (e.g. from [t] or [k] to [tʃ]). However, assuming this to be the case, the speaker could possibly avoid producing large changes because they are sensitive to possible lack of understanding that would result from such a change. There are some data suggesting that speakers avoid ambiguity in production by hyperarticulating cues that distinguish words from their minimal pair neighbors (Baese-Berk & Goldrick 2009, Wedel et al. 2013). On the other hand, there are also data suggesting that the speaker is not highly sensitive to homophony that results from an alternation, at least when learning alternations in the lab (Kapatsinski 2012, 2013).

Steriade (2009: 177) proposed a related perception-based bias account, arguing that: “innovators may aim to improve a sound system and... they do so in the safe regions of confusability identified by the P-map [a store of perceptual similarities between segments in context]. We assume, for instance, that speakers who initiate final devoicing have a choice of... spontaneously occurring speech variants to promote... and choose final devoicing because it involves the smallest departure from established speech norms.” On this account, the speaker introduces stem changes in the service of improving phonotactics but avoids introducing large changes because they may be noticed, and disapproved, by the listener.

when the listener’s task is speaker identification or old/new recognition but not if it is lexical decision (Theodore et al. 2015). Nonetheless, it appears clear that, at least for attended and potentially useful details, abundant storage in long-term memory is available.

12 Furthermore, if such a difference were found, we do not know how we would determine that it occurred for purely perceptual reasons (rather than the listener modeling production, as in analysis-by-synthesis and inverse model approaches).
There is a lively debate between speaker-internal and listener-oriented / audience-design explanations of speech production phenomena (e.g. Baese-Berk & Goldrick 2009, Galati & Brennan 2010, Jacobs et al. 2015, Kahn & Arnold 2015, Lam & Watson 2014). Our assessment of this debate is that, at the very least, the involvement of listener modeling in determining the articulatory details of pronunciation such as the degree of consonant voicing is unsubstantiated. The online involvement of perceptual representations in production is likewise not a settled matter (e.g. Perkell 2012). It is clear that modeling the perceiver is involved in slow and deliberately listener-oriented tasks such as poetry-writing (Steriade 2009). It must also be involved in retuning of production targets following target selection in auditory perturbation experiments (e.g. Perkell 2012, Purcell & Munhall 2006). However, such retuning is widely believed to be an offline process that follows the selection of the target (Villacorta, Perkell & Guenther 2007, Perkell 2012, cf. also Norris et al. 2003, Norris & McQueen 2008 for arguments against online feedback in perception). In contrast, the biases outlined by Kenstowicz (1996) and Steriade (2009) seem to require an influence of perceptual modeling prior to the selection of a production target during everyday speech production. We share the concerns of many that online perception guidance of production would be too slow to be consistent with how rapidly speech production, and other skilled motor action, unfolds (e.g. Elsner & Hommel 2001, Welsch & Llinas 1996). In any case, an online perceptual feedback mechanism appears unnecessary to account for word production, hence we would prefer not to rely on it to explain paradigm uniformity.

Moreton & Pater (2012a) identify possible analogs to the preference against large stem changes in work on category learning. In particular, they propose that acquiring an alternation involves categorizing the alternating sounds together and apart from sounds that do not participate in the alternation. For example, a [p]~[tʃ] alternation in the absence of [t]~[tʃ] and [k]~[tʃ] alternations would require categorizing [p] and [tʃ] into a single category that excludes [t] and [k]. Assuming that English [tʃ] is both [Coronal] and [Dorsal] (Yun 2006), this category might be [Labial] | ([Coronal] & [Dorsal]). On the other hand, a [k]~[tʃ] category excluding [t] and [p] would be describable simply as [Dorsal], a much simpler description. Arbitrary sound categories requiring multiple features to describe are harder to learn than those that can be described by a single feature (Cristià & Seidl 2008, Moreton et al. 2015, Pycha et al. 2003; see also Shepard et al. 1961, Feldman 2003 for non-linguistic categories). The connection between alternations and sound categorization is supported by the finding that the same sounds are perceived as being more similar by speakers of languages in which they are allophones of the same phoneme (e.g. Boomershine et al. 2008, Johnson & Babel 2010, Seidl et al. 2009).

The effects of perceptual category structure can plausibly account for a bias against large changes in judgments of mappings between forms (e.g., is [butʃi] the right plural for [bup]). When corresponding sounds in a judged word pair are all similar, the judge may think that the paired words go together. However, it is not clear why we should expect the same effect in production. It is not at all obvious that the alternating sounds also form a category in production, or that categorizing the alternating sounds together is useful for learning and, particularly, producing the alternation. Perceptual learning studies investigating transfer to production have so far focused on training participants to distinguish two sounds rather than training them to categorize two sounds together (e.g. Baese-Berk 2010, Bent 2005, Bradlow et al. 1997, Herd et al. 2013, Richtsmeier 2008, Wang et al. 2003). We are not aware of any work showing that training learners to categorize two sounds together (e.g. using the unimodal
distributional training procedure of Maye et al. 2002) would subsequently improve acquisition of an alternation involving the two sounds. In the absence of such data, increased perceptual similarity between sounds that alternate may also be a side effect of learning to produce the alternation.

**A production-internal learning bias against large changes**

As stated so far, our account does not necessarily predict a bias against *large* changes in production: while the speaker has a tendency to perseverate on the input, it is not clear that, say, replacing [p] with [tʃ] in [bup] → [butʃi] should be harder than replacing [k] with [tʃ] in [buk → butʃi]. In both cases, a change in the final position is required, and so CopyFin needs to be overridden. The question then is, overridden by what? We propose that learning to change an arbitrary structure X into an arbitrary structure Y often involves learning an association from X to Y, so that activating a form containing X would cause the activation of Y, allowing X to be changed into Y (see also Rumelhart & McClelland 1986). Interestingly, the associative learning literature suggests that acquiring an X → Y association is easier when X is similar to Y (Rescorla & Furrow 1977, Rescorla 1986; see also Moreton 2008a, 2012, Warker & Dell 2006, Warker et al. 2008 for related findings in phonotactic learning).

A plausible mechanism for why similarity should matter for associability has been proposed by Kapatsinski (2011), who notes that learning an association requires modifying synaptic connections between the associated representations. When the representations are dissimilar, they may be represented in different parts of the cortex, separated by more (or weaker) synaptic connections. More synaptic modification will then be required to associate the dissimilar representations with each other. This proposal is consistent with several findings in the literature on motor sequence learning where learning an association between X and Y appeared to involve changing the behavior of the cortical and subcortical areas separating X and Y (e.g. Hlušťík et al. 2004).

Kapatsinski (2011) and Moreton & Pater (2012a) thus both propose that similarity between alternants should influence learnability of the association but differ on 1) whether the effect of similarity on associability is mediated by categorization and 2) whether it is perceptual or articulatory similarity that is expected to matter (see also Cristià et al. 2013). While Moreton & Pater (2012a) are not explicit on the latter issue, Moreton’s (2008a: 85) conception of the phonological learner suggests that only perceptual similarity matters: only perceived forms serve as input to the learner module, which induces the grammar and the lexicon that drive production, and there is no feedback to the learner (‘analysis’) module from production except through the perception system. In our view, learning happens throughout the brain (modifiable synapses are everywhere), and most learning relevant for successfully performing alternations is within the production system. There also lives the perseveratory inertia to be overcome in performing a stem change. For this reason, we expect paradigm uniformity to be stronger in production than in judgment while the opposite might be expected if similarity is a learning bias localized to a learner module that deals with perceptual inputs.

**Summary of the theoretical proposal**

---

13 We say *often* because paradigmatic form-form associations are not necessary when the target meaning can only be expressed by a single surface form.
How do we form new forms of known words? Or, more generally, what happens when lexical retrieval cannot generate a context-appropriate form (in time)? On the present proposal, the speaker generates a form to express the desired meaning by using

1) meaning→form associations; (similar to the product-oriented/first-order schemas of Bybee 1985, 2001, Kapatsinski 2012, 2013, Nesset 2008 and constructions of Goldberg 2003);
2) paradigmatic form-form associations (similar to second-order schemas necessary for arbitrary paradigmatic mappings emphasized by Booij 2010, Gouskova & Becker 2013, Nesset 2008, Pierrehumbert 2006);
3) copying from the activated wordforms (similar to output-output faithfulness constraints of Benua 1997 and Kenstowicz 1996); and
4) a mechanism for maintaining and re-creating serial order (here assumed to be a word-sized prosodic template growing with learning to fit the range of experienced wordforms, cf. Redford 2015, Vihman & Croft 2007).

In the present work, we focus on the interaction between mechanisms (2) and (3) but see Kapatsinski (Submitted) for the full model. We propose that paradigmatic form-form associations are harder to learn when the to-be-associated forms involve different articulators, and thus highly dissimilar neuromotor representations. Further, we propose that these associations compete with a perseveratory tendency to simply output the activated neuromotor representations, which varies in strength by context in a language-specific manner as a function of linguistic experience. When a paradigmatic association is too weak to override the competing perseveratory tendency, a stem change is leveled and paradigm uniformity arises.

**Palatalization: The test case**

In order to establish a production locus for a learning bias against major changes to the base, the present paper examines the learning of miniature artificial languages differing in the magnitude of base changes they require (cf. also Peperkamp et al. 2006, Skoruppa & Peperkamp 2011, White 2013, 2014, and White & Sundara 2014). We focus on the acquisition of palatalization processes, in which stops become alveopalatal affricates before the suffix [i].

Cross-linguistically, palatalization before [i] is more common than palatalization before any other vowel (Bateman 2007, Kochetov 2011). The relative ubiquity of palatalization before [i] has been argued to be motivated on both articulatory (e.g. Anttila 1989: 72-73, Hock 1991:73-77) and perceptual grounds (Guion 1998, Ohala 1989:183-185, 1992: 320). It has also been argued to be easier to learn than palatalization before other vowels in both artificial and natural language learning (Mitrović 2012, Wilson 2006).

Articulatorily, [tʃ] and [dʒ] feature both coronal and dorsal articulations (Yun 2006), which makes them similar to both coronal and dorsal (velar) but not labial stops. These relations are also paralleled by patterns of perceptual similarity. Typologically, palatalization of coronals and velars is about equally common while palatalization of labials is extremely rare (Bateman 2007, Bhat 1978, Chen 1973, Kochetov 2011). Kochetov (2011) proposes an implicational universal: labial palatalization implies palatalization of coronals and/or velars while Bateman (2007) suggests that there are no clear cases of productive labial palatalization in which the labial articulation is fully suppressed in the palatalized outcome.
Typological frequency often maps on to how easy a pattern is to learn (e.g. Finley 2008, Mitrović 2012, Schane et al. 1975, White 2013, 2014, White & Sundara 2014, Wilson 2006). The typological data thus strongly suggest that there may be a learning bias against labial palatalization that would cause labial palatalization to either not be learned or to be overgeneralized to coronals and/or velars, resulting in patterns obeying the implicational universal.

In addition, American English listeners have been shown to often misperceive [ki] as [tʃi] in noise, while misperceiving [gi] as [dʒi] is significantly less common (Guion 1998). Guion (1998:20) suggests that this perceptual difference accounts for an implicational universal: while 40% of velar palatalization cases involve only the voiceless [k], there are no cases in which only the voiced velar [g] is palatalized (Bhat 1978). If typological frequency maps onto learnability here as well, palatalization of [k] should be harder to learn than palatalization of [g], at least in the context of a following [i]. However, there are also cases in which typological frequency does not seem to correlate with a difference in learnability (e.g. Olejarczuk & Kapatsinski 2015 on the relation between stress and syllable weight, Pycha et al. 2003 and Skoruppa & Peperkamp 2011 on vowel harmony vs. disharmony, Seidl & Buckley 2005 for intervocalic lenition vs. hardening, Zaba 2008 on consonant harmony vs. vowel harmony; see Moreton & Paternoster 2012 for a review). Given these data, it seems plausible that only some differences in typological frequency of synchronic alternation patterns are due to differences in learnability of these patterns. Often, typological frequency differences are instead due to differences in the frequencies of diachronic change pathways resulting in the alternations (Blevins 2006, Bybee 2001).

In the present case, the difference in learnability of palatalization is expected for [k] and [g] if learnability of palatalization is driven by perceptual (dis)similarity between the source and the product (Hayes & White 2015, Kenstowicz 1996, Steriade 2009). However, in previous work, Wilson (2006) found English speakers to palatalize [g] more than [k], contrary to this expectation. Articulatorily, [k] and [tʃ] are as similar as [g] and [dʒ]. Therefore, no difference in learnability is expected between [k]→[tʃi] and [g]→[dʒi] if learnability of palatalization is due to articulatory (dis)similarity between the input and the output. Any observed differences must be due to first-language experience. Wilson’s (2006) finding that [g] is easier to palatalize for English speakers might be explained by the fact that the English letter <g> is often pronounced [dʒ], while <k> and <*> are rarely pronounced as [tʃ] (Gontijo et al. 2003). The influence of orthography is apparently insufficient to override the greater acoustic similarity between [ki] and [tʃi] in its effects on perception. So [gi] and [dʒi] are still perceptually more distant for American English listeners than [ki] and [tʃi] (Guion 1998). However, orthography appears to be enough to override any effect of perceptual similarity on the learnability of the alternations. While it cannot rule out an effect of perceptual similarity on associability, a replication of Wilson (2006) would suggest it to be relatively minor.

Previous experimental data

Only a handful of papers have experimentally examined the learning of alternations while manipulating the distance between the alternates. Skoruppa et al. (2011) and White (2013, 2014) examined the learning of alternations involving place, voicing and manner of obstruents (e.g., [p]~[t] vs. [p]~[s]). Both studies found that alternations involving a change in only one feature were easier to learn (and perform) than alternations involving a change in more than one feature. In both studies, changes in more than one feature also involved ‘jumping over’ an intermediate sound (a saltatory alternation). However, no study has explicitly compared learnability of the same change for learners that possess the intermediate
sound in their first-language sound inventory, and those that do not. Therefore it is not yet clear whether the existence of such an intermediate sound is necessary for a large change to be harder to learn or perform than a smaller change. We therefore take these previous results as indicating that large changes are harder than small changes generally.

Under the Perseveration Hypothesis: large changes are harder to perform than small changes. In addition, learning an association between dissimilar sounds is harder, resulting in a slower learning rate. Skoruppa et al. (2011) examined the learning rate for the to-be-changed sounds and found that it was indeed slower for larger changes.

White (2013, Chapter 3, 4.5; 2014) instead trained participants to criterion on the alternating sounds and examined overgeneralization to intermediate sounds, e.g. whether a [p]~[v] alternation is taken to imply a [b]~[v] alternation. He found that exposure to a 'saltatory' change like [p]→[v], which involves 'jumping over' [b] or [f], led participants to produce [b]→[v]. In contrast, exposure to a 'non-saltatory' change [b]→[v] led to very few productions of [p]→[v]. White (2013, 2014) takes these results as suggesting that large saltatory changes are taken by participants to imply that intermediate sounds 'jumped over' in training also change, e.g. [p]→[v] implying [b]→[v]. However, White (2013, p.83) notes that large changes may also be harder to perform because many participants in the saltatory (large change) condition had to be excluded for never reaching criterion accuracy on the trained segments or failing to perform changes on the trained segments during test (17/33 vs. 2/22).

In our view, training participants to criterion on the to-be-performed changes is helpful for revealing categorization biases but obscures the nature of the bias behind PU, which is that large changes are hard to perform and dissimilar sounds are hard to associate. Indeed, White’s (2013) model of the results suggests that large changes should be harder to learn without necessarily implying overgeneralization to sounds 'jumped over': White (2013) suggests that learning to change [p]→[v] involves changing the weight of a *Map(p,v) constraint (Zuraw 2007) while changing [b]→[v] involves changing the weight of a *Map(b,v) constraint, and that the former constraint is a priori stronger than the latter because [p] is less perceptually similar to [v] than [b] is. The Perseveration Hypothesis agrees that the source of difficulty with changing [p] into [v] is the magnitude of the change, while suggesting that the magnitude of the change should be measured in articulatory, rather than perceptual terms.

White (2013, Chapter 5; White & Sundara 2014) presents additional results in support of a bias against saltatory alternations in two-year-old infants that cannot easily be accounted for by the Perseveration Hypothesis. Infants were exposed to training stimuli in one of four conditions described in (6)-(9). Crucially, infants exposed to (6) and (7) were exposed to saltatory alternations, either [p]~[v] in (6) or [t]~[z] in (7), and tested on trials involving [b]~[v] and [d]~[z] alternations. Participants exposed to (6) paid more attention -- as measured by looking time -- on test trials featuring [d]~[z] alternations (e.g. *dagu-*zag*u*) while those exposed to (7) paid more attention to test trials featuring [b]~[v] alternations (e.g., *bagu-*vag*u*). White & Sundara conclude that participants exposed to (6) learned that [p], [b] and [v] are one phoneme, while [t], [d] and [z] are different phonemes, while participants exposed to (7) learned the opposite. Thus, the sound 'jumped over' in a saltatory alternation was apparently grouped together with the sounds involved in the alternation.

(6) {rom;na} [t;z]VCV and rom pVCV and na vVCV
(7) {rom;na} [p;v]VCV and rom tVCV and na zVCV
(8) {rom;na} [d;z]VCV and rom bVCV and na vVCV
(9) \{rom;na\} \{b;v\}VCV and rom dVCV and na zVCV

We take White & Sundara’s (2014) results to support the proposal that alternating sounds are categorized together in perception, and that this category includes intermediate sounds (unless contradictory evidence is provided). In other words, we take this to be a special case of a bias against discontinuous categories in perception (e.g. Maddox et al. 2005, 2007). This bias is quite distinct from the production-internal biases that are the focus of the Perseveration Hypothesis, but we do not think that it obviates the need for a bias against large changes in production. The primary difficulty faced by learners of paradigms, once they start producing novel forms of known words, is not avoiding changing intermediate sounds but rather learning to change the sounds that should be changed.

Similarly, we suggest that saltatory alternations are rare in languages mostly not because the ‘jumped over’ sounds come to alternate along with the sounds that have originally participated in the alternation. Rather, they are rare because of a production bias against large changes, of which saltatory changes are a special case. The large changes that can ‘seed’ a saltatory alternation are more likely to never arise, or to die out once arisen, than to spread to other sounds.

Two studies examined a bias against large changes with palatalization patterns. First, Wilson (2006) failed to find a preference for [k] \(\rightarrow\) [tʃi] over [g] \(\rightarrow\) [dӡi] predicted by perceptual similarity / confusability. Second, in our own work (Stave, Smolek & Kapatsinski, 2013), we found a preference against [p] \(\rightarrow\) [tʃa] compared to [t] \(\rightarrow\) [tʃa] or [k] \(\rightarrow\) [tʃa]. As shown in Figure 2, participants palatalized [p] less than [t] and [k] despite palatalization of all kinds being phonetically unmotivated in this context (between [a]’s).

Together with Skoruppa et al. (2011), these results suggest that large changes are harder to perform than small changes, even when the changes being compared are ‘phonetically unnatural’ because they do not improve the phonotactics of the form.

**Figure 2.** Palatalization rates between [a]’s in Stave, Smolek & Kapatsinski (2013)
The contrast between the null effects of change magnitude in Wilson (2006) and very strong effects of change magnitude in Stave, Smolek & Kapatsinski (2013) suggests that it is articulatory change magnitude -- controlled in Wilson (2006) -- that affects how easy the change is to (learn to) perform (see also Cristià et al. 2013), as predicted by the Perseveration Hypothesis. In the present paper, we examine the influence of articulatory and acoustic change magnitude within a single experiment, by comparing labial and lingual palatalization in the natural pre-[i] context.

A preview of results

In the present paper we contrast labial palatalization (a large change) with lingual palatalization (smaller changes) in the pre-[i] context. The results are consistent with the results in the between-[a] context in Figure 2 and suggest a production basis for paradigm uniformity. Participants in the labial condition tend to palatalize nothing while participants in other conditions palatalize the consonants they are trained to palatalize. At the same time, palatalized outputs are preferred in judgment following labial training, despite being rarely produced. We argue that the results are problematic for perceptual accounts of the bias against large changes. The high perceptual similarity of [k] and [tʃi] compared to [g] and [dʒi] does not help palatalization of [k] or retard palatalization of [g]. In fact, as in Wilson (2006), palatalization of [g] is slightly preferred over palatalization of [k].

Experimental data

In the experiment, participants were exposed to either Labial Palatalization, Alveolar Palatalization or Velar Palatalization, as illustrated in Table 1. Singulars were always C(C)V(C) forms ending in oral stops. Plurals were formed by adding the plural suffixes –i and -a. Depending on the condition to which a participant was assigned, either Labial, Alveolar or Velar stops were palatalized, becoming [tʃ] if voiceless and [dʒ] if voiced.

<table>
<thead>
<tr>
<th></th>
<th>Labial Palatalization</th>
<th>Alveolar Palatalization</th>
<th>Velar Palatalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singular</td>
<td>Plural</td>
<td>Plural</td>
<td>Plural</td>
</tr>
<tr>
<td>...p</td>
<td>...tʃi</td>
<td>...{pi;pa}</td>
<td>...{pi;pa}</td>
</tr>
<tr>
<td>...b</td>
<td>...dʒi</td>
<td>...{bi;ba}</td>
<td>...{bi;ba}</td>
</tr>
<tr>
<td>...t</td>
<td>...{ti;ta}</td>
<td>...tʃi</td>
<td>...{ti;ta}</td>
</tr>
<tr>
<td>...d</td>
<td>...{di;da}</td>
<td>...dʒi</td>
<td>...{di;da}</td>
</tr>
<tr>
<td>...k</td>
<td>...{ki;ka}</td>
<td>...{ki;ka}</td>
<td>...tʃi</td>
</tr>
<tr>
<td>...g</td>
<td>...{gi;ga}</td>
<td>...{gi;ga}</td>
<td>...dʒi</td>
</tr>
</tbody>
</table>

Participants’ knowledge of the language was tested using elicited production and judgment tests requiring generalization to novel singulars. The production test involved hearing a novel singular and saying the corresponding plural form. The judgment test involved hearing singular-plural pairs and judging whether the plural form is the right one for this singular form.

We had several expectations for this experiment, some expected under any account in which PU influences learnability of alternations, others expected only under some accounts of PU. In particular,
under the Perseveration account of PU, we expect Hypotheses 1, 2 and 3 below to be supported, and Hypotheses 4 and 5 to lack support. The categorization explanation for PU based on featural simplicity (Moreton & Pater 2012a) predicts support for Hypotheses 2 and 4 and lack of support for Hypotheses 1, 3, and 5. The perceptual explanations for PU (Kenstowicz 1996, Steriade 2009) predict support for Hypotheses 1, 2, 4, and 5, and lack of support for Hypothesis 3.

**Hypothesis 1:** A learning bias against labial palatalization in production. Labial palatalization should be harder to learn than alveolar or velar palatalization. In particular, the difference between palatalization rates for the to-be-palatalized consonant and the not-to-be-palatalized consonants should be smaller in the Labial condition than in the other conditions. This finding is expected if PU influences learnability, for whatever reason. Furthermore, if the bias is a bias against changing labials, [p] should be palatalized less in the Labial condition than [t] is palatalized in the Alveolar condition or [k] is palatalized in the Velar condition. In addition, [p] should be palatalized less in the Alveolar and Velar conditions than [t] and [k] are palatalized in the Labial condition. This outcome is expected if there is a bias against large changes, which is predicted by the Perseveration Hypothesis and by the perceptual approach (Kenstowicz 1996, Steriade 2009). Whether the relevant notion of similarity is articulatory or perceptual, labials are less like alveopalatalas than linguals are. Therefore, a bias against large changes should result in lower rates of palatalization for the to-be-palatalized consonants in the labial training condition and for not-to-be-palatalized labials compared to non-labials in the other conditions.

**Hypothesis 2:** Labial palatalization is hard to learn because of faithfulness, not markedness. Judgments of unfaithful mappings will differ across conditions but judgments of faithful mappings will not. In other words, p~tʃ is worse than k~tʃ or t~tʃ but k~ki and t~ti are as good as p~pi. This finding must hold for us to argue that the bias against labial palatalization is a bias against certain changes. If Hypothesis 2 does not hold, any condition differences we obtain may be due to markedness differences between [pi] on the one hand and [ti] and [ki] on the other, which are fixed by palatalization. In particular, if [ki] and [ti] are worse (i.e., more marked) than [pi], palatalization of [t] and [k] is expected to be easier to learn than palatalization of [p] on these grounds alone (Pater & Tessier 2006).

**Hypothesis 3:** The learning bias is internal to the production system and is therefore stronger in production than in judgment. The difference between the Labial condition and the other conditions in (1) should be greater in the production test than in the judgment test. This is expected under the perseveration hypothesis advanced in the present paper. Particularly, we expect labial palatalization to be judged acceptable but rarely produced.

**Hypothesis 4:** Labial palatalization is especially likely to be overgeneralized. Under this hypothesis, there should be more palatalization of [k] in the Labial Palatalization condition than in the Alveolar Palatalization condition, and more palatalization of [t] in the Labial Palatalization condition than in the Velar Palatalization condition. This outcome is expected on Moreton & Pater’s (2012a) categorization account (see also White 2013). Under this account, the learner in the labial condition attempts to acquire a simple conjunctively defined category subsuming [p] and [tʃ], e.g. [-continuant]. This category would also subsume [t] and [k], causing them to be palatalized at similar rates to [p]. In contrast, a category subsuming [t] and [tʃ] ([Coronal; -continuant]) may not subsume [k] and a category subsuming [k] and [tʃ] ([Dorsal; -continuant]) may not subsume [t].

**Hypothesis 5:** The bias against labial palatalization is a bias in favor of perceptual similarity between alternants. Under this hypothesis, [k] should be palatalized less than [g] in all conditions. This is
expected under perceptual explanations for paradigm uniformity (Hayes & White 2015, Kenstowicz 1996, Moreton & Pater 2012a, Steriade 2009) but not under the Perseveration account.

**Methods**

**Participants**

107 undergraduate students in introductory psychology or linguistics courses at the University of Oregon were recruited through the Human Subject Pool and received partial course credit for participation. 11 subjects were excluded for producing plurals that didn’t correspond to the patterns in the training\(^{14}\). This left 32 subjects in the Alveolar condition, 31 in the Labial, and 33 in the Velar. All reported being native speakers of English, with no speech, hearing, language or learning disabilities.

**Training**

In every condition, there were 28 unique singular forms of words which were randomly paired with Spore creatures selected from the database of user-created images at sporepedia.com. Each creature was shown at least once alone, and at least once as part of a group (the same image copied and pasted multiple times); the images containing a single creature were accompanied by a recording of the singular form, and the images containing groups of creatures were paired with the corresponding plural form. All word-picture pairings appeared in random order, so the corresponding singulars and plurals were rarely temporally adjacent. In previous work (Kapatsinski 2012), this ordering was found to encourage overgeneralization compared to presenting singular-plural pairs, presumably by making it less obvious which singular-final consonants map onto [	extipa{tʃ}] and [	extipa{dʒ}].

The singular forms for all tokens were single-syllable roots of C(C)V(C) structure, where the word-final consonant was an oral stop. 12 out of the 28 training words ended with the to-be-palatalized consonants (Alveolar contained 6 each of stems ending with [t] and [d], Labial of [p] and [b], and Velar of [k] and [g]), with the remaining 16 split evenly between the other two place*voice combinations. Stimuli are shown in the appendix.

**Production Test**

For the production test portion of the experiment, we selected an additional 92 creatures from the Spore Creature Creator database. As before, we copied and pasted the creatures such that, for each creature, there was one slide showing a single creature and one slide showing a group of creatures. The singular slide was again accompanied by the recording of the singular form, but the plural slide played no recording. Instead, subjects were told to say what they thought the plural form for the singular should be, which we recorded for later coding. 36 of the 92 singulars ended with the consonants subjects had been trained to palatalize (half voiced, half voiceless), with the remaining 56 split evenly between the other two place*voice combinations. Stimuli are shown in the appendix.

Each plural recording from the subject was saved as a separate file and coded by either the first author or undergraduate RAs, whose codings were all checked by the first author for accuracy. We coded in particular for the stem-final consonant and the word-final vowel in the plural form. If a participant replaced the stem-final consonant with either of the palatals (e.g. blastp → blast{tʃ;dʒ}{i;a}), we coded

\(^{14}\) e.g., by repeating the singular form, adding the English plural –s, or as in the case of one particularly memorable subject, producing plurals only tangentially related to the singular form, as in klip → Cleopatra.
the form as being palatalized; if the participant retained the same consonant in the plural (e.g. \textit{blap} \rightarrow \textit{blap(i;a)}) , the form was coded as not palatalized. If the stem consonant was retained but the palatal was also added (e.g., \textit{blap} \rightarrow \textit{blappt(i;a)}), we also categorized the form as not palatalized. While these products fit the “plurals should end in tʃi” schema, they preserve the final consonant of the base, obeying PU. Biases to preserve aspects of the base should therefore favor these kinds of plurals alongside plurals like \textit{blap} \rightarrow \textit{blap(i;a)}. Rarely, a subject would replace the stem consonant with another non-palatal (e.g. \textit{blaip} \rightarrow \textit{blaida}); such responses are excluded from all analyses. Instances where subjects added the English plural -\textit{s} were also excluded, and subjects who produced a majority of such responses were excluded entirely.

\textit{Judgment Test}

Since the judgment task exposes participants to plural forms that contradict training, the judgment task followed the production task. The judgment session was the same across all conditions. 30 new singualrs were created, divided equally between the three places and two voices. For each singular, we recorded 4 plural forms, palatalizing vs. not before –i vs. before –a (e.g., for the alveolar stem \textit{prut}, the plural forms were \textit{pruti}, \textit{pruta}, \textit{prutʃi}, and \textit{prutʃa}). For each ratings trial, the picture of a single creature appeared first with the recording of the singular form, followed by a picture of a group of the same creatures with the recording of one of the plural forms.

Subjects were instructed to indicate, using a button box, whether the plural they heard was the right plural form for the singular. Each subject received a different random order of the 120 singular-plural pairings.

\textit{Dependent Variables for Judgment Data}

The button box had five buttons but the participants’ responses were strongly bimodal: 59% of the ratings were either ‘1’ or ‘5’, 29% were ‘2’ or ‘4’ and only 12% were ‘3’. Because of this, and for comparison with the inherently binary production task data, we transformed the obtained ratings into binary dependent variables. Two such variables were created:

1. Absolute Judgment, which was simply the binarized rating, so for every singular-plural pair we entered ‘1’ (accepted) for ratings above 3 or ‘0’ (rejected) for ratings below 3. We excluded the 12% of ratings that were ‘3’, since we took them as an indication of indecision.

2. Relative Judgment, which was the binarized difference of ratings between the palatalized and non-palatalized plurals with the same base and the same suffix by a particular subject, e.g. the rating of \textit{bup}~\textit{bupi} minus the rating of \textit{bup}~\textit{butʃi}. It was possible for a subject to rate both plurals equally, in which case the trial was excluded.

The Absolute Judgment variable allows us to examine the effects of condition on judgments of palatalization and faithful mappings separately. This is essential to evaluate Hypothesis 2, that the observed bias is driven by faithfulness rather than markedness.

The Relative Judgment variable was created to compare judgment to production in order to evaluate Hypothesis 3, which requires comparing production to judgments. It embodies the hypothesis that palatalization is produced when it is more acceptable than lack of palatalization. However, given Hypothesis 2 (no judgment differences for faithful mappings across conditions), we expected to see the
effect of Condition to be larger in Production than in Judgment whether absolute or relative Judgments are used. This was indeed the case in our data, so only absolute judgment results are reported here.

Statistical analysis

All statistical analyses were conducted using generalized (logistic) linear mixed-effects models by means of the lme4 package, (version 1.1-9, Bates et al. 2015) in R (version 3.1.1, R Development Core Team, 2014). We included random intercepts for Subjects and Bases, and random slopes for Condition (Training on Alveolar vs. Velar Palatalization; Labial vs. Alveolar or Velar Palatalization) within Base, and for Plural Vowel (-i vs. –a), Test Place (To-be-palatalized vs. Not-to-be-palatalized given the training), Test Voice (Voiced vs. Voiceless), Test Type (Production vs. Judgment), and interactions between them within Subjects. We followed Barr et al. (2013) in including all possible random slopes for all predictors and interactions we hope to find significant under the perseveration hypothesis (To-be-palatalized vs. Not-to-be-palatalized Test Place, Test Type, Labial vs. Other Palatalization Training Condition). When a model with maximal random effects structure would not converge, we removed slopes associated with predictors that should be significant under other researchers’ explanations for paradigm uniformity (Plural Vowel, Test Voice, Test Voice/Test Place interaction). This biases the analysis against our hypothesis by giving predictors we expect not to be significant a leg up over ones we do hope to find significant. Tested models are presented in footnotes throughout the results section. The full dataset and analysis code are available at.

https://www.dropbox.com/sh/rdzn3vitrfficaw/AABjRrDLaVxEcJJTkzF4h37va?dl=0.

Results and Discussion

Hypothesis 1: A bias against large changes (changing labial into alveopalatals) in production

In order to evaluate Hypothesis 1, we compared the differences in palatalization rates and judgments between to-be-palatalized and not-to-be-palatalized consonants in the Labial Palatalization condition and the two lingual palatalization conditions. For production data (palatalization rates), the results of this comparison are shown in Figure 3. Figure 3 shows that there is a large difference between Labial Palatalization and the lingual conditions in the expected direction: participants learned to palatalize velars when trained on velar palatalization, and alveolars when trained on alveolar palatalization. However, they did not learn to palatalize labials any more than non-labials when exposed to labial palatalization. In fact, participants exposed to labial palatalization palatalized labials slightly less than non-labials. Figure 3 also shows that the differences between conditions are largely due to how often the to-be-palatalized consonants are palatalized (light bars), as expected under Hypothesis 1, rather than in how much palatalization is overgeneralized to the not-to-be-palatalized consonants (dark bars), as one would expect from Hypothesis 4.

Statistically, we observe that to-be-palatalized base consonants are palatalized significantly more often (i.e., retained significantly less often) in the alveolar and velar palatalization conditions relative to the labial palatalization condition (Table 2), with no overall effect of voicing and no interactions with voicing.
Hypothesis 1, a bias against changing labials into alveopalatals, predicts that Labials should be palatalized more than the linguals when they should be palatalized, and also when they should not be palatalized. The first half of the claim is confirmed by Figure 3 and Table 2. The second half of the claim is evaluated in Figure 4 and Tables 3 and 4. As seen in Figure 4, the second half of the claim is also confirmed: labials are palatalized in error less than other stops are palatalized in error.
**Figure 4.** Overgeneralization of palatalization. Overgeneralization to labials in dark bars vs. overgeneralization to linguals in light bars. Left panel: Overgeneralization of alveolar palatalization to labials (dark) vs. overgeneralization of labial palatalization to alveolars (light). Right panel: Overgeneralization of velar palatalization to labials (dark) vs. overgeneralization of labial palatalization to velars (light).

### Table 3. Alveolars are palatalized more after training on labial palatalization than labials are palatalized after training on alveolar palatalization.

<table>
<thead>
<tr>
<th></th>
<th>$b$</th>
<th>$se(b)$</th>
<th>$z$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>4.2555</td>
<td>0.9343</td>
<td>4.555</td>
<td>5.24E-06 ***</td>
</tr>
<tr>
<td>Voiceless</td>
<td>0.7503</td>
<td>0.7457</td>
<td>1.006</td>
<td>0.31432</td>
</tr>
<tr>
<td>Alveolar Stem</td>
<td>-3.3044</td>
<td>1.0073</td>
<td>-3.28</td>
<td>0.00104 **</td>
</tr>
</tbody>
</table>

### Table 4. Velars are palatalized more after training on labial palatalization than labials are palatalized after training on velar palatalization.

<table>
<thead>
<tr>
<th></th>
<th>$b$</th>
<th>$se(b)$</th>
<th>$z$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>3.8124</td>
<td>0.784</td>
<td>4.863</td>
<td>1.16E-06 ***</td>
</tr>
<tr>
<td>Voiceless</td>
<td>0.9275</td>
<td>0.7749</td>
<td>1.197</td>
<td>0.23132</td>
</tr>
<tr>
<td>Velar Stem</td>
<td>-2.3129</td>
<td>0.8975</td>
<td>-2.577</td>
<td>0.00996 **</td>
</tr>
</tbody>
</table>

To summarize the results so far, the data strongly support a bias against labial palatalization, driven primarily by how difficult labials are to change into alveopalatals. This difficulty lowers the palatalization rate of to-be-palatalized labials for participants trained on labial palatalization as well as the not-to-be-palatalized labials in the other training conditions.
Hypothesis 2: The bias against labial palatalization is not due to markedness

Before we conclude that the bias against labial palatalization is about changes, we need to consider the alternative interpretation that it is a side effect of the markedness of [ti] and [ki] compared to [pi]. Perhaps, [t] and [k] are easy to change into [tʃ] compared to [p] before [i] because participants come to the experiment ready to avoid [ti] and [ki] but have no pre-existing aversion to [pi]. If this were true, we would expect a difference in judgments of no-change plurals in which [t], [k], or [p] is retained before [i]. Under the markedness hypothesis, we would expect judgments of p~pi in the labial condition to be higher than judgments of t~ti in the alveolar condition or k~ki in the velar condition. However, Figure 5 shows no significant difference in ratings of no-change plurals for to-be-palatalized consonants across conditions and, if anything, a slight trend in the unexpected direction for not-to-palatalized consonants, which is also not significant. These results are inconsistent with the markedness explanation: the bias is against certain changes, not against certain output structures.

Figure 5. Judgments of faithful plurals do not significantly differ across conditions.

Table 5. There is no significant difference in judgments of incorrect faithful mappings for to-be-palatalized consonants across training conditions.

<table>
<thead>
<tr>
<th></th>
<th>b</th>
<th>se(b)</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.47852</td>
<td>0.26176</td>
<td>-1.828</td>
<td>0.0675</td>
</tr>
<tr>
<td>-a</td>
<td>0.29992</td>
<td>0.21329</td>
<td>1.406</td>
<td>0.1597</td>
</tr>
<tr>
<td>Labial Training</td>
<td>-0.05402</td>
<td>0.35812</td>
<td>-0.151</td>
<td>0.8801</td>
</tr>
<tr>
<td>Voiceless</td>
<td>-0.01851</td>
<td>0.15493</td>
<td>-0.12</td>
<td>0.9049</td>
</tr>
</tbody>
</table>

Table 6. There is no significant difference in judgments of correct faithful mappings for not-to-be-palatalized consonants across training conditions.

<table>
<thead>
<tr>
<th></th>
<th>b</th>
<th>se(b)</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.07396</td>
<td>0.22282</td>
<td>0.332</td>
<td>0.74</td>
</tr>
<tr>
<td>Labial Training</td>
<td>-0.18202</td>
<td>0.1598</td>
<td>-1.139</td>
<td>0.255</td>
</tr>
<tr>
<td>Voiceless</td>
<td>0.11153</td>
<td>0.17453</td>
<td>0.639</td>
<td>0.523</td>
</tr>
</tbody>
</table>
Hypothesis 3: The learning bias is internal to the production system. Hence it should be stronger in production than in judgment.

The perseveration account of paradigm uniformity suggests that the bias against labial palatalization should be stronger in production than in judgment. We do not rule out effects of perseveration in judgment to the extent that the judgment is based on a model of production (‘would I say this? let me try and see how hard it is’). Nonetheless, we expect these effects to be weaker, since the judge is primed with the output, palatalized form, hence whatever difficulty they would face producing the form is alleviated.

As we saw in Figure 5 above, ratings of faithful plurals do not differ across conditions. Thus, any judgment differences between conditions we could see must come from judgments of unfaithful forms featuring palatalization. These data are shown in the left panel of Figure 6, side by side with production data from Figure 3 repeated here in the right panel.

The comparison reveals several findings. First, there is a striking dissociation between production and judgment with training on labial palatalization (left bars): labial palatalization is typically accepted but not produced (as shown by the significant interaction in Table 7). This dissociation between judgments and production is not there for training on alveolar and velar palatalization: when trained to palatalize alveolars or velars, participants produce palatalization as often as they accept it; when trained to palatalize labials, they accept palatalization but do not produce it. This is the main piece of evidence for why we consider the bias against labial palatalization to be a production bias.

Second, also in agreement with the production-internal nature of the bias against labial palatalization, there is less of a difference between training conditions (Labial vs. Alveolar or Velar palatalization) in the judgment test. This is seen in the three-way interaction in Table 8. This interaction arises for three reasons. First, palatalization of to-be-palatalized consonants is produced about as often as it is accepted but palatalization of not-to-be-palatalized consonants is accepted half the time or more even with alveolar or velar training, where it is seldom produced. This is consistent with the position that the initial state of a language learner is that everything is acceptable but everything is hard to produce (Boersma 1998, Kapatsinski 2013). Finally, to-be-palatalized velars are palatalized more than to-be-palatalized labials but correct palatalization of velars is accepted as often as correct palatalization of labials. Overall, the results are in line with Hypothesis 3: changes that are rarely produced despite being observed in training are nonetheless judged acceptable, presumably because the resulting structure has been associated with the plural meaning (Kapatsinski 2013). The bias against changing labials is strong in production but is attenuated in judgment data.
Figure 6. Acceptance of labial palatalization in perception (left) vs. lack of labial palatalization production (right)

Table 7. The effects of training on Labial vs. Non-labial palatalization on judgment vs. production of palatalized forms

<table>
<thead>
<tr>
<th></th>
<th>b</th>
<th>se(b)</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.74682</td>
<td>0.257172</td>
<td>-2.904</td>
<td>0.00368 **</td>
</tr>
<tr>
<td>Voiceless</td>
<td>-0.00154</td>
<td>0.108858</td>
<td>-0.014</td>
<td>0.98875</td>
</tr>
<tr>
<td>Labial Training</td>
<td>4.075962</td>
<td>0.739281</td>
<td>5.513</td>
<td>3.52E-08 ***</td>
</tr>
<tr>
<td>Judgment Test</td>
<td>-0.10839</td>
<td>0.220618</td>
<td>-0.491</td>
<td>0.62322</td>
</tr>
<tr>
<td>Labial Training x Judgment Test</td>
<td>-4.18717</td>
<td>0.695858</td>
<td>-6.017</td>
<td>1.77E-09 ***</td>
</tr>
</tbody>
</table>
Table 8. The effects of training on Labial vs. Lingual palatalization on correct vs. erroneous palatalization and judgments of correct vs. erroneous palatalization. Bolded rows show effects of interest.

Hypothesis 4: Labial palatalization is especially likely to be overgeneralized because it causes learners to group together labials and linguals.

In order to evaluate this hypothesis, we ask whether alveolars are palatalized more when participants are trained to palatalize labials than when they are trained to palatalize velars. Similarly, are velars palatalized more when participants are trained to palatalized labials than when they are trained to palatalize alveolars? An affirmative answer is expected according to the categorization explanation for paradigm uniformity. If labials and alveopalatals are to be classified together into a natural class, then velars and/or alveolars should be part of the class as well, since alveopalatals are both [Dorsal] and [Coronal] (Yun 2006) and labials are neither. Therefore, overgeneralization of labial palatalization to velars ([Dorsal]) and/or alveolars ([Coronal]) is expected. In contrast, overgeneralization from [Coronal] – {t;d;ʃ;dʒ} – to [Dorsal] [k;g] or from [Dorsal] {k;g;ʃ;dʒ} to [Coronal] [t;d] is not strictly necessary.

Figure 7 summarizes the results. Contrary to Hypothesis 4, there was no significant difference in palatalization rates of alveolars between the labial and velar training conditions (b = 0.60, se(b) = 0.74, z = 0.81, p = 0.42). Likewise, there was no significant difference in palatalization rates of velars between the labial and alveolar training conditions (b = 0.42, se(b) = 0.63, z = 0.66, p = 0.51). The trends seen in Figure 5 are not significant and, in any case, in the unexpected direction under Hypothesis 4. Thus, Hypothesis 4 is not supported. The present data are therefore problematic for attributing the bias against labial palatalization to categorization of the alternating phones into conjunctive natural classes.
Figure 7. Overgeneralization of palatalization depending on change magnitude during training

Hypothesis 5: The bias against labial palatalization is a bias in favor of perceptual similarity between alternants, hence [k] should be palatalized more than [g]

The relevant data are shown in Figures 8-9. Figure 8 shows palatalization rates for the individual stops. Velar stops are shown by the darkest bars (furthest right), with [g] on the left and [k] on the right. As the figure shows, there is a trend towards higher palatalization rate for [g] compared to [k]. This is not an overall voicing effect as it is absent in alveolars and labials. The difference between [g] and [k] is not in the expected direction according to the perceptual bias hypothesis. It is not clear how much is to be made of the trend, as it only reaches significance in the alveolar training condition \((b = 1.80, se(b) = 0.58, z = 3.09, p = .002)\). If we believe the trend, this may be an effect of categorization (Moreton & Pater 2012b), which in this case is presumably based on feedback from orthography: /g/ and /dӡ/ are categorized together because both are often written as <g>.

Figure 9 shows the comparable judgment data. There is no trend towards a difference between [g] and [k] with training on labial or alveolar palatalization but training on velar palatalization seems to elicit higher ratings of g~dӡi than k~tʃi \((b = .59, se(b) = .29, z = 2.08, p = .037)\). The direction of this trend is again counter to the predictions of perceptual bias.
**Figure 8.** Differences in palatalization rates across individual stops and training conditions. Shading indicates place of articulation from labial (lightest) through alveolar (medium) to velar (darkest). Voiced consonants are on the left within shading, while voiceless are on the right.

**Figure 9.** Differences in judgments of palatalization across individual stops and training conditions.

**Limitations of the data:** The role of first-language experience

The main limitation of the data is that all participants in the experiment had the same language background: they were all American English speakers. Using this population allows us to compare results to perceptual data from Guion (1998) and previous results on palatalization learning obtained by Wilson (2006) and Kapatsinski (2009, 2012, 2013). However, it also leaves the observed biases susceptible to an explanation based on first language phonological experience rather than articulatory differences between the examined phones. In particular, English has alveolar palatalization in certain contexts, particularly phrases like *would you* and *bet you*. While this kind of palatalization has been argued to retain the closure location of the input /d/ or /t/ (Zsiga 1995), unlike our stimuli, it may increase the productivity of alveolar palatalization for our participants. This may explain why alveolar palatalization was more productive than velar palatalization for the participants despite the fact that the two processes have similar typological frequencies (Bateman 2007, Kochetov 2011).

Palatalization of /g/ may also have been favored over palatalization of /k/ due to a first-language influence, namely the influence of English spelling-sound correspondence patterns. Based on Gontijo et
al. (2003), <g> often maps onto [dӡ] (~30% of <g>'s are [dӡ] and ~70% are [g]). In contrast, <k> always maps onto [k], <c> maps onto a palatal only 3% of the time, and <ch> maps onto [tʃ] 87% of the time. Thus spellings of [k] and [tʃ] are largely distinct, while spellings of [g] and [dӡ] are often the same, which may lead literate English speakers to be more likely to categorize the latter together in both perception and production, to perceive them to be more similar, and to assign them more similar neural codes. Alternations between them may then be easier to learn because of the mechanisms proposed by Moreton & Pater (2012a), Kapatsinski (2011) and Rescorla (1986). It would be interesting to see whether the difference between /k/ and /g/ observed here and in Wilson (2006) is nullified or reversed with preliterate children or infants (e.g. using the paradigm of White & Sundara 2014).

More problematically, velar palatalization may also be favored over labial palatalization in production because /k/ can change into [s] in English, as in electric-electricity, while [p] and [b] never change into anything. However, we do not consider it likely that this is responsible for the aversion to changing labials we observed. First, it is only /k/ and not /g/ that changes but we see no preference for changing /k/ over changing /g/. Second, the process is productive in only very restricted circumstances, namely when the input is Latinate-sounding and a specific Latinate suffix is attached. With stimuli like ours, English speakers are very reluctant to change /k/ into [s] even before change-triggering suffixes like –ity (Pierrehumbert 2006). Nonetheless, future work should examine acquisition of palatalization by participants whose native language has no productive palatalization process and no tendency for alternations to involve non-labials.

**Challenges for theories of PU and its functional motivations**

The results confirm the learning bias against palatalizing labials reported in Stave, Smolek & Kapatsinski (2013). The bias therefore holds both in contexts where palatalization in motivated (in that it improves phonotactics) and where it is unmotivated (between [a]'s), suggesting that it is a bias against large changes (see also Skoruppa et al. 2011). The bias is particularly strong in production compared to judgment. In production, labials are palatalized less than other stops. When exposed to labial palatalization, participants tend to palatalize nothing or (more rarely) palatalize everything. At the same time, participants tend to accept palatalization of everything or, more rarely, nothing. When trained to palatalize alveolars or velars, participants learn to palatalize the correct consonants more than the others. Despite the fact that palatalization of [k] and [t] before [i] can be argued to improve phonotactics, the bias against palatalizing labials does not appear to be due to high markedness of [ti] and [ki], as [pi] is no more acceptable to the participants.

We remain somewhat cautious in attributing the observed biases to the inherent articulatory differences between labials and linguals rather than first-language experience. Nonetheless, we believe that the present data do demand an articulatory, production locus for the bias wherever it comes from. Labial palatalization is seldom produced but is judged acceptable (see also Zuraw 2000 for a real-language example of this dissociation). In contrast, alveolar and velar palatalization are produced about as often as they are judged acceptable. Thus production and judgment preferences dissociate after labial training but not after training on alveolar or velar palatalization, suggesting that it is in production that the bias against changing labials resides.
The data present several problems for the P-Map account of paradigm uniformity (Steriade 2009). First, the English speakers in our study, like the Tagalog speakers in Zuraw (2000) are unaware of any speech norms prohibiting stem changes. In fact, in both cases speech norms (as judged by the participants) favor stem changes. Nonetheless, stem changes are avoided. The P-Map account traces avoidance of change to existence of perceptual preferences favoring faithful forms, from which the speaker prefers not to deviate. We see no such preference in the current study. In addition, as in Wilson (2006), [k] is palatalized no more than [g], despite [k] sounding more like an alveopalatal before [i] (Guion 1998), disconfirming a prediction of perceptual accounts for the bias (Hayes & White 2015, Kenstowicz 1996, Steriade 2009).

Another way to account for a bias against labial palatalization within the P-Map would be to attribute the bias to markedness. According to Steriade (2009), speakers produce changes in order to improve phonotactics (see also Prince & Smolensky 2004). Palatalization of [k] and [t] before [i] would therefore be attributed to the speakers’ dislike of [ti] and [ki]. Lack of palatalization of [p] would be attributed to the goodness of [pi]. However, we found no differences in acceptability between [pi], [ti] and [ki]. In a previous experiment (Stave, Smolek & Kapatsinski, 2013), we also found that the bias against labial palatalization continues to hold if the palatalizing vowel is [a] (Figure 2), i.e. in a context where all kinds of palatalization are phonotactically unmotivated. While participants are more reluctant to palatalize in an unnatural context (the bars are lower in Figure 2 than in Figure 3, see also Mitrović 2012, Wilson 2006), the difference between labial and lingual conditions in Figure 2 is virtually identical to that in Figure 3. The bias against labial palatalization thus appears to genuinely be a bias against certain changes, rather than a bias against or in favor of certain structures.

Kenstowicz’s (1996) suggestion that PU serves to improve lexical access fares better than the P-Map, assuming that it results in a bias against large stem changes within the production system. As noted in the introduction, such a bias is not necessarily implied by the lexical access motivation: a word that has not been accessed in perception will then be unlikely to be re-used in production by the listener-turned-speaker (Ohala 1981). Under this interpretation, we would not have expected a bias against large changes to occur in the present experiment.

However, we could also interpret Kenstowicz’s (1996) proposal as suggesting a bias against large changes on the part of the speaker. Assume that a change from [p] to [tʃ] is harder for a listener to undo than a change from [t] or [k] to [tʃ] (perhaps, because the neural representations of [p] and [tʃ] are less closely connected). Then the speaker could possibly avoid producing large changes because they are sensitive to possible lack of understanding that would result from such a change.

There are some data suggesting that speakers avoid ambiguity in production by hyperarticulating cues that distinguish words from their minimal pair neighbors (Baese-Berk & Goldrick 2009, Wedel et al. 2013). However, it seems rather unlikely that this kind of pressure could lead to change avoidance in the present experimental paradigm because of the (somewhat surprising) effect of providing the learner of a language featuring velar palatalization (k~tʃi) with additional examples of [tʃ]-final singulars mapping onto [tʃi]-final plurals. These examples make it ambiguous whether a [tʃi]-final plural corresponds to a [k]-final or [tʃ]-final singular. Thus, based on Kenstowicz (1996) we would expect them to make speakers avoid producing [tʃi]-final plurals. Nonetheless, Kapatsinski (2009, 2012, 2013) reports these examples to actually increase the likelihood of the speakers producing [tʃi]-final plurals, at least when -- as in the
present experiment -- stem-sharing singulars and plurals are not presented next to each other. In addition, if perceptual similarity between alternants mattered for learning the alternation, we would have expected a difference in palatalization rates between [k] and [g], with [k] palatalizing more than [g]. However, no significant difference was observed, suggesting that the effect of perceptual similarity between alternants is at least not as strong as the effect of articulatory similarity.

The categorization account (Moreton & Pater 2012a) faces major difficulties in accounting for both the greater degree of avoidance of palatalization in production and the lack of support for Hypothesis 4: if palatalization requires grouping alternating sounds together into a category, and the observed bias is a bias in favor of categories definable by a conjunction of features, then labial palatalization should be overgeneralized to apply to either coronals or velars. In contrast, velar palatalization does not have to be overgeneralized to coronals, and coronal palatalization does not have to be overgeneralized to velars. Our data show no evidence for this hypothesis. When participants are not trained to criterion (as in White 2013, 2014), i.e. trained until they overcome their biases against large changes, participants do not generalize large changes to intermediate sounds any more than they generalize other changes to the same sounds. Rather, the main difference between small and large changes is that participants tend not to produce the large changes.

**Conclusion**

We believe that the best explanation for the present data is provided by perseveration within the production system, coupled with a bias against associating representations of articulatorily dissimilar gestures. Perseveration disfavors all kinds of stem changes, while the proposed learning bias makes it hard to strengthen associations implementing large changes so that they become strong enough to override perseveration. Because of these biases, stem changes are disfavored in languages. However, learners of a language can learn to change specific segments in specific positions. When learners learn when and on what to perseverate in specific, semantically defined environments (such as when they want to express plurality), perseveration ceases to become merely a part of the motor system and becomes part of the morphophonological grammar, giving rise to paradigm uniformity effects.

---

15 When stem-sharing singulars and plurals are presented next to each other, presentation of these examples favor k→ki over the trained mapping k→tʃi, while still favoring the untrained t→tʃi over the trained t→ti (Kapatsinski 2009, Chapter 3; see also Yin & White 2016). The disfavoring of k→tʃi may be taken as a bias against homophony (Yin & White 2016), although it then needs to be explained why this bias is present only when stem-sharing singulars and plurals are presented next to each other in training and why it does not apply to t→tʃi even under those circumstances. However, in Kapatsinski (Submitted), we show that the model outlined in Figure 1 accounts for the full pattern of results: examples of tʃ→tʃ favor CopyFin but they also favor mapping PLURAL onto [tʃi] and disfavor mapping it onto [ti] (by reducing its probability given the plural meaning). On balance, tʃ→tʃ examples benefit t→tʃ more than t→ti. Following Rescorla & Wagner (1972), we assume that participants do not learn anything about associations of cues or outcomes that are always absent. Given this assumption, examples of tʃ→tʃ say nothing about the probability of mapping PLURAL onto [ki] because [ki] is never presented during training with or without the additional examples. Note that palatalization of both [k] and [t] is favored by these examples if CopyFin is not detected by the learner, though [k] benefits less. For this reason, palatalization of both consonants is favored when stem-sharing singulars and plurals are not presented next to each other in training as in Kapatsinski (2009, Chapter 4; 2012) and the present experiment.

16 Voicing differences have often been reported to have a stronger effect on perceptual confusability than place differences for English listeners (e.g. Miller & Nicely 1955).
Acknowledgments

Parts of this work were presented at the 2013 meetings of the Linguistic Society of America, and the Cognitive Science Society, the 2013 American International Morphology Meeting, the 2013 Quantitative Approaches to Theoretical Linguistics workshop and at departmental colloquia at the University of Oregon and UCSD. We thank audience members at these events for their helpful feedback. In particular, we thank Adam Albright for suggesting we examine the influence of voicing on palatalization. We would also like to thank Matthew Stave for recording the stimuli for the present experiment, and for spearheading the pilot experiment that served as a fore-runner to the present study (Stave et al., 2013).

References:


Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*(2), 204-238.


Robins, R. H. (1959). In defence of WP. *Transactions of the Philological Society, 58*(1), 116-144.


Appendix
Training Stimuli:
{Labial;Alveolar;Velar}

<table>
<thead>
<tr>
<th>0→i/X</th>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>0→i/X</td>
<td>la(t;p;t) x2</td>
<td>la(t;p;t)i x2</td>
</tr>
<tr>
<td></td>
<td>pre(t;p;t)</td>
<td>pre(t;p;t)i</td>
</tr>
<tr>
<td></td>
<td>fer(d;b;d) x2</td>
<td>fer(d;b;d)i x2</td>
</tr>
<tr>
<td></td>
<td>he(d;b;d)</td>
<td>he(d;b;d)i</td>
</tr>
<tr>
<td></td>
<td>noʊ(k;k;p) x2</td>
<td>noʊ(k;k;p)i x2</td>
</tr>
<tr>
<td></td>
<td>spaɪ(k;k;p)</td>
<td>spaɪ(k;k;p)i</td>
</tr>
<tr>
<td></td>
<td>glu(g;g;b) x2</td>
<td>glu(g;g;b)i x2</td>
</tr>
<tr>
<td></td>
<td>roʊ(g;g;b)</td>
<td>roʊ(g;g;b)i</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>0→a/X</th>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>0→a/X</td>
<td>dae(t;p;t) x2</td>
<td>dae(t;p;t)a x2</td>
</tr>
<tr>
<td></td>
<td>flaɪ(t;p;t)</td>
<td>flaɪ(t;p;t)a</td>
</tr>
<tr>
<td></td>
<td>bri(d;b;d) x2</td>
<td>bri(d;b;d)a x2</td>
</tr>
<tr>
<td></td>
<td>tʃai(d;b;d)</td>
<td>tʃai(d;b;d)a</td>
</tr>
<tr>
<td></td>
<td>faɪ(k;k;p) x2</td>
<td>faɪ(k;k;p)a x2</td>
</tr>
<tr>
<td></td>
<td>tʃi(k;k;p)</td>
<td>tʃi(k;k;p)a</td>
</tr>
<tr>
<td></td>
<td>gwi(g;g;b) x2</td>
<td>gwi(g;g;b)a x2</td>
</tr>
<tr>
<td></td>
<td>her(g;g;b)</td>
<td>her(g;g;b)a</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X</th>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>broʊ(p;t;k) x4</td>
<td>broʊtʃi x4</td>
</tr>
<tr>
<td></td>
<td>slæ(p;t;k) x4</td>
<td>slætʃi x4</td>
</tr>
<tr>
<td></td>
<td>her(p;t;k)</td>
<td>herʃi</td>
</tr>
<tr>
<td></td>
<td>naz(p;t;k)</td>
<td>natʃi</td>
</tr>
<tr>
<td></td>
<td>gla(p;t;k)</td>
<td>glatʃi</td>
</tr>
<tr>
<td></td>
<td>sni(p;t;k)</td>
<td>sniʃi</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X</th>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>he(b;d;g) x4</td>
<td>heðʒi x4</td>
</tr>
<tr>
<td></td>
<td>paɪ(b;d;g) x4</td>
<td>paɪdʒi x4</td>
</tr>
<tr>
<td></td>
<td>va(b;d;g)</td>
<td>vadʒi</td>
</tr>
<tr>
<td></td>
<td>gi(b;d;g)</td>
<td>gidʒi</td>
</tr>
<tr>
<td></td>
<td>bri(b;d;g)</td>
<td>bridʒi</td>
</tr>
<tr>
<td></td>
<td>tʃai(b;d;g)</td>
<td>tʃaiðʒi</td>
</tr>
</tbody>
</table>