Velar palatalization in Russian and artificial grammar: Constraints on models of morphophonology*

VSEVOLOD KAPATSINSKI

University of Oregon

Abstract

Russian velar palatalization changes velars into alveopalatals before certain suffixes, including the stem extension -i and the diminutive suffixes -ok and -ek/ik. While velar palatalization always applies before the relevant suffixes in the established lexicon, it often fails with nonce loanwords before -i and -ik but not before -ok or -ek. This is shown to be predicted by the Minimal Generalization Learner (MGL), a model of rule induction and weighting developed by Albright and Hayes (2003), by a novel version of Network Theory (Bybee 1985, 2001), which uses competing unconditional product-oriented schemas weighted by type frequency and paradigm uniformity constraints, and by stochastic Optimality Theory with language-specific constraints learned using the Gradual Learning Algorithm (GLA, Boersma 1997). The successful models are shown to predict that a morphophonological rule will fail if the triggering suffix comes to attach to inputs that are not eligible to undergo the rule. This prediction is confirmed in an artificial grammar learning experiment. Under either model, the choice between generalizations or output forms is shown to be stochastic, which requires retrieving known wordforms from the lexicon as wholes, rather than generating them through the grammar. Furthermore, MGL and GLA are shown to succeed only if the suffix and the stem shape are chosen simultaneously, as opposed to the suffix being chosen first and then triggering (or failing to trigger) a stem change. In addition, the GLA is shown to require output-output faithfulness to be ranked above markedness at the beginning of learning (Hayes 2004) to account for the present data.

1. Introduction

A common phenomenon in language change is the loss of productivity by morphophonemic alternations. The phenomenon is so common that Bybee (2008: 114) considers it a diachronic universal, yet it is quite puzzling. Why would an alternation start accumulating exceptions and stop being extended to new words entering the language despite starting out with no exceptions and an abundance of examples

JLP 1 (2010), 361–393
DOI 10.1515/LABPHON.2010.019

1836-6346/10/01–361

supporting it in the lexicon? A particularly interesting historical development happens when an alternation has no exceptions in the lexicon but is not extended fully to new words entering the language. Thus, the alternation loses productivity while not gaining new exceptions.[1] In the present paper I will show that a particular alternation, velar palatalization (k → tʃ, g → ʒ), has lost productivity before some Russian suffixes but not others. Velar palatalization has no exceptions in the Russian lexicon, as depicted by dictionaries, before the verbal stem extension -i and the diminutive suffixes -ok and -ek/ik. However, with recent loanwords found in the discourse of Russian-speaking Internet users, velar palatalization fails about 50% of the time before -i and -ik while remaining fully productive before -ek and -ok. Thus Russian provides an opportunity to examine the factors influencing productivity by examining what makes velar palatalization productive before -ek or -ok but not before -i and -ik.

The Minimal Generalization Learner (MGL, Albright and Hayes 2003) is a computational model that induces rules from a lexicon and weights them relative to each other. The MGL proposes that the productivity of a linguistic rule is determined by its estimated reliability relative to other rules that can apply to the same input forms.[2] For instance, suppose we have a language where the velar consonant [k] at the end of a singular stem changes into the alveopalatal [tʃ] when followed by a plural suffix [i]. This singular-plural mapping can be stated as the rule k# → tʃi#. The reliability of a rule is defined as the number of words to which the rule applies divided by the number of words to which it could apply. In the case of k# → tʃi#, this is the number of singular-plural pairs in which a final [k] in the singular corresponds to [tʃi] in the plural divided by the number of singulars that end in [k]. If all singulars that end in [k] correspond to plurals that end in [tʃi], the rule is completely reliable. The more reliable a rule is, the more productive it is expected to be. The present paper shows that reliability-driven competition between rules predicts that a morphophonological rule will lose productivity if the triggering suffix comes to attach mostly to inputs that cannot undergo the rule. This prediction is confirmed both by Russian loanword adaptation data, where -i and -ik are the suffixes that tend not to attach to velar-final inputs whereas -ok and -ek are favored by velar-final inputs, and by data from an experiment with miniature artificial languages. In both loanword adaptation and experimental data, the degree of productivity of a palatalizing rule like k → tʃ/_i is explained by how often the triggering suffix (here, -i) attaches to consonants that can undergo palatalization relative to how often it attaches to consonants that cannot undergo palatalization.

The role of the artificial language experiment is to test the direction of causation. The Russian data establish a correlation between how often a suffix attaches to non-velars and how productively it triggers velar palatalization. However, correlation is not causation. Thus, one possibility is that low productivity of velar palatalization before a suffix, whatever its cause, causes speakers of Russian avoid using -i and -ik with velar-final inputs (e.g., Berg 1998: 230–233; Martin 2007; Thomason 1976). Another possibility is that some suffixes are bad triggers of velar palataliza-

tion because they tend not to be attached to velars, which is predicted by MGL (Albright and Hayes 2003) and Optimality Theory with the Gradual Learning Algorithm (Boersma 1997). The advantage of an artificial language is that one is able to manipulate all characteristics of its lexicon. In the present experiment, we hold productivity of velar palatalization in the training data constant (at 100%) while manipulating the rate of attaching -i to non-velars, and observe that the learnability of velar palatalization (i.e., its productivity at test) varies as a function of the manipulated variable (the rate of attaching -i to non-velars). Thus, we can say that the hypothesized direction of causation is supported.

The MGL is used as a representative example of models that embody the hypothesis of statistically resolved competition between source-oriented generalizations. Source-oriented generalizations specify a mapping between a specific category of inputs and a specific category of outputs, e.g., k → tʃi, rather than simply describing what the outputs should be like (Bybee 2001: 126–129). This assumption is also made in Analogical Modeling of Language (Skousen 1989), which relies on rule type frequency instead of reliability,[3] the account of productivity of English velar softening developed in Pierrehumbert (2006), and feedforward connectionist networks learning mappings between forms belonging to the same paradigm (e.g., Rumelhart and McClelland 1986). Product-oriented generalizations specify only the shape of the output, e.g., 'plurals end in /tʃi/', and are derived via generalization over outputs. Alternative accounts for the data using a combination of product-oriented and source-oriented generalizations based on Network Theory (Bybee 1985, 2001) and stochastic Optimality Theory / Harmonic Grammar learned using the Gradual Learning Algorithm (Boersma 1997) are presented in Section 4. Unlike the MGL, the extant product-oriented approaches do not provide an algorithm for extracting generalizations from the lexicon, specifying only how the already extracted generalizations are to be weighted. Thus, the MGL model is presented first. Nonetheless, the data place similar restrictions on all models of grammar, whether rule-based or constraint-based.

2. Loanword adaptation in Russian

2.1. *The pattern in the lexicon*

If one looks at a dictionary of modern Russian, velar palatalization appears to involve several exceptionless morphophonological rules, which can be stated simply as "velars become alveopalatals before the derivational suffixes X" where the relevant derivational suffixes either begin with a front vowel or used to begin with a front vowel historically. For the purposes of this article, we will be concentrating on Russian verbs with the highly productive stem extension -i, and on the diminutive suffixes for masculine nouns, -ik/ek and -ok, which obligatorily trigger velar palatalization in the lexicon, as depicted by dictionaries (e.g., Levikova 2003; Sheveleva 1974).

Example (1) shows that Russian verbs are derived from consonant-final nouns by adding the stem extension (in this case /i/) followed by verbal inflection (e.g., the infinitival marker t^j). As shown in (1)–(3), velars at the ends of noun roots change into alveopalatals when a verb is derived from the root. This does not happen with all stem extensions, or even all [i]-initial extensions, as evidenced by the existence of Russian verbs like n^jux-a-t^j, plak-a-t^j, and stalk-iva-t^j, but it always happens with the stem extension -i.

(1) k → tʃi[4]

blok	blotʃ-i-t^j
Root	Root-EXT-INF
'block'	'to put together (into a block)'
durak	duratʃ-i-t^j
Root	Root-EXT-INF
'fool'	'to fool'
polk	o-poltʃ-i-t^j-s^{j}a
Root	PFV-Root-EXT-INF-REFL
'regiment'	'take up arms'
jamʃtʃik	jamʃtʃitʃ-i-t^j
Root[5]	Root-EXT-INF
'coachman'	'to work as a coachman'

(2) g → ʒɨ

drug	druʒ-i-t^j
Root	Root-EXT-INF
'friend'	'to be friends'
dolg	o-dolʒ-i-t^j
Root	PFV-Root-EXT-INF
'debt'	'to lend / borrow'

(3) x → ʃɨ

grex	greʃ-i-t^j
Root	Root-EXT-INF
'sin'	'to sin'

The mappings between velar consonants and the corresponding alveopalatals are constant across Russian. Thus, if velars change into alveopalatals in some context, /k/ always becomes [tʃ], /g/ becomes [ʒ], and /x/ becomes [ʃ]. The Russian phone inventory does not contain [dʒ]. The phone [ɨ] cannot follow velars or [tʃ] while the phone [i] cannot follow [ʃ] or [ʒ]. Whether [i] and [ɨ] are allophones of /i/ and chosen during a separate allophone selection stage or separate stem extensions does not influence the qualitative results presented here. The reported graphs are

based on models that treat the choice between [i] and [ɨ] as happening after the morphophonological competition modeled.

In the Russian lexicon, as depicted by dictionaries, -a is favored over -i by velar-final roots while -i is favored elsewhere. The distribution in the diminutive system is quite different. Only masculine diminutive suffixes will be considered for the purposes of this paper, because the loaned English nouns in question end in a consonant, consequently being adopted into the masculine gender. There are three highly productive masculine diminutive suffix morphs, -ik, -ek, and -ok. The morphs -ek and -ik are in nearly complementary distribution in the established lexicon and thus can be considered allomorphs of a single morpheme. The suffixes that trigger palatalization in the lexicon, -ok and -ek, are heavily favored by velar-final nouns, with -ek attaching only to velar-final bases. The suffix -ik, on the other hand, almost never attaches to velar-final bases, thus one could argue that the Russian lexicon provides almost no evidence on whether -ik would trigger or fail to trigger velar-palatalization if it were to be attached to a velar-final base, although I will argue that the lexicon does in fact provide the relevant information and Russian speakers use this information in loanword adaptation.[6] Examples are shown in (4)–(8) (all from Sheveleva 1974).

(4) The diminutive -ok, illustrating the changes undergone by /g/ and /k/

lug	→	luˈʒ-ok
'meadow'		'little / nice meadow'
luk	→	luˈʧ-ok
'onion'		'little / nice onion'

(5) The diminutive -ek

ˈfartuk	→	ˈfartuʧ-ek
'apron'		'little / nice apron'
ʧelоˈvek	→	ʧelоˈveʧ-ek
'person'		'little / nice person'

(6) The diminutive -ik can (rarely) attach to velars. This is the only unambiguous example in Sheveleva (1974).

bank	→	ˈbanʧ-ik
'bank (financial)'		'little / nice bank'

(7) -ok and -ik in the same non-velar context (no semantic difference as far as I can tell)

ʧas	→	ʧaˈs-ok
'hour'		'nice hour'
ʧas	→	ˈʧas-ik
'hour'		'nice hour'

(8) -ek can attach to already diminutive forms, which end in [k]

noʒ	→	ˈnoʒ-ik	→	ˈnoʒ-itʃ-ek
'knife'		'little knife'		'nice little knife'
ʃag	→	ʃaˈʒ-ok	→	ʃaˈʒ-otʃ-ek
'step'		'nice / little step'		'small / nice little step'

2.2. *Methods*

2.2.1. *Data collection.* When an English verb is borrowed into Russian, it must be assigned a stem extension. In order to get a sample of such borrowings, I took all verbs found in the British National Corpus retrieved by searching for "*x.[vvi]" in the online interface provided by Mark Davies (http://corpus.byu.edu/bnc/) where 'x' is any letter. The resulting verbs were transliterated into Cyrillic.

For each verb, possible Russian infinitival forms were derived. For instance, if the English verb is *lock*, some possible Russian infinitives are /lotʃ-i-tʲ/, /lok-i-tʲ/, /lok-a-tʲ/, /lok-ova-tʲ/ and /lok-irova-tʲ/. Verbs for which an established Russian form already existed (e.g., format > /formatirovatʲ/) were excluded. Existence was determined by the occurrence in either the Reverse Dictionary of Russian (Shveleva 1974), Big Dictionary of Youth Slang (Levikova 2003), or the present author's memory. This yielded 472 different verbs. For 56 of them, the final consonant of the English form was a velar, for 99 it was a labial, and for 317 it was a coronal. In the case of the nouns, all existing English nouns ending in /k/ or /g/ and found in the British National Corpus were created and transliterated into Russian manually. Then possible diminutive forms were created from them and submitted to Google. An additional sample of non-velar-final nouns was then created by matching the distribution of final consonant types in terms of manner and voicing and preceding vowels in the sample of velar-final nouns.

The frequencies of the possible infinitives and nominative diminutives on the web were determined by clicking through the pages of results returned by Google to eliminate identical tokens and to allow Google to 'eliminate similar pages', which increases speaker diversity by eliminating results that come from the same server, e.g., different pages from the same bulletin board. In addition, clicking through is necessary when one of the possible forms has a homonym.

Finally, to have a reasonably reliable estimate of the likelihood of failure of velar palatalization before -i for each verb, velar-final verbs and nouns that had 10 or fewer tokens containing the palatalizing suffixes were excluded from the sample. This yielded 36 velar-final verbs and 19 velar-final nouns that could undergo velar palatalization and had a reasonably large number of tokens containing the relevant suffixes.

2.2.2. *Modeling* The Minimal Generalization Learner (Albright and Hayes 2003) is a computational model of rule induction and weighting. The model starts with a set of morphologically related word pairs as in (9).

(9) mot motat^{j}
tʃmok tʃmokat^{j}
drug druʒit^{j}
krug kruʒit^{j}
golos golosovat^{j}

For each word pair, the model creates a word-specific rule as in (10).

(10) Ø → a / mot__
Ø → a / tʃmok__
g → ʒi / dru__
g → ʒi / kru__
Ø → ova / golos__

Then, rules that involve the same change are combined. Contexts in which the same change, e.g., Ø → i, happens are compared by matching segments starting from the location of the change. If segments match, they are retained in the specification of the context for the change and the pair of segments further away from the change is compared. When this comparison process reaches the nearest pair of segments that do not match, the phonological features they share are extracted and retained in the specification of the context. Segments that are further away from the location of the change than the closest pair of non-matching segments are not compared and are replaced by a free variable in the specification of context.

For instance, the rules in (11) are combined into the rule in (12).

(11) g → ʒi / dru__
g → ʒi / kru__

(12) g → ʒi / [+cons;+cont;-son;-Labial]ru__

The resulting more general rules are then compared to each other and even more general rules derived if the same change occurs in multiple contexts, eventually resulting in quite general rules, such as Ø → i / C__. However, all rules are retained in the grammar. Instead of removing non-maximally-general rules from the grammar, the MGL weights each rule by its reliability. Reliability is defined as the number of words to which the rule applies divided by the total number of words to which it *could* apply. For instance, the reliability of the rule in (12) is the number of words of the form in (13) that are derived from words with the shape in (14) divided by the total number of words with the shape in (14) in the lexicon.

(13) [-cons;-cont;-son;-Labial]ruʒi
(14) [-cons;-cont;-son;-Labial]rug

A reliable rule is more likely to apply to a novel word than a less reliable rule. For instance, if the rule in (15) is more reliable than the rule in (16), and these are the only rules that can apply to the novel verb /dig/, the verb should be more likely to be borrowed as /diʒi/ than as /diga/.

(15) g → ʒi / V__
(16) Ø → a / Vg__

The model treats reliability as a value that is conservatively estimated based on the sample of experienced words. It requires a choice of confidence level, which determines the width of the confidence interval around the raw reliability value derived from the lexicon. It then takes the lower bound of the confidence interval as the estimate of reliability. This makes rules that can apply to many words (rules with high scope) more reliable than rules that apply to few words (low scope), other things being equal. The weight of scope relative to raw reliability is determined by the width of the confidence interval, which in turn is determined by how confident we want to be that the true reliability of the rule is equal to or greater than the estimated value. For the present data, the three default values (75%, 90%, 95%) produced qualitatively indistinguishable results. The quantitative results reported here are for the 75% confidence interval (which is consistent with Albright and Hayes 2003).

The set of rules extracted from the lexicon, i.e., the grammar, is used only on novel words entering the lexicon. Existing morphologically complex words are stored in memory and retrieved from the lexicon as wholes rather than being generated by the rules of the grammar (Bybee 1985, 2001; Halle 1973; and Vennemann 1974, inter alia). Storage and retrieval of morphologically complex words is essential for a rule to be able to lose productivity while not gaining exceptions. If existing words were generated by the grammar, they would not continue to obey a rule as the rule loses productivity.

For the purposes of the present paper, this model has four essential features: 1) the model generalizes over input-output mappings, as opposed to just outputs (Bybee 2001: 126–129; Pierrehumbert 2006), 2) input-output mappings compete for inputs, 3) the outcome of this competition is driven by reliability, and 4) morphologically complex words are retrieved from the lexicon in production.

The model of the stem extension process was presented with the set of stem-verb pairings found in the Reverse Dictionary of Russian (Sheveleva 1974) and/or the Big Dictionary of Youth Slang (Levikova 2003). The Reverse Dictionary contains 125,000 words extracted from the four major dictionaries of Russian that existed in 1965 (Sheveleva 1974: 7). The Slang Dictionary is much smaller, containing 10000 words. The main results presented below held regardless of whether the Reverse Dictionary, the Slang Dictionary, or both were used. Only the results based on the full training set will be presented. Only stems that occurred independently as separate words were included. No stem extensions were excluded from the training set. Thus, aside from verbs featuring the highly productive -i and -a, verbs having -ova, -irova, -izirova, and -e were also included. The full training set consisted of 2396 verb-stem pairs, of which 286 stems had final /k/ and 85 had final /g/. There were 22 examples of g → ʒi and 62 examples of k → ʧi. The model of diminutive formation was trained on a set of 1154 diminutive nouns ex-

tracted from the Reverse Dictionary of Russian. All diminutive nouns whose base ends in a consonant were extracted regardless of the diminutive suffix used. The Slang Dictionary contains only a very small number of diminutives and thus was not used. Neither dictionary contains any examples of failure of velar palatalization before any of the crucial suffixes. Thus velar palatalization is exceptionless in the training set.

The learner models competition between input-output mappings. Therefore it is crucial to define what is meant by the input and the output. For the present paper, we are interested in modeling competition between input-output mappings in which some mappings require velar palatalization. The input form for these mappings may or may not have the stem extension already specified. If it does, rules specifying that a velar changes into an alveopalatal compete with rules that say that the consonant stays the same in the context of a stem extension that triggers velar palatalization in the lexicon (e.g., k → tʃ / __i vs. k → k / __i). If not, rules specifying that a velar changes into an alveopalatal also specify the stem extension. Thus a rule like k → tʃi would compete with k → ka as well as C → Ci.

In addition, the output of the competition can be either a phonetic form, specifying the allophone of /i/ used, or a phonemic form, which does not include this specification. Both of these possibilities were examined in modeling but the choice between phonetic and phonemic outputs did not influence the qualitative results. In the case of the diminutive suffixes -ek and -ik, which can be considered allomorphs (or even orthographic variants), it also did not make a difference whether the choice between -ek and -ik followed the stage in which the decision on whether to palatalize the stem was made.

The model is presented with the set of English verbs found to be borrowed into Russian in the corpus study. To estimate the probability of a given verb undergoing velar palatalization given that a particular suffix is chosen we can divide the reliability of the most reliable rule that requires palatalization by the sum of its reliability and the reliability of the most reliable rule that does not require palatalization but still attaches the same suffix. For instance, suppose the verb is /dig/ and the model has extracted the rules in (17) with reliability estimates shown in parentheses. The only rules that can apply to /dig/ are (a), (d), (e), (h), (i), and (j). Of these, the only rules that require velar palatalization are rules h and i. Rule h is more reliable than rule i, so it would get to apply. Its reliability is 0.272. The rule that attaches -i without palatalizing the stem-final /g/ is rule j. Its reliability is 0.232. Therefore, the predicted probability that the final consonant of /dig/ will be palatalized, given that -i is selected as the stem extension, is 0.272 / (0.272 + 0.232) = 54% (cf. Albright and Hayes 2003: 128).

(17) a. Ø → a / {i;l}g__ (.723)
b. Ø → a / Cag__ (.718)
c. Ø → a / {l;r}eg__ (.718)
d. Ø → a / {i;l;n;r}g__ (.670)

e. Ø → a / [velar]__ (.641)
f. g → ʒi / $V_{[-back;+high]}$__ (.475)
g. g → ʒi / $V_{[-high]}$__ (.350)
h. g → ʒi / V__ (.272)
i. g → ʒi / [+voice]__ (.195)
j. Ø → i / $C_{[+voiced]}$__ (.232)

It is also possible to derive predicted rate of application by summing up reliabilities of all rules requiring a given change and dividing that value by the sum of reliabilities of all rules requiring -i. Overall, assessing the probability of palatalization in this alternative way favors palatalization over non-palatalization because there is only one rule that can produce a non-palatalized velar for each velar-final input. Consider again the rules in (17). Then to determine the probability of changing the final /g/ in, let's say /dig/, we would sum rule reliabilities in the sequence of rules that can apply to /dig/ and palalize the /g/, which would be g → ʒi / V__ and g → ʒi / [+voice]__ (0.272 + 0.195 = 0.467). Dividing the result by the total probability of adding -i (0.467 + 0.232) we receive a palatalization rate of 67%. Aside from a somewhat higher predicted rate of velar palatalization under this alternative method of calculating reliability, the two methods make the same predictions for the data in the present study.[7]

2.3. *Results*

Figure 1 shows that most velar-final verbs are highly unlikely to take -i while most labial-final and coronal-final verbs are very likely to take -i. Thus, the stem extension that triggers a stem change in the lexicon is disfavored by the stems that can undergo the change.

Since the population distribution is skewed and bimodal, there is no monotonic transformation that will restore normality, which makes standard statistical tests

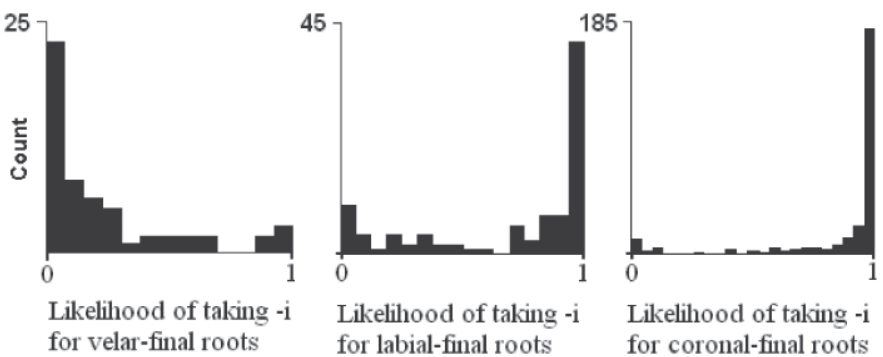


Figure 1. *Histograms of loanword adaptation data from Google showing that most velar-final stems are unlikely to take -i while most labial-final and coronal-final stems tend to take -i.*

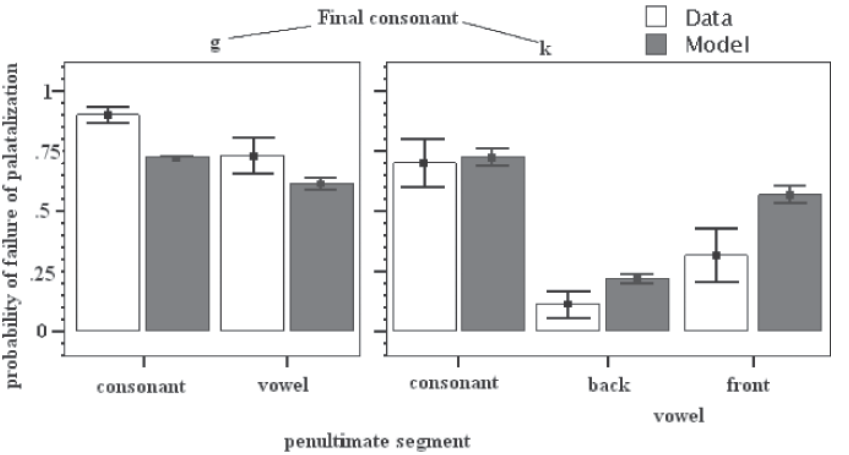


Figure 2. *Observed (white bars) vs. predicted (grey bars) probabilities of failure of velar palatalization in loanword adaptation before the stem extension -i depending on segmental content of the stem.*

inapplicable, which means that bootstrapping should be done. For this test, I treated the labial-final roots and coronal-final roots as the null population and generated 2000 samples of 56 verbs from this population, calculating mean rate of taking -i in each sample. The mean rate of taking -i in the sample of velar-final stems (33%) falls very far outside the distribution of 2000 samples of 56 verbs from the null population, thus the rate of taking -i for velars is significantly different from the rate of taking -i for labials and coronals at $p < 0.0005$ (1/2000). All versions of the model are able to predict that -i is less productive with velar-final stems than with coronal-final and labial-final stems.

Figure 2 shows just the velar-final stems that take -i as the stem extension. These are the only stems that undergo velar palatalization in the data, suggesting that the speakers are using a source-oriented generalization mapping velars onto alveopalatals, rather than a purely product-oriented generalization requiring alveopalatals before -i (Pierrehumbert 2006). A product-oriented generalization specifies only the shape of the output, thus imposing no restrictions on what changes can be done to the input to produce the output (for examples of such product-oriented behavior, see Bybee 2001: 126–129).

The white bars show the observed likelihood of failure of velar palatalization before -i in various contexts while the dark bars show probabilities of velar palatalization failure predicted by the model. Figure 2 shows that velar palatalization is more likely to fail with /g/ than with /k/ ($p = 0.009$ according to the Wilcoxon test). For, stems ending in [k], the rate of velar palatalization is higher if the stem ends in Ck rather than Vk ($p = .005$ according to the Wilcoxon test). Notably, in these data, the proportion of CC sequences relative to VC sequences is higher

when the final consonant of the input is [k] (1/2) than when it is [g] (2/9), thus the effect of final consonant is not due to the effect of the penultimate consonant. There is also a trend for the rule to fail more often after front vowels than after back vowels but it is not statistically significant ($p = 0.12$ according to the Wilcoxon test). In other words, speakers tend to retain the velar if it is /g/ and if it is preceded by a consonant. They tend to replace the velar with an alveopalatal if it is a /k/ preceded by a vowel, especially if the vowel is back.

Despite the fact that the model is trained on a lexicon in which velar palatalization is exceptionless, the model predicts that velar palatalization will not be exceptionless with the borrowed verbs. Mean rate of failure of velar palatalization varies between 43% and 62% depending on parameter settings and approximates the actual mean rate of failure of velar palatalization in the data (56%).

While the mean predicted rate of failure for velar palatalization is similar to the observed rate of failure, the model's predictions are less variable than the data. In order to make them comparable, failure rates predicted by the model were rescaled so that $p_{rescaled} = 1/(1 - \exp(\ln((1 - p)/p)/\text{temperature}))$ where temperature was set to 0.45.[8] The qualitative results shown in Figure 2 hold for all versions of the model. These versions of the model correctly predict that velar palatalization is more likely to fail when the stem ends in a consonant cluster than when it ends in a single consonant, that penultimate front vowels disfavor palatalization compared to back vowels, and that /k/ is more likely to be palatalized than /g/ (however, all versions of the model underestimate the difference between /k/ and /g/).[9]

Figure 3 shows the predicted and observed velar palatalization probabilities before -i for all verbs in the study. The model is able to account for 41% of the variance among verbs. The correlation between observed and expected probabilities of velar palatalization is highly significant ($t(38) = 3.38$, $p = 0.002$).[10] The words on the lower right of the graph, with which the model has the most trouble, are verbs

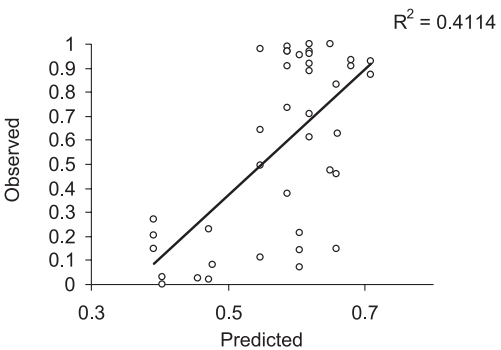


Figure 3. *Observed (in loanword adaptation) vs. predicted probabilities of velar palatalization before the stem extension -i. The predicted probabilities are raw (unscaled).*

ending in [ik] (e.g., 'tweak', 'freak', 'speak') and one verb ending in [inʲk] ('drink'), which undergo palatalization less often than predicted.

Observed and predicted rates of failure of velar palatalization in front of diminutive suffixes are shown in Figure 4. As with the stem extensions, velars are the only consonants that change into alveopalatals, suggesting a source-oriented generalization. The rate of failure of velar palatalization is significantly higher before the suffix -ik (mean rate of failure = 40%) than before the suffix -ok (mean rate of failure = 0%), according to the paired-samples Wilcoxon signed ranks test ($Z(16) = 3.516$, $p < 0.0005$). Failure of palatalization (which only happens before -ik) is more likely with /g/ (67%) than with /k/ (29%), $p < 0.05$ according to the Wilcoxon test). The likelihood of using -ik is lower after /k/ than after /g/ ($p < 0.005$ according to the Wilcoxon) and is higher after non-velars than after velars ($p < .005$ according to the Wilcoxon). Thus, the suffixes -i and -ik tend to attach to non-velar-final inputs and often fail to trigger velar palatalization. The suffixes -ek and -ok tend to attach to velar-final inputs and are strong triggers of velar palatalization. Furthermore, in both the domain of verbal stem extensions and nominal diminutives, the productivity of k → tʃ is greater than the productivity of g → ʒ.

The model successfully learns that -ik is disfavored by velars and that palatalization is likely to fail only if -ik is chosen as the suffix, although the rate of failure of velar palatalization before -ik is overestimated. It predicts that -ek should be more productive with bases ending in /k/ than with bases ending in /g/, a numerical trend in the data. It fails to predict that /k/ is more likely to undergo palatalization and less likely to be followed by -ik than /g/. These predictions are parameter-independent, holding for all versions of the model.

2.4. *Explaining successes and failures of the model*

In the present study, the MGL is used as only an example of a general class of models that postulate that input-output mappings are involved in a competition that is resolved by the mappings' relative reliability. Therefore it is important to determine the extent to which the successes and failures of the MGL are due to its reliance on this assumption.

In order to explain why the model performs the way it does let us examine the rules that it abstracts from the lexicon and uses when a velar-final verb is presented. The full list of applicable rules for [g]-final verbs is shown in (11) above. For both [k]-final and [g]-final verbs, there is only one rule that favors adding -i and leaving the final consonant of the stem unchanged. For /g/-final roots, this is the rule $C_{[+voiced]} \rightarrow C_{[+voiced]}i$ and for /k/-final roots this is the rule $C \rightarrow Ci$. Thus, in order for the more specific rules requiring /k/ to change into /tʃ/ or /g/ to change into /ʒ/ to fail, they must lose to an extremely general rule. For this outcome to be likely, 1) a very general rule must be extracted from the lexicon, 2) it should be quite reliable relative to the less general rules requiring stem changes, and 3) it must compete with those rules.

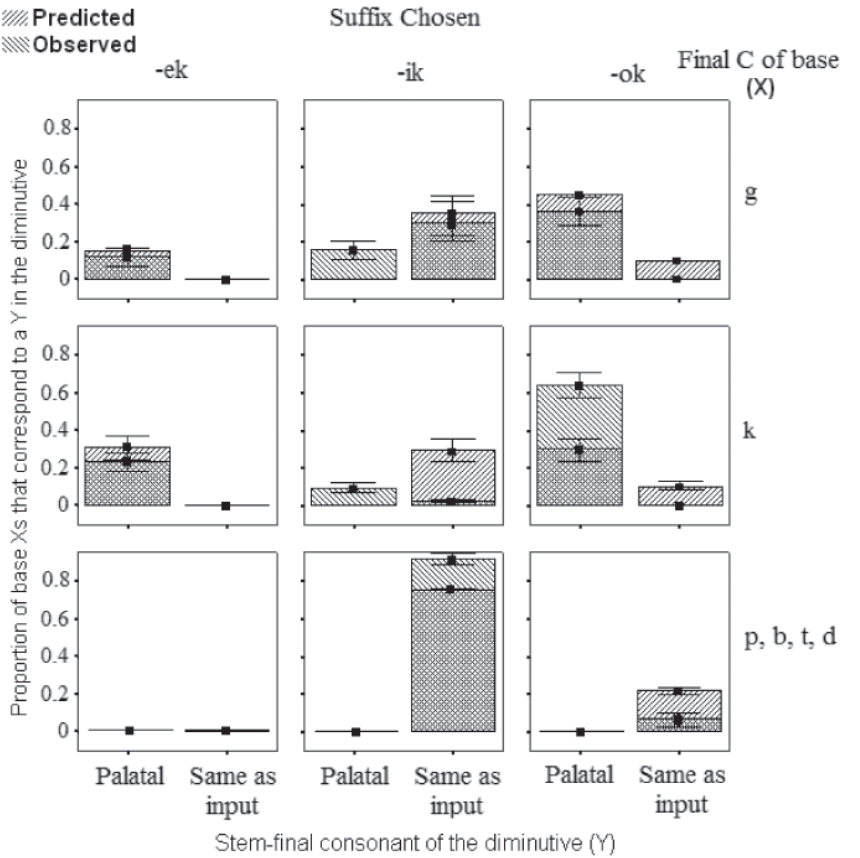


Figure 4. *Relative observed (in loanword adaptation on the web) and predicted likelihoods of various base-diminutive mappings for velar-final and non-velar-final bases of diminutive nouns (likelihoods sum to one across each row of panels).*

In the Russian lexicon used to train the model, coronal-final and labial-final stems tend to take -i while velar-final stems tend to take -a. Since most stems in the lexicon end up taking -i, the model extracts a very general rule C → Ci and assigns it a moderate reliability. On the other hand, the fact that velar-final stems favor -a drives down the reliabilities of rules that add other stem extensions to velar-final stems. This includes the rules that add -i and change the root-final consonant. As a result, these rules will sometimes lose the competition for application to the more general rule C → Ci. Thus, the model predicts that velar palatalization will often fail before an affix if and only if the affix is more productive after non-velars than after velars. This holds for the stem extension -i and the diminutive suffix -ik but not for the diminutive suffixes -ek and -ok. Therefore, the model correctly predicts that velar palatalization should fail often before -i and -ik and rarely before -ek and

-ok. This prediction follows directly from the hypothesis that input-output mappings compete, with the outcome determined by reliability.

The model systematically fails to capture the difference in rate of palatalization between /k/ and /g/, which is observed in both stem extension and diminutive formation (before -i, -ik, and -ok). In all cases, the rate of palatalization is underestimated for /k/. Palatalization of /k/ to [tʃ] is more phonetically natural than palatalization of /g/ to [ʒ]. Bhat (1974: 41) notes that velar stops generally become affricates or remain stops as a result of palatalization and if a language palatalizes voiced velars, it also palatalizes voiceless velars but not necessarily vice versa, which suggests that the g → ʒ change is typologically marked. Hock (1991: 73–77) proposes that palatalization arises when a fronted velar stop develops a fricative release, with the conversion of the resulting affricate to a fricative being a later development. In addition, the voiceless velar stop [k] is more acoustically similar to [tʃ] than [g] is to [dʒ] in terms of peak spectral frequency and duration of aperiodic noise (at least before front vowels), leading listeners to misperceive [ki] as [tʃi] much more often than they misperceive [gi] as [dʒi] (Guion 1998). Thus, [g] and [ʒ] can be argued to be more perceptually and articulatorily distinct than [k] and [tʃ] and the g → ʒ alternation can be argued to be less phonetically natural than the k → tʃ alternation. Phonetic naturalness has been argued to influence learnability of an input-output mapping when the reliability of the mapping is controlled (Finley and Badecker to appear; Wilson 2006). The [k]/[g] asymmetry observed in Russian may be another case of this phenomenon. If the palatalization rule for [g] is more difficult to learn than the rule for [k] and the diminutive suffixes -ok and -ek preceded by a velar sound unacceptable, the speaker is driven to choose -ik as the diminutive suffix after [g] more often than after [k], accounting for the relatively high productivity of -ik following [g]. An alternative explanation is provided by type frequency of the to-be-produced pattern (Bybee 1985, 2001), which is much higher for [tʃi], [tʃok] and [tʃik] than for [ʒɨ], [ʒok], and [ʒɨk]. In any case, phonetic naturalness alone cannot account for the present data because velar palatalization is much more likely before -ok than before -ik, despite the fact that [o] is a less natural trigger of palatalization than [i].

Another shortcoming of the model is that it underpredicts the rate of velar palatalization before the suffix -ik, especially when -ik attaches to a /k/-final noun. This prediction follows (in the MGL model) from the fact that -ik almost never attaches to velar-final inputs in the native lexicon and thus is predicted not to trigger velar palatalization. There are at least two possible explanations for why it should still sometimes trigger velar palatalization. First, the alveopalatal stem-final consonant may be used as a diminutive marker in its own right, especially when the consonant is /tʃ/. This hypothesis is supported by the fact that some labial-final bases take -tʃik rather than -ik as the diminutive marker, e.g., sup 'soup' → suptʃik. Secondly, -ik and -ek are extremely phonetically similar due to being unstressed and may be phonetically identical for at least some speakers, although Shvedova et al. (1980: 28) maintain that the distinction is only partially neutralized. Despite this

(near-)identity in sound to -ik, -ek is a much stronger trigger of velar palatalization, thus the two suffixes seem to be different choices in phonology (unless palatalization before -ek is driven by orthotactics). However, it is possible that some instances of -ik in the (written) data can be cases in which the speaker chose -ek (which triggered velar palatalization) and misspelled it as the more frequent -ik.

2.5. *Are the affix and the stem shape chosen simultaneously?*

Perhaps, the most interesting parameter in the MGL is the sequence of stages assumed in modeling morphophonological processing. Interestingly, the penultimate segment effect on palatalization rate (shown in Figure 2) allows us to distinguish between the two models in (18). Each stage in (18) is modeled by a separate Minimal Generalization Learner trained on the relevant input-output mappings. The two-stage model assumes that the stem extension (-i vs. -a) is chosen first, followed by the decision on whether to change the stem. The one-stage model assumes that the stem shape and the affix are chosen simultaneously.

(18) Two-stage Model:
Stage I:
Choose the suffix based on the borrowed base:
Ø → suffix / Base__

Stage II:
Modify the base to fit the suffix:
/Base/ → [Base] / __suffix

One-stage Model:
Stage I:
Choose the suffix based on the borrowed base and modify the base to fit the suffix:
/Base/ → [Base] - suffix

The two-stage model fails to predict the strong effect of the penultimate segment shown in Figure 2. Let us now examine why this is the case. In the one-stage model, the palatalizing rules that are applicable to a given stem differ in their reliability, with some rules being more likely to outcompete the general non-palatalizing rule than others. For instance, the stem /overlok/ is likely to undergo velar palatalization because the most reliable palatalizing rule that can apply to it (k → tʃi / [+cons;+son] o__) is very reliable (.805) and can easily outcompete the applicable general rule (C → Ci) with its .2 reliability. By contrast, the most reliable palatalizing rule that can apply to the stem /drink/ ([+son]k → [+son]tʃi) has a reliability of only 0.125, which means that it is likely to lose to the more general rule C → Ci whose reliability is .2, resulting in failure of palatalization.

Suppose instead that the suffix has already been chosen and it is -i. The model now needs to decide whether to palatalize the stem. Interestingly, although the

rules changing k → tʃ and g → ʒ are exceptionless and thus have a reliability value of 1, they can still sometimes lose to the more general rule "do nothing" because the reliability of "do nothing" is also quite high (86%). This is because most stems in the lexicon take -i and remain the same after the addition of -i.

However, with the stem change choice following affix choice raw reliability predicts no effect of penultimate segment identity. In this model, the reliabilities of all stem-changing rules are at 1, regardless of penultimate segment identity because velar palatalization never fails before -i in the lexicon on which the model is trained. Therefore, the model can capture segmental context effects only if they correspond to differences in rule type frequency (i.e., the number of word pairs supporting the rule) and therefore confidence-adjusted reliability. In this case, there is a difference in type frequency in the expected direction, which produces a difference in confidence-adjusted reliability but it is a very small one: even with confidence set to 95%, the estimated reliability of the palatalizing rule for back vowel contexts is 0.89–0.92, compared to 0.87 for front vowel contexts, and 0.82–0.85 for consonantal contexts. It appears unlikely that such small differences in estimated probability of palatalization would produce the relatively large penultimate segment effect observed in the data. Thus, the effect of the penultimate segment is accounted for by the model only if the stem change and the affix are chosen during a single decision stage in which the palatalizing rules compete with rules adding other stem extensions, such as -a (the one-stage model).

3. Artificial grammar learning

3.1. *Introduction*

The data from Russian strongly suggest that the productivity of velar palatalization is connected to whether the palatalizing affix is used mostly with inputs that can undergo velar palatalization or with inputs that cannot. However, the data are correlational in nature, so the direction of causation is uncertain. It is possible that the low productivity of velar palatalization before -ik and -i, whatever its cause, makes speakers of Russian avoid using -i and -ik with velar-final inputs (e.g., Berg 1998: 230–233; Martin 2007; Thomason 1976). If changing an input consonant in front of a certain suffix is difficult while keeping the consonant unchanged results in a suboptimal output, the best course of action may be to avoid using the suffix altogether. Furthermore, the dictionary is not a perfect model of the Russian lexicon as it exists in the mind of a Russian speaker. Therefore, the data on whose basis velar palatalization is acquired by the model are different from the data on whose basis velar palatalization is acquired by Russian speakers.

A way to address both of these issues is provided by artificial grammar learning. By training the subjects and the model on the same language featuring velar palatalization, we can maximize the similarities between their relevant learning

experiences. Furthermore, by varying the lexical distribution of the palatalizing suffix and keeping all other aspects of the competing rules constant, we can determine whether the distribution of the palatalizing suffix can influence the productivity of palatalization. Of course, this would only provide support for one direction of causation (i.e., the tendency to attach the palatalizing suffix to stems that cannot be palatalized rather than to stems that can influencing the productivity of palatalization) but will not rule out the opposite direction of causation. The influence in the opposite direction may also exist, resulting in a self-reinforcing loop in language change, but it is not tested in the present experiment.

Native English speakers were randomly assigned to two groups. Both groups were presented with an artificial language featuring two plural suffixes, -a and -i, and an exceptionless rule that palatalized velars before -i, turning /k/ into [tʃ] and /g/ into [dʒ]. The suffix -a never attached to velars. Since the same input-output mappings are used in both languages, phonetic naturalness is controlled. In both languages, velar-final singulars always corresponded to plurals ending in -tʃi or -dʒi. Both subject groups were presented with 30 singular-plural pairs in which the singular ended in a velar. Therefore, the palatalizing rule has the same type (and token) frequency in both languages. The difference between the two languages was that in Language I -i was not very productive with non-velar-final singulars, being used in only 25% of the cases with -a being used 75% of the time. In Language II, the rates were reversed: -a was used 25% of the time with non-velar-final bases while -i was used 75% of the time. In both cases, 40 non-velar-final bases were used. Just like velar-final bases, the non-velar-final bases ended in oral stops (/p/, /b/, /t/, and /d/). Non-velar consonants did not change when a suffix was added. The languages are shown in Table 1.

The only rules that are applicable for novel velar-final bases and are extracted by the MGL upon exposure to the two languages are presented in Table 1. As the table shows, the two languages differ only in the reliabilities of the rules that do not

Table 1. *The two languages presented to the participants. The first column shows the most general minimally competing rules, which perfectly describe the languages presented to learners but predict no difference in productivity of velar palatalization between them. The second column shows the corresponding rules extracted by the MGL. The next two columns show the difference between the two languages. N's indicate numbers of different singular-plural pairs exemplifying a rule in each language. MGL values are confidence-adjusted rule reliabilities.*

Language rule	MGL rule	Language I	Language II	Sample stimuli
{k;g} → {tʃ;dʒ}i / V__	k → tʃi / V__	N: 30		drɪk → drɪtʃi
	g → dʒi / V__	MGL: 0.85		drɪg → drɪdʒi
Ø → i / {p;b;t;d}__	Ø → i / $C_{[-son;-cont]}$__	N: 8	N: 24	drɪp → drɪpi
		MGL: 0.18	MGL: 0.57	drɪd → drɪdi
Ø → a / {p;b;t;d}__	Ø → a / $C_{[-son;-cont]}$__	N: 24	N: 8	drɪp → drɪpa
		MGL: 0.57	MGL: 0.18	drɪd → drɪda

require velar palatalization. The rule attaching -i without changing the preceding consonant is much more reliable in Language II than in Language I. Therefore, velar palatalization is predicted to fail before -i in Language II more often than in Language I. Importantly, in both Language I and Language II, the most reliable rules that can apply to a velar-final input are palatalizing. Thus if subjects always used the most reliable applicable rule, there would be no difference between the two languages. Thus, the model predicts a difference between Language I and Language II only if the choice between rules is probabilistic.

3.2. *Methods*

The experiment consisted of a training stage and a testing stage. In the training stage, subjects repeated singular-plural pairs presented to them auditorily over headphones. The auditory stimuli were recorded by me in a sound proof booth onto a computer. The stimuli were sampled at 44.1 kHz and leveled to have the same mean amplitude. They were presented to the learners at a comfortable listening level of 63 dB. The aural presentations of the words were accompanied by visual presentations of the referents on a computer screen. The training task is shown schematically in Figure 5 below. In the testing stage, participants were presented with novel singular forms (not presented during training) and asked to orally produce the plural.

Participants were recruited from the Indiana University subject pool. All participants reported being native English speakers with no knowledge of Russian or any other Slavic language and no history of speech, language, hearing, or learning impairments. The participants' repetitions of training stimuli were used to assess whether the training stimuli were perceived correctly. If a participant made perception errors on more than 5% of singular-plural pairs, s/he was excluded from the experiment. Two participants, both exposed to Language II, were excluded from the experiment on the basis of this criterion. Participants who used -i fewer than 10 times with velar-final singulars were excluded from the present analyses. Two participants, both exposed to Language I, were excluded based on this criterion. The

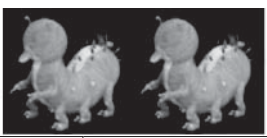
Video:						
Audio:		[bouk]			[boʊtʃi]	
Learner action:	Watch	Watch and listen		Watch	Watch and listen	Repeat aloud, then click
Duration:	300 ms	500-900 ms	500 ms	300 ms	500-900 ms	0-10s

Figure 5. *The training task.*

analyses below are based on thirty participants, half of whom were exposed to each language.

Participants were tested one at a time. The participant was seated in a sound-proof booth. The audio stimuli were delivered via headphones, while the participant's speech was recorded onto a digital audio tape using a head-mounted microphone. The experimenter was seated outside the booth and was able to hear the audio presented to the participant as well as the participant's productions. The participant was unable to see the experimenter. The participant's productions were scored by the experimenter online, as the participant was producing them. The stimuli were presented and ratings recorded using PsyScript experiment presentation software on Mac OS9.2. The order of presentation of the stimuli was randomized separately for each participant.

3.3. *Results and discussion*

First, it is important to note that, in novel items, velars become alveopalatals much more often than other consonants do for participants trained on both languages (the rate of labials changing into alveopalatals is 0% for both languages; for coronals, it is 2% for Language I and 8% for participants exposed to Language II, a non-significant difference). The difference between rate of coronal palatalization and the rate of velar palatalization is significant for both Language I (Wilcoxon signed ranks test, $Z(14) = 3.05$, $p < 0.01$) and Language II ($Z(14) = 2.63$, $p < 0.01$). Thus, there is evidence that participants really have acquired input-output mappings specifying that velars change into alveopalatals, rather than simply learning that -i should be preceded by an alveopalatal.

The participants were able to discover the distribution of -i and -a in the lexicon. Participants exposed to Language I used -i after alveolars and labials 30% of the time while participants exposed to Language II used -i 67% of the time ($t(28) = 4.4$, $p < 0.001$). Thus the training was successful in making -i more productive after non-velars in Language II than in Language I (the proportions of -i use by the subjects in the two groups are similar to proportions in the data to which they were exposed: 25% for Language I and 75% for Language II). More interestingly, participants exposed to Language I, the language predicted to favor velar palatalization by virtue of disfavoring the use of -i with non-velar-final singulars, palatalized the velar before -i 67% of the time, while participants exposed to Language II palatalized the velar before -i only 38% of the time ($t(28) = 2.316$, $p < 0.05$). Thus, the predictions of rule reliability are confirmed: even if the rules changing velars into alveopalatals before -i are exceptionless in the language, the more productive -i is with non-velar-final bases, the more likely velar palatalization is to fail.

Like speakers of Russian, subjects exposed to the artificial languages do not simply match the rate of velar palatalization to which they are exposed (100% for all subjects, regardless of whether they were exposed to Language 1 or Language 2). Rather, learners appear to be sensitive to the reliability of the 'just add -i' rule

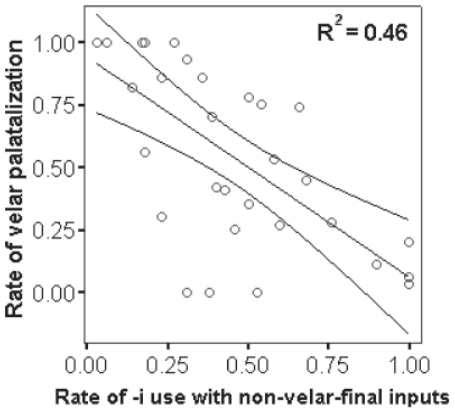


Figure 6. *Subjects for whom -i is productive with inputs that cannot undergo velar palatalization are the subjects for whom velar palatalization is unproductive. Curves show the 95% confidence region for the regression line.*

relative to the palatalizing rule. Figure 6 shows that there is a strong and significant negative correlation ($r = -0.68$, $p < 0.001$) between how much a subject uses -i with non-velar-final inputs and how likely s/he is to palatalize a velar before -i.

Once the correlation between rate of velar palatalization exhibited by a subject and his/her rate of -i use with non-velar-final inputs (Figure 6) is taken into account, the difference between subject groups no longer contributes towards explaining between-subject differences in velar palatalization productivity (according to an ANCOVA with rate of velar palatalization as the independent variable, the rate of -i use with non-velar-final inputs as a covariate, and Language as a fixed factor, rate of -i use is significant, $F(1,27) = 14.23$, $p < .001$, while Language is not, $F(1,27) = .082$, $p > .5$). Thus, the difference in productivity of -i with non-velar-final inputs, which is the independent variable predicted to influence productivity of velar palatalization by the MGL, accounts for all differences in productivity of velar palatalization between subjects that can be attributed to the artificial language they are exposed to.

4. Constraint-based alternatives

4.1. *Positive product-oriented generalizations*

So far, we have examined and confirmed the predictions of a rule-based model, which relies exclusively on source-oriented generalizations. We will now examine whether the same predictions can be derived from a model that uses product-oriented generalizations. The simplest product-oriented model is one in which the possible generalizations have the form 'outputs must have X' (Bybee 2001). In the

case of our two artificial languages, the relevant palatalizing schema would be 'plurals end in {tʃ;dʒ}i'. However, the type frequency of this generalization is equal across the two artificial languages. What then is responsible for the difference in the rate of velar palatalization across the two languages?

One possibility is that the participants rely on *conditional* product-oriented palatalizing schemas of the form 'if the plural ends in -i, the preceding consonant must be {tʃ;dʒ}'. The reliability of this generalization can be calculated as the number of plurals that end in {tʃ;dʒ}i divided by the number of plurals that end in -i, i.e., the backwards transitional probability from -i to {tʃ;dʒ}. This transitional probability differs between the two languages, since the denominator is much greater in Language II than in Language I, thus palatalization is correctly predicted to fail more often in Language II than in Language I. Some evidence that language learners can calculate transitional probability statistics was provided by Aslin et al. (1998), who found that infants are able to distinguish low-transitional-probability syllable sequences from high-transitional-probability sequences while controlling for sequence frequency.

While Aslin et al.'s data provide strong evidence for the learners being able to compute and use transitional probability statistics when other sources of data are unavailable, they do not indicate that learners preferentially rely on transitional probability. Alternatively, the participants may attempt to simultaneously satisfy 'plurals must end in -{tʃ;dʒ}i' and 'plurals must end in -Ci'. Bybee (1985, 2001) suggests that simple, or unconditional, generalizations like these can be weighted by their type frequency. Then the statistical support for the second generalization is greater in Language II than in Language I, thus it will be satisfied more often in Language II than in Language I. The support for the first generalization is the same across the two languages, thus it would be satisfied equally often. Thus the proportion of times a plural ending -i features velar palatalization is expected to be lower in Language II than in Language I.

The participants in the experiment can map an input [k] onto either [ki], [tʃi], [ka], or [tʃa]. The rate of velar palatalization is calculated as the frequency of mapping [k] onto [tʃi] divided by the frequency of mapping [k] onto either [tʃi] or [ki]. Thus rate of palatalization is low if *either* [k] is often mapped onto [ki] or [k] is rarely mapped onto [tʃi] or both. In Section 3, we observed that the more often a learner attaches -i to [t] and [p], the lower the rate of velar palatalization. The question then is, does attaching -i to non-velars lower the probability of mapping [k] onto [tʃi] and/or does it increase the probability of mapping [k] onto [ki]? The two alternative product-oriented accounts make contrasting predictions.

If learners rely on the reliability of 'if the plural is -i, the preceding consonant must be {tʃ;dʒ}', then the less reliable this generalization, the less often the preceding velar must be changed into {tʃ;dʒ}. Thus, the probability of the {k;g} → {tʃ;dʒ}i mapping should correlate with the probability of {t;d;p;b} → {t;d;p;b}i mapping. By contrast, the product-oriented model based on unconditional product-oriented generalizations claims that velar palatalization should fail whenever the partici-

pant obeys 'plurals must end in -i' while disobeying 'plurals must end in {tʃ;dʒ}i'. Since the two generalizations are weighted by type frequency, only the strength of the former generalization is affected by how often -i is attached to {t;d;p;b}. Thus, the probability of {t;d;p;b} → {t;d;p;b}i is predicted to correlate positively with the probability of {k;g} → {k;g}i. The same prediction is made by the MGL, where adding -i to alveolars and labials supports C → Ci without providing disconfirmation for {k;g} → {tʃ;dʒ}i.

In the present artificial language data, the negative correlation between the probability of producing {t;d;p;b} → {t;d;p;b}i and the probability of producing {k;g} → {tʃ;dʒ}i is weak and non-significant ($r = -.07$, $p = .71$). By contrast, the correlation between the probability of {t;d;p;b} → {t;d;p;b}i and {k;g} → {k;g}i is very strong ($r = .76$, $p < 0.00001$). Thus, the predictions of non-conditional product-oriented generalizations and the MGL are supported, while the predictions of conditional product-oriented generalizations are not.

A problem for the product-oriented account is the lack of restrictions on inputs that can give rise to outputs ending in {tʃ;dʒ}i (Pierrehumbert 2006). A possible solution, albeit one relying on generalization over word pairs, is that the palatalizing product-oriented generalization is in competition with some version of paradigm uniformity constraints (see Downing et al. 2005, for possible formalizations), stipulating that the singular and the plural have the same value on the place feature of a stem-final consonant, such as Ident-[velar].

Since this constraint-based account weights product-oriented generalizations by type frequency, one-stage and two-stage versions of the model do not differ in their ability to account for penultimate segment effects seen in Figure 2. In both the one-stage and the two-stage version of the model, the productivity of velar palatalization in context X is determined by the number of verbs ending in X{tʃ;dʒ}i (followed by the inflection) relative to the number of verbs ending in XCi (followed by the inflection). The type frequencies stay constant regardless of whether the source-product mappings featuring palatalization of the velar compete with mappings featuring -a. Among Russian verbs found in Sheveleva (1974), 63/507 (12%) verbs ending in $V_{[+back]}Ci + t^j$ end in $V_{[+back]}tʃi + t^j$, compared to only 8/125 (6%) for $V_{[-back]}Ci + t^j$ and only 12/185 (6%) for $CCi + t^j$.[11] Thus, the constraint-based account featuring positive product-oriented constraints weighted by type frequency may be able to account for the relatively high rate of palatalization following back vowels in the Russian data regardless of the number of stages assumed but, even more than the MGL, would incorrectly group front vowel contexts with consonantal contexts rather than with back vowel contexts.

One way the constraint-based account using positive product-oriented generalizations and paradigm uniformity constraints can be distinguished from the rule-based account offered by the MGL is by examining the effect of adding examples of tʃ → tʃi mappings to training for subjects exposed to one of the artificial languages presented in Section 3. For the rule-based model, these are examples of 'just add -i', which vote against velar palatalization. For the product-oriented

model, these exemplify 'plurals must end in {tʃ;dʒ}i', thus their addition favors velar palatalization.

4.2. *Negative product-oriented generalizations and Optimality Theory*

The artificial language with the largest difference between the frequency with which [ki] is observed (zero) and the frequency with which it is expected to be observed (Language II) is the language in which [ki] is produced relatively often. At first glance, this finding is inconsistent with the idea that the speakers avoid producing [ki] and [gi] to the extent that they are certain that [ki] and [gi] are systematically absent from the training data. A plausible way to estimate the likelihood that the absence of [ki] is not accidental is to calculate the difference between how often [ki] is expected to occur and how often it actually occurs. The actual frequency of occurrence of [ki] is zero across the two artificial languages. However, other [Ci] sequences occur much more often in Language II than in Language I. Thus, if the learner estimated how often [ki] would occur if it were just like the other [Ci] sequences, s/he would estimate a higher frequency in Language II than in Language I. Thus, s/he should be more confident that [ki] is to be avoided in Language II than in Language I. For example, Xu and Tenenbaum (2007) find that learners presented with three examples of the novel word *fep* paired with a picture of a Dalmatian infer that *fep* means 'Dalmatian' rather than 'any dog' more often than if only one *fep*-Dalmatian pairing is presented. They argue that the learners detect a suspicious correlation between *fep* and pictures of Dalmatians, which would be unexpected if *fep* could refer to any dog. If phonology learning worked the same way, we would expect that exposure to many examples of -i attaching to [p] and [t] in Language II would restrict -i to attaching only to [p] and [t], rather than any stop, compared to Language I. However, the learners actually avoid [ki] much *less* often in Language II than in Language I. This finding suggests that the learner does not act on the basis of an estimate of the likelihood that unobserved sequences are systematically absent, at least, as we shall see, if that estimate must be calculated over products rather than source-product pairs.

The dominant current approach to phonology is Optimality Theory (Prince and Smolensky 2004 [1993]). In Optimality Theory, the grammar consists of negative product-oriented generalizations (markedness constraints), which stipulate that certain segments or sequences are to be avoided, and source-oriented generalizations enforcing identity between morphologically related forms (paradigm uniformity constraints).[12] Standard Optimality Theory cannot account for the present data because competition between generalizations is resolved by strict ranking (i.e., the stronger generalization is obeyed all the time) while in the present data competition between generalizations is resolved stochastically. If conflicts between constraints voting for palatalization of the velar and constraints voting for keeping the velar unchanged were resolved via strict ranking, the palatalizing constraint would always win in both artificial languages as well as in front of any ex-

amined Russian suffix since the velar is observed to always undergo change in the training set. However, a stochastic version of Optimality Theory (e.g., Harmonic Grammar, Smolensky and Legendre 2006) is consistent with this aspect of the present data. Furthermore, if stochastic Optimality Theory is coupled with Boersma's Gradual Learning Algorithm (GLA, Boersma 1997, Boersma and Hayes 2001), it can account for the difference between Language I and Language II.

The GLA takes in a lexicon of source-product pairs and attempts to derive each product from the corresponding source given the current constraint weights (by adding a fixed amount of noise to the weights and converting the resulting noisy weights to a ranking). If the derived, i.e., expected, product does not match the observed product, the algorithm reduces the weights of constraints that are violated by the observed product and obeyed by the expected product and increases the weights of constraints that are violated by the expected product and obeyed by the observed product. All weights are adjusted by the same amount, which decreases over the course of training.

This procedure correctly predicts the difference between Languages I and II as long as we assume that the learner has extracted a general constraint like *i ('Do not produce products ending in -i') or *$C_{[-cont]}$i ('Do not produce products ending in a stop followed by -i'). These constraints are in competition with the constraint militating against products ending in -a (*a). In Language II, *i or *$C_{[-cont]}$i loses to *a more often than *a loses to *i or *$C_{[-cont]}$i. This is why these languages have more products ending in an alveolar or a labial followed by -i than by -a. The opposite is true for Language I. Whenever *$C_{[-cont]}$i (or *i) loses to *a, the weight of *$C_{[-cont]}$i is reduced and products ending in -i, and specifically, an -i preceded by a stop, including [k], become more likely to be produced. This increase in [ki] likelihood happens more often in Languages II than in Languages I, predicting the retention of [k] before -i to be more common when -i tends to often attach to non-velars. Importantly, even if there is a more specific constraint against [ki] (*ki), its strength is identical in both languages. Thus, the GLA, like MGL and unconditional product-oriented generalizations, correctly predicts that the addition of -i to labials and alveolars should positively correlate with the addition of -i to velars without necessarily depressing the absolute probability of mapping velars onto alveopalatals. Representative constraint weights are shown in Tables 2–3. These weightings were obtained in Praat (Boersma and Weenink 2009) with evaluation noise at 0, plasticity set to 10 with a decrement of 1 and all other settings at default values. As the tables show, the difference in acceptability (shown to the left of the table) between the [ki] form and the [tʃi] form is smaller in Language II than in Language I.

Close examination of Tables 2–3 reveals that the model correctly predicts that [ki#] should be produced more often in Language II than in Language I, provided that the choice between forms is stochastic. However, stochastic choice also makes incorrect predictions. In particular, the difference in harmony between [ki] and [tʃi] is much greater than the difference in harmony between [ti] and [tʃi] or the

Table 2. *Language I under the grammar containing *C$_{[-cont]}$i (a.k.a. *Stopi). The hand shows the best product for each source. The constraints are arranged in decreasing weight order from left to right. The 'C' in the table stands for [tʃ].*

k	*Stopi	Ident-alveolar	Ident-labial	*a	Ident-velar	
ki	*					−110.954
ka				*		−93.084
☞ Ci					*	−79.661

t	*Stopi	Ident-alveolar	Ident-labial	*a	Ident-velar	
ti	*					−110.954
☞ ta				*		−93.084
Ci		*				−109.125

p	*Stopi	Ident-alveolar	Ident-labial	*a	Ident-velar	
pi	*					−110.954
☞ pa				*		−93.084
Ci			*			−107.177

difference between [pi] and [tʃi]. Thus, the learner is predicted to erroneously palatalize labials and alveolars before -i (turning them into [tʃi]) much more often than s/he erroneously fails to palatalize the velar before -i. This directly contradicts the data from both loanword adaptation and artificial languages where palatalization of alveolars and labials is very uncommon. The prediction is made because in the simulations reported in Tables 2–3 all constraints are initialized with the same weight (100). The weights of Ident-Labial and Ident-Alveolar never change during training from their initial values. Thus the problem is solved if, instead of starting out with all constraints having the same weights, (output-output) faithfulness constraints are initialized as having higher weights than markedness constraints (e.g., 200 for faithfulness vs. 100 for markedness with plasticity at 10), as suggested by Hayes (2004). The high initial ranking of faithfulness seems uncontroversial for the artificial grammar learning experiment given that the to-be-learned stem changes are unattested in the learners' native language. More controversially, we also need to assume it for learners of Russian as a native language to avoid overgeneralization of velar palatalization to non-velars while allowing for the failure of velar palatalization in loanword adaptation before suffixes that are observed to always palatalize the velar in the lexicon.

Table 3. *Language II under the grammar containing *$C_{[-cont]}$i (a.k.a. *Stopi). The 'C' in the table stands for [tʃ].*

k	Ident-labial	Ident-alveolar	*a	*Stopi	Ident-velar	
ki				*		−96.801
ka			*			−103.358
☞ Ci					*	−79.743

t	Ident-labial	Ident-alveolar	*a	*Stopi	Ident-velar	
☞ ti				*		−96.801
ta			*			−103.358
Ci		*				−109.435

p	Ident-labial	Ident-alveolar	*a	*Stopi	Ident-velar	
☞ pi				*		−96.801
pa			*			−103.358
Ci	*					−110.662

Unlike taking the difference between observed and predicted sequence probabilities given the observed frequencies of occurrence of parts of the sequence across contexts, the GLA allows unobserved sequences to profit from the existence of similar sequences, including sequences in which some element(s) are shared with the unobserved sequence. Table 4 illustrates this feature of the GLA for a source sequence CD, which is mapped onto a product that either shares one of its elements with AB or shares neither element with it.

As can be seen from Table 4, the harmony of an AB product is increased if AD and CB products occur, provided that AD, CB, and AB are derived from the same source. The reasoning is that if AB consists of subsequences that are avoided, then it will also be likely to be avoided. By contrast, the observed-expected difference punishes sequences that are not observed if there are similar sequences that are observed because the observed sequences count as evidence for the systematicity of the gap. The reasoning is that if AB consists of subsequences that are unobserved, the fact that the entire sequence is unobserved is not good evidence for the avoidance of the sequence. In the GLA (and Harmonic Grammar / Optimality Theory with paradigm uniformity more generally), avoidance of a sequence is not estimated based on the products but rather based on the source-product mappings, i.e., it is source-oriented. This appears consistent with the present data.

Table 4. *The Gradual Learning Algorithm rewards sequences for the existence of similar sequences (iff the similar sequences are derived from the same source). The lower the harmony, the worse the sequence is estimated to be.*[13]

Probability during training				Harmony of CD → AB after training
CD → CB	CD → AD	CD → CF	CD → ED	
50	50	0	0	–366
0	0	50	50	–433

While the MGL extracts generalizations from the training data, the GLA does not, starting out with a set of generalizations that it simply weights relative to each other during training. It is important to note that not any set of generalizations would succeed in accounting for the difference between Language I and Language II. Specifically, if the learner starts out with constraints against specific Ci phoneme sequences (*ki, *pi, *ti) and no more general constraint (*i or *$C_{[-cont]}$ i), no difference in rate of velar palatalization across the two languages is predicted. For Russian, generalizations need to be affix-specific because some affixes starting with the same phoneme palatalize the preceding consonant, while others do not. We cannot rely on Universal Grammar to supply the generalizations since they do not always go with universal markedness (cf. Mielke 2008), as evidenced by the finding that -ok and -ek are better triggers for velar palatalization in Russian than -ik and -i are. Thus, to provide a full account for the present data, the GLA needs to be supplemented with some algorithm for generating the right generalizations to be weighted.

Given that the GLA needs to be supplied with generalizations, we can ask the same question we asked of the MGL: should we assume that the suffix is chosen first with the choice of whether or not to change the stem following, or are both chosen at the same time? Turning back to the context effects in Figure 2, is there a way to account for those context effects in a two-stage model consisting of two stacked Harmonic Grammars trained using the GLA? Let us imagine that the first grammar has already chosen the suffix, and the suffix is -i. Now we need to decide whether, given ok + i or ik + i in the input, we should output a sequence ending in [tʃi] or a sequence ending in [ki]. A simplified GLA training set is shown in (19). It is apparent that the training set contains no information for producing an effect of the penultimate stem segment on probability of velar palatalization. Thus, a two-stage GLA learner would not acquire the effect of context on velar palatalization shown in Figure 2.

(19) A simplified training set for the GLA deciding on whether to palatalize [k] before the stem extension -i in Russian:

aki	atʃi	100%	at	ati	100%	ap	api	100%
iki	itʃi	100%	it	iti	100%	ip	ipi	100%

Let us now consider a one-stage model. Here, the probabilities would be approximately as shown in (20).

(20) A simplified training set for a one-stage GLA learning Russian:

ak	aka	10%	at	ata	30%	ap	apa	50%
ak	atʃi	90%	at	ati	70%	ap	api	50%
ik	ika	50%	it	ita	30%	ip	ipa	50%
ik	itʃi	50%	it	iti	70%	ip	ipa	50%

The training set in (20) contains the information needed to distinguish between the different contexts, as long as the learner weights generalizations that make reference to those contexts. Thus, as shown in Section 2.5, the Minimal Generalization Learner would predict velar palatalization to be more productive following [a] than [i] because the rule 'k → tʃi / i__' is less reliable than the rule 'k → tʃi / a__'. Table 5 shows that the GLA is able to learn the distribution if it is equipped with constraints that refer to the transition from the penultimate segment of the stem to the final segment of the stem. Here, a vowel and a consonant obey Harmony (and disobey *Harmony) if the vowel is front while the consonant is alveopalatal or the vowel is back and the consonant is not alveopalatal.[14] Since both Harmony and *Harmony constraints were available to the learner and had equal weights at the beginning of training, the learner must have induced the right grammar from the training data. Thus, segmental context effects, like the ones observed in Russian palatalization, require a single-stage morphophonological grammar whether it contains rules or constraints.

A crucial difference between Harmonic Grammar (Smolensky and Legendre 2006) learned using the GLA (Boersma 1997) and the Minimal Generalization Learner (Albright and Hayes 2003) to be addressed in future research is that the GLA predicts that the existence of a competitor suffix (-a) is necessary for the difference between Language I and Language II to be obtained, since the weight of *i is changed only when it is in competition with another constraint (*a). Minimal Generalization Learner predicts that the removal of examples of -a from training should not eliminate the difference between the two types of languages, since the learner has more confidence in the productivity of 'just add -i' when it is supported by many examples. The same prediction is also obtained for positive product-oriented generalizations which are weighted by type frequency.

5. Conclusion

The hypothesis that rules compete for inputs with the outcome of this competition determined by differences in reliability or type frequency between the competing rules (Albright and Hayes 2003) predicts that a morphophonemic rule will lose productivity if the triggering affix comes to be used increasingly with inputs that

Table 5. *The one-stage GLA can capture the difference between stems containing [i] and [a] in productivity of velar palatalization: the [tʃi] form is better than the [ki] form following [a] while the [ki] form is better following [i].*

ak	Ident-alveolar	Ident-labial	*VCHarmony	CVHarmony	*VHarmony	VHarmony	*CVHarmony	VCHarmony	Ident-velar	
aki			*	*		*				−309.644
☞ aCi						*	*	*	*	−292.246
aka			*		*		*			−309.128
aCa				*	*			*	*	−298.541

ik	Ident-alveolar	Ident-labial	*VCHarmony	CVHarmony	*VHarmony	VHarmony	*CVHarmony	VCHarmony	Ident-velar	
iki				*	*			*		−293.762
iCi			*		*		*		*	−313.907
☞ ika						*	*	*		−287.467
iCa			*	*		*			*	−314.424

cannot undergo the rule due to not being in the class of inputs to which the rule can apply. This hypothesis is supported by loanword adaptation data in Russian as well as experimental data from artificial grammar learning. An alternative account of the effect is provided by competition between product-oriented generalizations and paradigm uniformity constraints (given that the right product-oriented generalizations are supplied). The product-oriented generalizations can be either unconditional positive, as in Bybee (1985, 2001) or negative, as in Optimality Theory / Harmonic Grammar (Prince and Smolensky 2004 [1993], Smolensky and Legendre 2006). The present data place several restrictions on models of grammar. First, the affix and the 'triggered' stem change appear to be chosen at the same time, rather than the affix being chosen first and then triggering or failing to trigger a stem change, unless palatalization is assumed to be triggered by unconditional positive product-oriented generalizations weighted by type frequency. Second, the choice between rules in the rule-based model or output forms in the constraint-based model must be stochastic, rather than the subjects always applying the most reliable applicable rule or producing the output with the highest harmony. Finally, given that the choice between forms to be produced is stochastic, existing morphologically complex words must be stored in memory and retrieved for production (e.g., Bybee 1985, 2001; Halle 1973; Vennemann 1974) if they are to be produced correctly close to 100% of the time. A restriction that is specific to stochastic Optimality Theory / Harmonic Grammar acquired using the Gradual Learning Algorithm (Boersma 1997) is that the relevant (output-output) faithfulness constraints must be ranked above markedness constraints at the beginning of learning (Hayes 2004) in order to avoid overgeneralizing stem changes to consonants with which it is not observed while frequently failing to palatalize consonants in front of suffixes that are observed to palatalize them.

Correspondence e-mail address: vkapatsi@uoregon.edu

Notes

* This research was supported by funds from NIDCD Research Grant DC-00111 and NIDCD T32 Traning Grant DC-00012 to David Pisoni and the Speech Research Laboratory at Indiana University. Many thanks to Luis Hernandez for technical assistance, and to Ken de Jong, Bruce Hayes, Janet Pierrehumbert, Kie Zuraw, and an anonymous reviewer for helpful comments.

1. Later stages of productivity loss have been documented in wug tests by Zimmer (1969) and Zuraw (2000).
2. Throughout this paper, by 'input' and 'output' I mean the input to the rule and the output of the rule, which could both be either underlying or surface forms. A rule here is an input-output mapping, in which both the input and the output are classical categories.
3. While Analogical Modeling of Language is not usually described as rule-based, (supra)context-outcome pairings are input-output mappings in which the input and the output are classical categories.
4. 'EXT' – stem extension, 'INF' – infinitive, 'PFV' – perfective, 'REFL' – reflexive.

5. According to Fasmer (2004), historically from Turkic 'jamtʃɨ' plus -ik, but could also be analyzed as being formed from jamsk(-oj) with the addition of -ik and velar palatalization where 'jam' was the name for horse stables used to change horses by messengers, and -sk is an adjectivizing suffix.
6. Since -ek and -ik are unstressed, because of vowel reduction they have a very similar phonetic realization, so the choice between them may be part of orthography, although Shvedova et al. (1980: 28) report that the -ik/-ek neutralization is incomplete. However, the answer to the question of whether the choice is made in orthography or in phonology is not relevant to the modeling of output stem shape as long as the choice of the allomorph follows the decision on whether to change the stem.
7. Theoretically, this alternative suggests that all rules get to apply all the time and compete with each other in a single processing stage with competition being resolved stochastically, which appears attractive. The alternative assumed by Albright and Hayes (2003) and adopted here suggests that rules are stacked so that rules that produce the same input-output mapping compete for application with each other in a separate winner-take-all stage such that only the most reliable rule gets to compete for application with rules that map the same input to other outputs, with the outcome of this second stage being resolved stochastically.
8. Assuming $p(pal) = \exp(-energy(pal)) / (\exp(-energy(pal)) - \exp(-energy(non\text{-}pal)))$, $p(pal)_{rescaled} = \exp(-energy(pal)/t) / (\exp(-energy(pal)/t) - \exp(-energy(non\text{-}pal))/t)$. Many thanks to Kie Zuraw for suggesting this formula, which avoids the problem of the rescaling producing negative probabilities.
9. This problem is exacerbated when impugnment is used. While versions of the model without impugnment are able to predict that /k/ is more likely to be palatalized than /g/, versions with impugnment incorrectly predict the opposite result except for stems with a penultimate back vowel.
10. The test was done on ranked scores to control for non-normality.
11. Stems ending in [ɨC] were not included because it is unclear if [ɨ] is an allophone of /i/ and thus if it should be coded as [+back], which it is phonetically, or [−back], which it might be phonemically.
12. The job of paradigm uniformity constraints can also be (and often is) done by faithfulness constraints, which enforce identity to an underlying form, which is an abstraction over morphologically related forms, rather than directly enforcing identity between observed morphologically related forms. The difference is not essential for the present discussion.
13. The table was created by a grammar containing *A, *B, *C, *D, *E, *F, *AB, *EF, Ident-C, and Ident-D. The relative ordering of the harmony values and the importance of attestation of similar strings rather than frequency are constant across parameter settings. For this simulation, evaluation noise was set to 0, initial weights were at 100, with default settings for all parameters.
14. A different (and language-specific) constraint would of course be needed to capture the fact that palatalization is even less productive after consonants.

References

Albright, Adam, & Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90. 119–161.

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

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

Bhat, D. N. S. 1974. A general study of palatalization. *Working Papers on Language Universals* 14. 17–58.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* 21. 43–58.

Boersma, Paul & Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32. 45–86.
Boersma, Paul & David Weenink. 2009. Praat: doing phonetics by computer (Version 5.1.07). [Computer program]. Retrieved May 19, 2009, from: http://www.praat.org.
Bybee, Joan. 1985. *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins.
Bybee, Joan. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.
Bybee, Joan. 2008. Formal universals as emergent phenomena. In Jeff Good (ed.), *Linguistic universals and language change*, 108–124. Oxford: Oxford University Press.
Downing, Laura J., T. A. Hall, & Renate Raffelsiefen (eds.), 2005. *Paradigms in Phonological Theory*. Oxford: Oxford University Press.
Fasmer, M. 2004. *Etimologicheskij slovar' russkogo jazyka* [Etymological dictionary of Russian]. Moscow: Astrel'.
Finley, Sara & William Badecker. To appear. Towards a substantively biased theory of learning. *Berkeley Linguistics Society* (*BLS*) 33.
Guion, Susan G. 1998. The role of perception in the sound change of velar palatalization. *Phonetica* 55. 18–52.
Halle, Morris. 1973. Prolegomena to a theory of word formation. *Linguistic Inquiry* 4. 3–16.
Hayes, Bruce. 2004. Phonological acquisition in Optimality Theory: the early stages. In René Kager, Joe Pater & Wim Zonneveld (eds.), *Fixing Priorities: Constraints in Phonological Acquisition*, 158–203. Cambridge: Cambridge University Press.
Hock, Hans H. 1991. *Principles of historical linguistics*. New York: Mouton de Gruyter.
Levikova, S. I. 2003. *Bol'shoj slovar' molodezhnogo slenga* [Big dictionary of youth slang]. Moscow: Fair-Press.
Martin, Andrew Thomas. 2007. *The evolving lexicon*. Los Angeles, CA: UCLA dissertation.
Mielke, Jeff. 2008. *The emergence of distinctive features*. Oxford: Oxford University Press.
Pierrehumbert, Janet B. 2006. The statistical basis of an unnatural alternation. In Louis Goldstein, D. H. Whalen, and Catherine T. Best (eds.), *Laboratory Phonology* 8. 81–107. Berlin: Mouton de Gruyter.
Prince, Alan & Paul Smolensky. 2004 [1993]. *Optimality Theory: Constraint interaction in generative grammar*. Malden, MA: Blackwell.
Rumelhart, David & James McClelland. 1986. On learning the past tenses of English verbs. In David Rumelhart, James McClelland & the PDP Research Group (eds.), *Parallel Distributed Processing,* vol. 2, 216–271. Cambridge, MA: M.I.T. Press.
Sheveleva, M. S. 1974. *Obratnyj slovar' russkogo jazyka* [Reverse dictionary of Russian]. Moscow: Sovetskaja Enciklopedija.
Shvedova, N. Ju., N. D. Arutjunova, A. V. Bondarko, V. V. Ivanov, V. V. Lopatin, I. S. Uluxanov & F. I. Filin. 1980. *Russkaja grammatika* [Russian grammar], vol. I. Moscow: Nauka.
Skousen, Royal. 1989. *Analogical modeling of language*. Dordrecht: Kluwer.
Smolensky, Paul & Geraldine Legendre. 2006. *The harmonic mind: From neural computation to Optimality-Theoretic Grammar*. Cambridge, MA: M.I.T. Press.
Thomason, Sarah G. 1976. What else happens to opaque rules? *Language* 52. 370–381.
Vennemann, Theo. 1974. Words and syllables in natural generative phonology. In Anthony Bruck, Robert A. Fox & Michael W. La Galy (eds.), *Papers from the Parasession on Natural Phonology*, 346–374. Chicago: Chicago Linguistic Society.
Wilson, Colin. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30. 945–982.
Xu, Fei & Joshua B. Tenenbaum. 2007. Word learning as Bayesian inference. *Psychological Review* 114. 245–272.
Zimmer, Karl E. 1969. Psychological correlates of some Turkish morpheme structure conditions. *Language* 45. 309–321.
Zuraw, Kie. 2000. *Patterned exceptions in phonology*. Los Angeles, CA: UCLA dissertation.