# Comparative Phonotactics[*]

Bruce Hayes
UCLA

September 2014

Submitted to the Proceedings volume for the 50th Meeting
of the Chicago Linguistic Society

This version is reformatted for web distribution and includes the abstract.

## Abstract

The phonotactic learner of Hayes and Wilson (*LI* 2008) discovers what could be called **absolute phonotactics**: using a maxent framework, it selects and weights constraints so as to maximize the predicted probability of the set of existing words against a backdrop of all possible strings. The same apparatus can be used for **comparative phonotactics**: given two populations of strings, A and B, we seek a grammar whose output probabilities accurately indicate the likelihood that any given novel string will belong to A or B.

Do language-acquiring children learn comparative phonotactics? I think it likely that they do, and indeed that they do so for multiple purposes. Such would include part-of-speech prediction (work of Christiansen), prediction of gender (work of Lyster and others), and two areas I focus on here. I put forth a comparative phonotactics that singles out words of the Latinate **vocabulary stratum** of English (Chomsky and Halle 1968), distinguishing it from the native stratum, and ponder how the presence of such strata in a language could be detected by a bootstrapping process. I will also use comparative phonotactic analysis to carry out **stem sorting** in the sense of Becker and Gouskova: the population of stems in a language can be sorted according to which affix allomorphs they take. The pattern predicted by stem-sorting-cum-allomorph-selection is often indistinguishable from the result of ordinary GEN+EVAL phonology, but in one area of Hungarian vowel harmony, the evidence is quite clear that the pattern must be the result of stem sorting.

[*] For helpful comments and advice I thank Michael Becker as well as colleagues who provided feedback on spoken versions of this paper at UCLA and at the CLS meeting. Special thanks to Kie Zuraw, who prepared the lexical search of Hungarian stems reported in §5.2. The web site for this paper, with supplemental materials, is www.linguistics.ucla.edu/people/hayes/ComparativePhonotactics.

## 1 Two kinds of phonotactics

I contrast the notions of “absolute phonotactics” and “comparative phonotactics.”

**Absolute phonotactics** is the study of well-formedness in phonology. The topic has a long history, but was laid out with particular cogency by Chomsky and Halle (1965), who noted that speakers have phonotactic judgments even of words they have never heard before; thus *blick* [blɪk] is non-existent but well-formed, while *bnick* [bnɪk] is non-existent and ill-formed. Further work (Scholes 1966, Chomsky and Halle 1968:416-418, and many others) has suggested that absolute phonotactics is gradient; for example, there are words (e.g. *poik* [pɔɪk]) that sound neither terrible nor perfect. In the approach to be taken here, gradience is captured by assigning words numerical scores, which have an explicit interpretation in the theory of probability.

For **comparative phonotactics**, we assume two populations of strings, which we can call A and B, and a grammar whose outputs are likewise probabilities, but which specifies whether any given string will belong to Population A or Population B. That is, there will be a probability assigned to the outcome “this form belongs to A”, a probability assigned to the outcome “this form belongs to B”, and the two probabilities sum to one. In order to do this well, we will need to select constraints that single out traits that successfully distinguish the A and B populations; perhaps exceptionlessly, perhaps only probabilistically.

This article explores the question of whether comparative phonotactics is a useful idea for phonology. I cover relevant theoretical background, then go through two case studies, and finally address some general theoretical questions raised by the concept.

## 2 Some earlier work in absolute phonotactics

My own earlier work on absolute phonotactics with Colin Wilson (Hayes and Wilson 2008) was directed toward the development of an explicit theory of phonotactics, under which grammars can be constructed that can accurately predict phonotactic well-formedness intuitions. We also sought to shed light on the question of learnability in phonotactics by developing a computational model that would learn a phonotactic grammar, based solely on exposure to a large set of well-formed words, much as is assumed to happen in real childhoods.

The theoretical foundation for our work was the theory of **maxent grammars** (see, e.g. Goldwater and Johnson 2003, Wilson 2006), which are a stochastic variant of the more general approach known as **harmonic grammar** (Smolensky 1986, Legendre et al. 1990). The content of our grammars closely follows mainstream Optimality Theory (OT; Prince and Smolensky 1993) in consisting of formalized constraints (Markedness only). Given some learning data, the model has the capacity to select constraints from a large space of logical possibilities, then assign them appropriate weights, i.e. nonnegative real numbers that reflect their strength in penalizing forms that violate them. A simple formula converts the constraint violations of a form, as multiplied by the weights and summed, into a probability value, which is held to be proportional to the word’s phonotactic well-formedness and constitutes the overt prediction of the theory. During learning, weights are set (using a provably convergent algorithm; Berger et al. 1996) to maximize the predicted probability of the set of existing words against a backdrop of all possible strings.

We offered two forms of empirical support for our model. First, to some extent it succeeds in replicating the phonotactic descriptions arrived at by skilled linguists through study. Second, to some degree the model is able to predict the phonotactic intuitions of human participants in experimental settings.

## 3 Extending the model to comparative phonotactics

Comparative phonotactics is in principle much simpler: for any input form we have just two candidates (Population A and Population B), representing a binary decision. As before, constraints will be like the Markedness constraints of Optimality Theory, except that they should say something like PREFER {[A], [B]} IF Y, where [A] and [B] are the designated populations and Y is some phonological configuration. In the output of the grammar, the hypothesis that a form will belong to Population A will be assigned a probability value between zero and one, and the hypothesis of Population B will be one minus the probability assigned to Population A.

The maxent math remains otherwise the same and follows the procedure outlined in (1)-(2); see Hayes and Wilson (2008) for fuller discussion. **Harmony** is defined as in (1).

$$(1)\ \mathrm{H}(x) = \sum_{i=1}^{N} w_i\, C_i\,(x)$$

where $x$ is some candidate, $\mathrm{H}(x)$ is the harmony value being computed for that candidate, $w_i$ is the weight of the $i$th constraint, $C_i\,(x)$ is the number of times that $x$ violates the $i$th constraint, and $\sum_{i=1}^{N}$ denotes summation over all constraints ($C_1$, $C_2$, … $C_N$).[1] Intuitively, harmony is the weighted sum of the constraint violations. The **probability** of a candidate (as expressed for a simple two candidate system) is calculated as in (2):

$$(2)\ \mathrm{p}(\mathit{Cand1}) = \frac{\exp(-\mathrm{H}(\mathit{Cand1}))}{\exp(-\mathrm{H}(\mathit{Cand1})) + \exp(-\mathrm{H}(\mathit{Cand2}))}$$

where $\mathrm{p}(x)$ = the predicted probability of candidate x; $\exp(y) = e^y$, where $e$ is the base of natural logarithms, about 2.718; and $\mathrm{H}(x)$ = the harmony of $x$, given in (1). The overall effect is that the probability of a candidate is lowered by constraint violations, more so with more highly-weighted constraints; and is raised by the constraint violations (again, more so with higher weights) of its rival.[2]

[1] Hayes and Wilson (2008) adopt slightly different terminology, calling (1) the formula for "scores".

[2] The equations in (1)-(2) may look familiar to many readers who know some statistics; they embody the formula for logistic regression, a commonly used technique in statistics used extensively by sociolinguists (Gorman and Johnson 2013). In other words, maxent reduces to logistic regression in a two-candidate system.

## 4 Application I: the Latinate stratum of English

In *SPE*, Chomsky and Halle (1968:373) proposed that languages with heavy admixtures of loanwords develop synchronically arbitrary **lexical strata** — groupings of vocabulary that have a purely diachronic origin (native vs. adapted foreign words) but are nevertheless apparent to native speakers as a synchronic phenomenon. In English, the principal strata are thought to be Native and Learned/Latinate, arguably with a Greek subdivision of the latter.

Of course, the strata cannot be justified as entities of synchronic grammar on etymological grounds, as most speakers do not know the historical origin of the words of their vocabulary. Rather, the words convey their stratal memberships in some way that emerges from their form. Under this view, words can actually belong to a different stratum from their etymological source. Thus *dish*, *mile*, *noon*, *pillow*, *sack*, and *wine* sound native, but are early loanwords from Latin, thoroughly nativized in their phonology over time (*OED*). For similar Japanese examples, see Ito and Mester (1995: 836).

Why should speakers internalize such stratal divisions? I suggest that this knowledge is crucial to command of **style**. Fluent speakers know that certain contexts (education, science, bureaucracy) call for using Latinate vocabulary and other contexts (vernacular ones) call for not using it; they use their knowledge of Latinity to guide their productions (as well as their expectations in perception) across contexts.[3]

Ito and Mester (1995:821) suggest that membership in lexical classes is gradient. This corresponds to my own intuitions with regard to Latinity. For example, the following words (to which we will return later on) strike me as very Latinate indeed: *objectionable*, *veterinarian, protectionism, sexuality, vegetarian, reactionary, perfectionism, confectionery, naturalistic,* and *heterogeneity*. In contrast, these words strike me as being not at all Latinate: *smooth*, *yield*, *swish*, *wield*, *dwarf*, *swab*, *yarn*, *wind* ([waɪnd]), *gift*, and *twelfth*. I find the following forms to be somewhat Latinate but not really that strongly Latinate: *palate*, *taxi*, *motor*, *stupid*, *suitor*.

What could constitute the language learner's evidence for strata? There are several possibilities.

First, there is the characteristic **cooccurrence patterns of morphemes**. Thus, for instance, in the data considered below, when a word begins in a Latinate prefix, then it is more likely than otherwise to end with a Latinate suffix.[4] We will further consider such cooccurrences below.

Second, as emphasized in *SPE*, Latinate forms tend to undergo **phonological alternations** resisted by Native forms. For instance, the rule of Trisyllabic Shortening (*SPE*, 180) is generally applicable only in the Latinate stratum of English.

[3] The many Native/Latinate (near-)synonym pairs in English attest to this stylistic need; *begin*/*commence*, *job*/*occupation*, *baby*/*infant*, *refill*/*replenish*, *forecast*/*projection*, etc. It also seems possible that speakers who have learned what is Latinate use this information to predict other things, such as which verbs take double object constructions (Gropen et al. 1989).

[4] Of approximately 4,252 forms with a Latinate prefix in the lexical database described below, 2,359, or 55.5%, have a Latinate suffix. Of 13,492 forms without a Latinate prefix, 4,224, or 31.3%, have a Latinate suffix.

While both of the above criteria may have some validity, here I will focus on the evidence from **phonotactics**, a domain discussed insightfully by Plag (2003:83) and Ito and Mester (1995). In what follows, we study the ways in which the Latinate and Native strata of English differ phonotactically.

### 4.1 The comparative phonotactics of the English Latinate and Native strata: getting started

To study this topic, we need first to break out of a circularity: the comparative phonotactic analysis we will do requires a labeling of forms as Latinate or Native, but this status is justified in part by the phonotactic evidence. This circularity is both a theoretical issue and a practical one. To get started, let us solve the practical issue by fiat: I will suppose that any word of at least seven letters ending in one of these suffixes (defined orthographically) is Latinate: *-able*, *-acy*, *-al*, *-ance*, *-ancy*, *-ant*, *-ary*, *-ate*, *-ated*, *-ation*, *-ator*, *-atory*, *-ence*, *-ency*, *-ent*, *-graphy*, *-ia*, *-iac*, *-ian*, *-ible*, *-ic*, *-ical*, *-ician*, *-ific*, *-ify*, *-ine*, *-ism*, *-ist*, *-ity*, *-ium*, *-ive*, *-ize*, *-ular*, *-logy*, *-or*, *-ory*, *-ous*, *-sis*, *-tion*, *-ure*, *-us*.[5] This is basically the longest list of Latinate suffixes (including Greek as a subcategory) that I could think of when embarking on the problem.

Is this acceptable as a heuristic criterion? I checked this by implementing it in a spreadsheet containing my lexical database and inspecting the result. The lexical database is my edited version (`linguistics.ucla.edu/people/hayes/EnglishPhonologySearch`) of the Carnegie-Mellon Pronouncing dictionary (`speech.cs.cmu.edu`); it consists of all the words in the CMU database that have a lemma frequency of at least one in the CELEX lexical database (Baayen et al. 1995). It has been worked over attempting to repair some of the many transcription errors found in the original CMU database. Examining the words that would qualify as Latinate or Native by the above criterion, I felt that it was doing not too badly, well enough to serve as the basis of some comparative phonotactic exploration.[6]

### 4.2 Setting up the grammar

Let us review the grammar type we are working with. It consists of Markedness constraints that assess penalties for particular output candidates. The candidate set is very simple, consisting of just two candidates per input, the output classifications [Latinate] and [Native]. The constraints will take the form PREFER [LATINATE] IF X (or PREFER [NATIVE] IF X), where X is some sort of phonological configuration. The input to the grammar is simply a word, which I will assume appears in its surface representation.

In a fully-principled approach, the constraints used for a comparative phonotactic grammar would be located by algorithm (as in Hayes and Wilson 2008) from a huge set of logically possible constraints. At this exploratory stage, however, I felt it would be useful to study instead the constraints for which I had some reason for thinking *a priori* that they would be effective. I worked on the problem by experimentation, trying a wide variety of constraints and discarding

[5] In compiling this list of suffixes I was assisted by the careful description in Marchand (1969).

[6] The reader may check for herself by examining the full dataset at the article website.

those which did not work well. I drew on my knowledge of English and Latin historical phonology,[7] as well as the intensive analysis of English segmental phonology in *SPE*.

## 4.3 The constraints

It proved easy to find constraints that penalized Latinate status: by and large, Latin had a stricter phonotactics than English, so that the English-specific patterns absent in Latin serve as a good basis for predicting Native status.[8] A simple case involves word-initial clusters of the form **[s] + nasal**. These are known to have been obliterated in the early stages of Latin by a sound change that effaced the sibilant in this position. Thus the earlier form **sniks* had become *niks* 'snow' by the stage of Latin from which we take our borrowed vocabulary. The constraint in my comparative phonotactic grammar is stated as PREFER [NATIVE] IN [$_{\text{word}}$ s [+nasal], which is to be read, "Assess a penalty to the [Latinate] candidate for any word that begins with [s] followed by a nasal."

There are many similar cases. Latin had no [f] before obstruents (Hayes and White 2013), although [ft] is reasonably common in English words. More straightforwardly, there are single sounds of English that correspond to no Latin sound (or, more accurately, no evolved version of a Latin sound). Thus, we can set up PREFER [NATIVE] IN [ʊ] and PREFER [NATIVE] IN [aʊ].

Various Latin sounds do appear English, but with restricted distributions due to sound change (either in the history of Latin/Romance prior to borrowing, or in the history of English). Thus, Latin *w* appears widely in English words, but only after the velars [k] and [g] (*quiescent*, *sanguine*); elsewhere, it shows up as [v] instead: *convivial*. Thus, a constraint system penalizing Latinity in the presence of non-postvelar [w] will help identify Native forms. I treat this contextual variation with the maxent equivalent of OT constraint ranking: we place both PREFER [NATIVE] IN [w] and PREFER [LATINATE] IN {k,g} + w in the grammar. With the weights given below, the two will largely cancel each other out in *velar + w* sequences, the difference resulting only in a mild preference for Latinity in this context. Latin [k] and [g] likewise have a contextual outcome in English, since when they appear before what once was a nonlow front vowel, they underwent historical Velar Softening (*SPE* §4.5) and appear instead as [s] and [dʒ]: *concision*, *cogent*. The pattern was obscured by the later Great Vowel Shift, whereby the English nonlow front vowels [iː, i, eː, e] evolved into modern [aɪ, ɪ, iː, ɛ]. The result is that a penalty for velar stops before the set of modern triggers will identify Native status in words like *kite* or *geese*.

Long and short Latin [u] are distributed thus: before nonfinal coda consonants they show up as [ʌ] (*ungulate*), else [uː] after coronals (*duplicate*),[9] else [juː] (*circuitous*). Thus the configuration [uː] after noncoronals will help identify Native words (*pooch*, *cool*); and since there is basically no source for [ʌ] in Latinate words other than *u*, [ʌ] in open syllables will also identify Native forms (*cousin*, *buffalo*). Before nonfinal codas, [uː] is a cue for Native status;

[7] Some good sources: Jespersen (1909) for English and Sturtevant (1920) and Allen (1978) for Latin.

[8] A prominent area where Latin was phonotactically more permissive than English was in permitting long vowels in non-final closed syllables (Allen 1978, ch. 3). But these appear to have been loan-adapted as short; so that the constraint environment Vː CC (Table 1, #7 below) actually turns out to favor Native, not Latinate status.

[9] This provision holds for most varieties of American English; British and some other varieties retain the [j] in many cases after coronals; see Wells (1983:247).

indeed this is only part of a larger pattern based on the classification of [uː] as a long vowel. Exceptions to the general ban on long vowels before nonfinal coda consonants are more common in the native vocabulary (*shield*, *angel*).

The [juː] corresponding to orthographic *u* is the primary source for [j] in Latinate words; otherwise historical Latin *j* appears as [dʒ] (*judicious*). Restricting the analysis to initial position (because glide formation processes obscure the pattern medially), the analysis proposed here posits conflicting constraints PREFER [NATIVE] IN INITIAL [j] and PREFER [LATINATE] IN INITIAL [juː].

The palato-alveolars [ʃ, ʒ, tʃ, dʒ] were not phonemes in Latin, but in the target dialect of American English they are abundant in Latinate words, due to historical Alveolar Palatalization, which merged earlier [sj, zj, tj, dj] to [ʃ, ʒ, tʃ, dʒ]; *nation*, *vision*, *natural*, *gradual*. However, Alveolar Palatalization applied only in the ambisyllabic position (medial pre-atonic) that conditions so many processes of English phonology (Kahn 1976, Gussenhoven 1986). Thus, palatoalveolars occurring other than in ambisyllabic position diagnose Native status, as in *ship*, *chip*, *jail*, *lash*, *patch*, *badge*.[10] As with [w] and velars, we can describe the pattern well with conflicting constraints: PREFER [NATIVE] IN PALATOALVEOLARS is in conflict with PREFER [LATINATE] IN AMBISYLLABIC PALATOALVEOLARS.

Here are some miscellaneous configurations preferring Native status: {t,d}l (*antler*, *bedlam*), final main stress,[11] dental or alveolar obstruents before [w] (*twin*, *dwell*), [ŋ] (allophonic, and rare, in Latinate; *sanctify*) and [θ] (*thumb*; Latinate cases rare and limited to the Greek substratum; *theologian*).

Let us turn to the constraints that penalize Native status and favoring Latinity. At first blush these might be expected to be missing, since the Latin loanwords have existed in English for a long time and might be expected to have been gradually cleansed of their foreign phonological character. Yet there is one property that blatantly singles out Latinate words, which is their length. The native vocabulary at least originally followed a pattern of one maximally disyllabic trochaic metrical foot (stressed syllable plus sometimes an additional stressless syllable) and Latinate words often exuberantly exceed this limit. Indeed, it seems to be part of our folklore to use the term "long words" to characterize not just words that are literally long, but words that are fancy, learned, and so on — i.e. which bear the stylistic mark of Latinate words. Constraints penalizing long [Native] words and short [Latinate] words are thus effective and I implemented a suitable set.

There are a few characteristically Latinate phoneme sequences that escaped regularization and serve as diagnostics of Latinity. [mn] appears widely in Latinate (or Greek subtype) words (*alumnus*, *damnation*, *amnesia*), but only rarely in Native forms (*chimney*). Similar sequences that are strongly Latinate are [pʃ] and [kʃ] as well as stressless [iə] and [ɚə].[12] Certain individual

[10] The absence of examples with [ʒ] results from this sound being largely limited to ambisyllabic position, with exceptions (*Zhivago*, *soupe du jour*, *rouge*, *garage*) only in the Foreign stratum.

[11] In a full grammar this would have to be limited to nonverbs, since Latinate verbs with final stress are abundant (*perfect*, *consist*, *reside*, etc.).

[12] Other than these two cases, hiatus of two stressless vowels moderately prefers Native status.

phonemes, characteristically unmarked ones, also are weak cues favoring Latinity: [ə], [n], [t], [v].

All of the above I formalized as constraints, using a simple program (www.linguistics.ucla.edu/people/hayes/EnglishPhonologySearch/) to assess the violations of all constraints for every form in the corpus. For the complete list, see Table 1 below.

### 4.4 The constraint weights

The weights chosen for the analysis are those that provide the best fit to the observed data; that is, that best predict my heuristic classification of words into Latinate and Native. The setting of the weights that accomplishes this goal can be done in many different ways; I used a script in R (R Core Team. 2013); for details the reader may consult this script, which is posted online at the article website. The software follows a standard search criterion, that of maximizing the predicted probability of the observed data.

Weighting made use of the full set of 17,744 database words, each construed as an input form and provided with output candidates [Latinate] and [Native]. One of the candidates was labeled as the "winner", according to the criterion of Latinity laid out in §4.1; and each candidate was provided with the full set of violations for all constraints used.

### 4.5 Results

Table 1 gives the results. Column 1 describes the content of each constraint; all are discussed above in §4.3. Column 2 gives what candidate the constraint prefers ("N" is Native; "L" Latinate), and Column 3 lists the indices of whatever constraints it conflicts with. Column 4 gives the calculated weight, and column 5 gives the result of a Wald significance test. Generally, in creating the grammar I kept only constraints that were significant at the .05 level; but as a matter of scientific interest I also included some non-significant constraints that were assigned very high weights.[13] Columns 6 and 7 give the percentage of Latinate and Native forms in the database in which the constraints are violated; to obtain actual counts multiply these values by 6,583 and 11,161 respectively.

**Table 1**. Comparative-phonotactic grammar for [Latinate] status in English

| Constraint | Prefers | Conflicts with | Weight | *p* | Lat. % | Native % |
|---|---|---|---|---|---|---|
| 1. $[_{\text{word}}$ s [+nasal] | N | 22 | 12.47 | 0.935 | 0.00% | 0.56% |
| 2. Monosyllabic | N | | 6.53 | <.001 | 0.00% | 31.53% |
| 3. $\begin{bmatrix}\text{alveolar}\\ \text{stop}\end{bmatrix}$ l | N | 24 | 2.57 | 0.013 | 0.02% | 0.36% |
| 4. $[_{\text{word}}$ j | N | 31 | 2.50 | 0.013 | 0.70% | 0.92% |
| 5. [ft] | N | 24 | 1.66 | 0.106 | 0.02% | 0.49% |
| 6. Disyllabic | N | | 1.52 | <.001 | 12.72% | 47.18% |

[13] I judge that it is a matter for future empirical research to what extent standard significance tests line up with the generalizations actually internalized by native speakers.

| | | | | | | |
|---|---|---|---|---|---|---|
| 7. V:CC | N | | 1.28 | <.001 | 0.21% | 1.67% |
| 8. [ʊ] | N | | 1.22 | 0.031 | 0.06% | 0.68% |
| 9. Final main stress | N | | 1.22 | <.001 | 2.25% | 45.26% |
| 10. $\begin{bmatrix}+\text{cor}\\-\text{son}\end{bmatrix}$ w | N | 24 | 1.17 | 0.103 | 0.05% | 0.95% |
| 11. $\begin{bmatrix}\text{velar}\\\text{stop}\end{bmatrix}$ [i, ɪ, eɪ, aɪ] | N | | 1.10 | <.001 | 0.93% | 2.61% |
| 12. [w] | N | 29 | 1.09 | <.001 | 2.13% | 4.95% |
| 13. [aʊ] | N | | 1.00 | 0.001 | 0.24% | 2.12% |
| 14. [ʌ] in open syllable | N | | 0.98 | <.001 | 0.88% | 2.38% |
| 15. [−cor] u | N | | 0.67 | 0.091 | 0.18% | 1.12% |
| 16. Default preference for [Native] | N | 21-33 | 0.62 | <.001 | | |
| 17. [ŋ] | N | | 0.58 | 0.002 | 1.05% | 2.88% |
| 18. $\begin{bmatrix}\text{V}\\-\text{stress}\end{bmatrix}\begin{bmatrix}\text{V}\\-\text{stress}\end{bmatrix}$ | N | 30, 32 | 0.41 | 0.002 | 12.75% | 2.24% |
| 19. [θ] | N | | 0.38 | 0.018 | 2.08% | 2.46% |
| 20. Palato-alveolar | N | 26, 33 | 0.17 | 0.044 | 30.02% | 14.07% |
| 21. [ə] | L | | 0.16 | 0.002 | 81.07% | 33.40% |
| 22. [n] | L | 1 | 0.38 | <.001 | 57.22% | 29.14% |
| 23. [v] | L | | 0.60 | <.001 | 13.70% | 6.61% |
| 24. [t] | L | | 0.84 | <.001 | 54.95% | 30.59% |
| 25. [mn] | L | | 1.14 | 0.04 | 0.40% | 0.04% |
| 26. V $\begin{bmatrix}\text{Palato-}\\\text{alveolar}\end{bmatrix}\begin{bmatrix}\text{V}\\-\text{stress}\end{bmatrix}$ | L | 20 | 1.17 | <.001 | 20.72% | 2.71% |
| 27. At least 5 syllables | L | | 1.24 | <.001 | 50.18% | 3.75% |
| 28. At least 4 syllables | L | | 1.36 | <.001 | 87.28% | 21.30% |
| 29. $\begin{bmatrix}\text{Velar}\\\text{stop}\end{bmatrix}$ w | L | 12 | 1.46 | <.001 | 1.82% | 1.33% |
| 30. [ĭə] | L | 18 | 1.52 | <.001 | 6.90% | 0.57% |
| 31. [$_{\text{word}}$ ju | L | 4 | 1.85 | 0.076 | 0.67% | 0.46% |
| 32. [ ɚ̆ə] | L | 18 | 1.94 | <.001 | 3.39% | 0.26% |
| 33. {p,k}ʃ | L | 20 | 3.71 | <.001 | 2.84% | 0.07% |

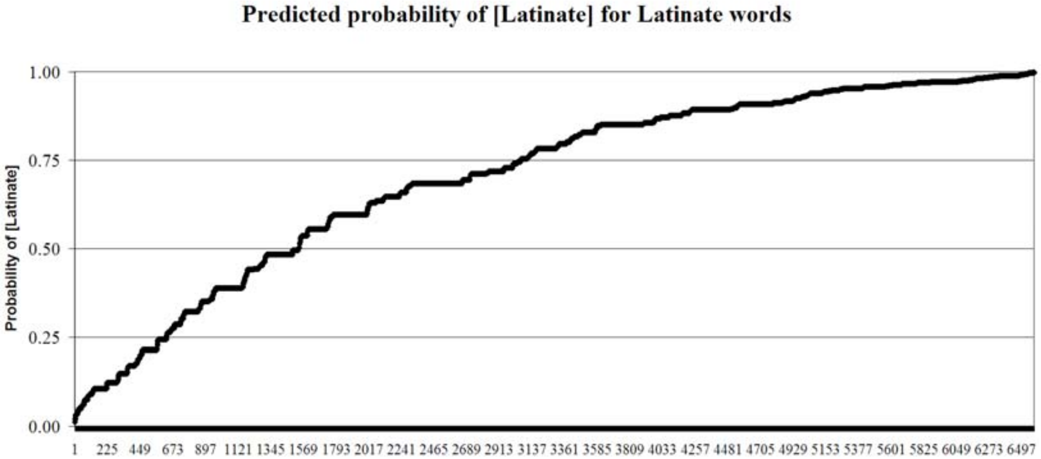
I illustrate how the grammar works by computing the probability of Latinate status for one particular form: *frustration* [ˌfrʌsˈtɹeɪʃən], following (1)-(2). *Frustration* violates five constraints penalizing Native status, given here with their weights: Prefer [Latinate] if [ə] (0.16), Prefer [Latinate] if [n] (0.38), Prefer [Latinate] if [t] (0.84), Prefer [Latinate] if V $\begin{bmatrix}\text{Palato-}\\\text{alveolar}\end{bmatrix}$ $\begin{bmatrix}\text{V}\\-\text{stress}\end{bmatrix}$ (1.17), and Prefer [Latinate] if at least 4 syllables (1.36). Each constraint is violated just once, so we can sum up the harmony of the Native candidate (cf. (1)) as 3.91. *Frustration* also violates two constraints penalizing Latinity, Prefer [Latinate] if Palato-alveolar (0.17) and the default preference for Native status (0.62), so the harmony of the Latinate candidate is 0.79. Plugging these harmony values into formula (2), we find that the predicted probability of the Latinate output is 0.958. The upshot is that *frustration* is claimed to be quite Latinate, but not maximally Latinate.

## 4.6 Performance of the Latinity grammar

Let us first check the performance in a purely intuitive way, returning to the words I had rated myself at the start of §4. The ten words I had asserted to be highly Latinate are indeed the ten words with the highest Latinity probabilities according to the grammar (all above 0.997): *objectionable*, *veterinarian*, *protectionism*, *sexuality*, *vegetarian*, *reactionary*, *perfectionism*, *confectionery*, *naturalistic*, and *heterogeneity*. The bottom 150 words on the list output by the grammar as ranked by Latinity all have scores less than 0.00005; the forms listed above (*smooth*, *yield*, *swish*, *wield*, *dwarf*, *swab*, *yarn*, *wind* ([waɪnd]), *gift*, and *twelfth*) are representative.[14] The forms I judged to be intermediate in Latinity (*palate*, *taxi*, *motor*, *stupid*, *suitor*) all have scores of about 0.21. Plainly, experimental work is needed to make serious claims here, but nevertheless I would judge that the results are on the right track.

The grammar also commits plenty of errors, but many of them strike me as resulting from the insufficiency of the heuristic criterion I had used to assess Latinity, not the grammar itself. A characteristic example is *sardine*, which my criterion had flagged as Latinate due to the orthographic final *-ine*; disyllabicity and final stress leads the grammar to classify it as non-Latinate (score 0.048). The heuristic was also defective in that I missed a few rare Latinate suffixes, such as *-iac*. The full predictions of the grammar may be downloaded from the article website.

To assess the aggregate performance of the grammar, I separated out the Latinate and non-Latinate words (by my heuristic classification), then sorted each by ascending predicted probability of Latinity. In a perfect grammar, the curve for the Latinate words will hug the top of the chart, and that for the Native words will hug the bottom; the actual result is at least substantially in this direction.


Predicted probability of [Latinate] for Latinate words

[14] These were selected by hand for variety, as the ten predicted most native words all begin with *s + nasal* clusters.

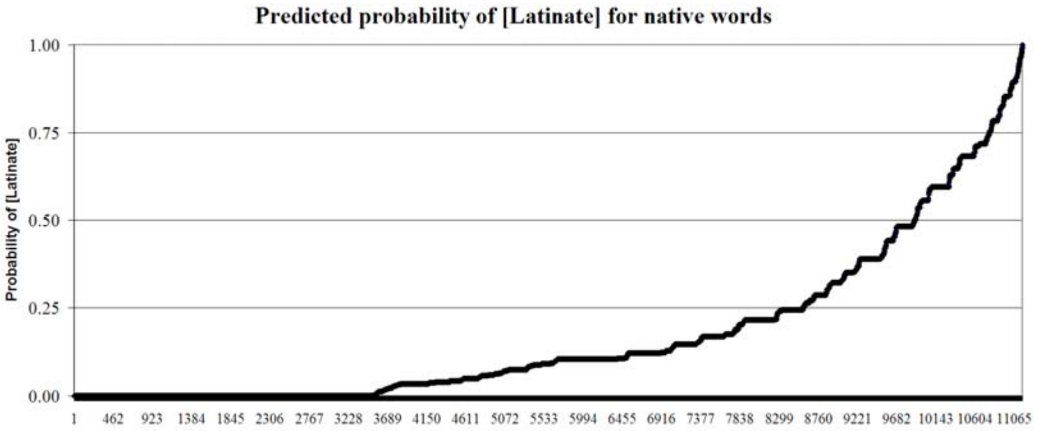


**Figure 1:** Sorted predicted probabilities for Latinate and Native words

The average error (distance from 1 for Latinate forms, and from 0 for Native forms) is 0.219.[15]

### 4.7 General discussion

I wish to make only modest claims for the analysis just given, of which the main one is as follows: quite a few observations made in traditional scholarship and in *SPE* receive empirical backup when, as here, they are implemented in a formal grammar and tested against a corpus. That is, many of the data generalizations found by earlier scholars appear to be good ones.

The question, "Is Latinity predictable on phonotactic grounds?" is trickier. For one thing, while I believe that word length plays a major role in speakers' intuitions about Latinity, my use of word length constraints when coupled with a suffix-based heuristic criterion for Latinity is hardly valid statistically. Words with suffixes are, all else being equal, likely to be longer, which makes the reasoning circular. When the model is deprived of its length constraints, average error rises from 0.219 to 0.271.

The heuristic I used for Latinity is itself ad hoc. It would be far better, I think, to undertake psycholinguistic experiments for Latinity on educated native speakers. I'm not sure what the best methodology for such testing would be; perhaps one could ask "Please rate on a scale of 1-7 whether this imaginary word would be likely to be used in scientific or scholarly writing." The reader can imagine herself a subject in such a test with a comparison like *temnication* (model prediction of Latinity for this word: 0.992) vs. *pookichation* (model prediction: 0.397).

### 4.8 How could the model be improved?

To begin, I doubt that the basic analytic work of finding effective constraints is complete; it is likely that further study, along with methods of machine search, would find more.

Second, the model does not use morphological information. I experimented briefly with this, expanding the grammar to include constraints based on the Latinate prefixes listed in Marchand

[15] N.B. random guessing yields an average error of 0.5.

(1969). (I could not use suffixes, since they were already the basis of the Latinity heuristic). To my surprise, this procedure turned out to help very little, reducing average error only to 0.215 (from 0.219). I have not yet tried to isolating Latinate stems (e.g., as stems that cooccur with Latinate suffixes), suggested by Ito and Mester (1995:818).

Going beyond the actual properties of words (for which we can write grammars), I suspect that modeling Latinity would be greatly aided if we could use a data source that is currently unavailable. Specifically, I think speakers may apprehend Latinity in part using the source from which they learned a word in the first place: words learned at the parent's knee are likely to be Native, words learned from books (especially harder books) are likely to be Latinate; such intuitions might then be refined later on as the child gathers enough data to develop a comparative grammar using both phonotactic and morphological constraints. It is possible that databases of whole childhoods, which are only now starting to be gathered (Roy et al. 2006) will make the study of the role of source information in vocabulary strata more feasible.

### 4.9 A theoretical point about the phonotactics of lexical strata

The Latinity pattern of English is evidence against theories (e.g. Ito and Mester 1995) that assert that the vocabulary strata are nested (Native words fill a subset of the phonotactics of the foreign words). Plainly, in the analysis here there is no subset relation in either direction, given that there are constraints that penalize both Latinate and Native status. Indeed, one might argue that the lexical strata of Japanese are likewise not nested; this is indicated by the evidence gathered by Kawahara et al. (2005), though these authors are cautious about asserting the point from their data.

As noted above, lexical strata usually emerge historically from substantial loanword importation. In light of this, I think we should expect *a priori* that there would be no subset relation, given that source and recipient languages are likely to be phonotactically complex in different ways.

## 5 Application II: finding the environments for phonological processes by sorting the stem inventory

I follow here in modified form an idea put forth by Becker and Gouskova (2012) about how to learn environments for phonological processes: in many cases it appears helpful to proceed by **stem-sorting**. Suppose we have some affix that exists in two allomorphic forms **a** and **b**. The stems that take these allomorphs can be considered as populations, or more specifically sublexicons; in the terms of Becker (2009), they are the "**a-takers**" and "**b-takers**". The idea is that language learners sometimes perform comparative phonotactics on the two populations and use the result to distribute the affix allomorphs.

This is very different from the classical approach to allomorph distribution proposed in Optimality Theory. In the latter, one supposes appropriate underlying forms, one per morpheme, and the GEN component actually *creates* the affix allomorphs. The EVAL component, consisting of ranked constraints, looks at whole-word surface candidates, rather than sorting the stems.

In many cases I think the OT approach works very well and is also highly principled — it relates alternations to phonotactics in the simplest possible way. But in other cases, it seems possible that the stem-sorting approach is nonetheless correct. I will give a Hungarian example here.

Hayes, Zuraw, Siptár, and Londe (2009) is a study of patterns in Hungarian vowel harmony, devoted to an issue orthogonal to those addressed here (the status of unnatural constraints in phonology). The authors examined the Hungarian vowel harmony pattern in quantitative detail, considering both the classical environments for harmony, which involve neighboring vowels, and also some surprising environments involving neighboring consonants.

Our research took what seemed to us a harmless shortcut: instead of examining the whole phonology of the language (as classical OT tells us to do), we used stem-sorting, sorting out the stems of our Hungarian lexical database into those taking front-vowel suffixes ("front-takers") and those taking back-vowel suffixes ("back-takers"). For the vowel-based constraints, it turned out not to matter whether one used classical OT or stem-sorting. For instance, any stem whose last vowel is back will be a back-taker; indeed, exceptionlessly so. This pattern could be derived either by an exceptionless principle of stem-sorting, or in classical OT with an undominated constraint requiring that back vowels be followed by back (or neutral) vowels. Where Hungarian vowel harmony becomes interesting from the present point of view is when one addresses the harder cases. These arise in what we called the **zones of variation**, consisting of the stems (about 900) that fit particular phonological descriptions, such as ending in Back + Neutral or ending in Back + Neutral + Neutral. For such stems, harmony is unpredictable, and for every stem the language learner must memorize whether it is a front-taker or a back-taker. However, there are strong tendencies within the zones of variation that render harmony semipredictable. These include the patterns whereby stem-final consonants statistically affect harmony. Specifically, there are four consonant environments, each of which favor front suffixes within the zones. These are given in (3).

(3) *The frontness-preferring consonant environments of Hungarian*

a. stem-final bilabial consonant
b. stem-final sibilant
c. stem-final coronal sonorant
d. stem-final consonant cluster

The last of these overlaps with the first three, since the last of two consonants can be any of the first three classes. The effect of these consonant environments is surprisingly large. In the zones of variation, when none of these environments is met, the data show about 1/3 back suffixes. But when two such environments are present at once, the backness frequency is close to zero. Moreover, the consonant effects are not only robust in the data, they are also apprehended by native speakers. In our nonce-probe testing (Hayes et al. 2009, §8), we found statistically significant effects for all four unnatural environments.

Let us return now to the question of stem-sorting. In the original work, we crunched the data using stem-sorting as a time-saving procedure, and skipped the step of testing the implications of the consonant environments for the language as a whole. What happens when one carries out this missing step?

First, it appears that very little changes when one considers the constraints that are based on vowels; apart from a small number of disharmonic stems and a few non-alternating suffixes, stem-sorting and classical OT lead to similar predictions.

The surprise comes with the consonant-based constraints. As it turns out, they have essentially zero validity, other than in the already-established function of predicting the behavior of stems in the zones of variation. In other words, the consonant environments are very good environments if one's goal is to predict whether a given Hungarian stem in the zones of variation is a back-taker or a front-taker. But they are not good environments at all for predicting the distribution of vowels in Hungarian in general. Below I give the two key arguments that support this conclusion.

## 5.1 Evidence for stem-sorting I: suffix behavior

A fair number of Hungarian suffixes begin with a consonant in one of the four classes of (3). An example is the dative suffix *-nak* ~ *-nek*, which begins with a coronal sonorant. However, these suffixes do not take front allomorphs more often than the other suffixes; if anything, it is the reverse. This is shown in Figure 2, which sorts the various Hungarian suffixes by consonant, and within categories by tendency to take back suffixes.

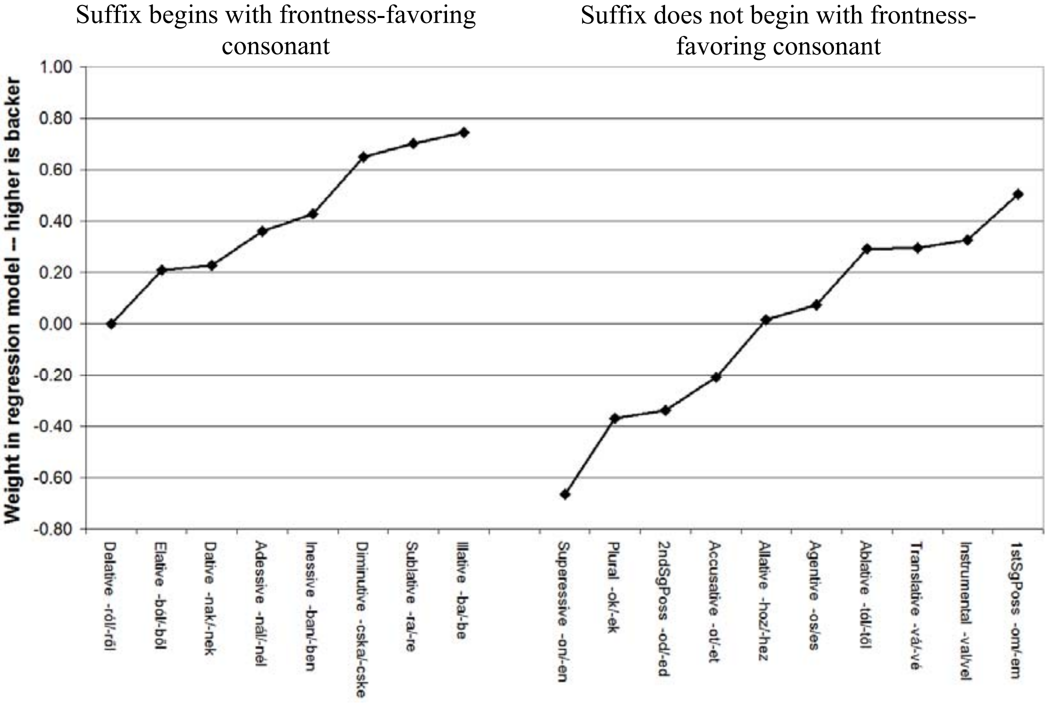


**Figure 2:** Statistical behavior of Hungarian suffixes attached to zone-of-variation stems

If the consonant environments were true across-the-board phonology, we would have expected the opposite.

### 5.2 Evidence for stem-sorting II: stem behavior

If the consonant effects seen in the zone-of-variation stems were applicable across the board, we would expect to see their effects within the stem inventory. A simple count of stem vowels in the relevant environments shows this is simply not true: after consonants that favor frontness in suffixes the percentage of front vowels is 42.4%; whereas after consonants that do not favor frontness in suffixes the percentage is 44.7%.

Summing up, it seems clear that the relevant constraints, however stated, must embody generalizations like: "Use front vowels after stems that end in bilabial consonants." They do not embody broader generalizations like "Use front vowels after bilabial consonants." Therefore, the Hungarian zones of variation could be learned well by stem-sorting, but an effort to learn them as general phonology would be stymied.

### 5.3 The scope of stem-sorting in phonology

It is hard to tell to what degree language learners have to use stem-sorting when they learn phonology, particularly since in the past most analytic work has not been on the lookout for this possibility. One recent clear case is given by Gouskova and Becker's (2013) experimental study of a (likewise stochastic) area of Russian phonology, namely the deletion of the mid vowels that historically originated in the "jer" vowels of Proto-Slavic. The authors observe that XVCVCC stems greatly resist jer drop, but XVCCVC stems do not. Ordinary markedness constraints cannot distinguish the two, since in each the Faithfulness violation is of MAX(V) and the markedness of the deleting output (XVCCC-V in each case) is identical. Stem sorting would straightforwardly identify the cluster-final stems as resistant to jer-drop.

I conjecture that to the extent that stem-sorting is used in phonological learning, it serves as a backup strategy. Where alternations are governed directly by surface-true phonotactic constraints, there would be no point in struggling to find environments that are, in effect, already known. The Hungarian and Russian examples share the property that they are neither surface true nor, indeed, predictable except at the stochastic level. Stem-sorting may well be a strategy used when the language learner faces a tough challenge for predictability, and thus must engage in problem-specific toil.

## 6 Three general questions about comparative phonotactics

### 6.1 Does comparative phonotactics solve some problems better than "absolute" phonotactics would?

To address this question, we need an alternative, which is aptly supplied in the work of Becker and Gouskova (2012): comparative phonotactics is compared absolute phonotactics. This idea works as follows; given a Population A and Population B, we learn the absolute phonotactics of Population A, and also the absolute phonotactics of Population B. Then, the probability that a form *x* belongs to A is defined as in (4):

(4) *Comparative phonotactics deduced from absolute phonotactics*

$$\frac{x\text{'s phonotactic probability construed as A}}{x\text{'s phonotactic probability as A} + x\text{'s phonotactic probability as B}}$$

This idea strikes me as intriguing but oblique — why not solve the problem as directly as possible? A reason we might actually prefer a direct solution is that any kind of phonotactic learning is vulnerable to noise, which will accumulate more, the more steps we take along the way to deriving the outcome.

I made an exploratory effort to compare the two approaches, implementing absolute phonotactics with the Hayes/Wilson 2008 computational learner, and using a truncated constraint set due to computational limitations. I found that direct comparative phonotactics did indeed achieve a lower average error, namely 0.269 as opposed to the 0.288 achieved using compared absolute phonotactics .

## 6.2 Why would it be sensible for language learners to engage in comparative phonotactics?

I think the answer to this question is simple but perhaps underappreciated: grammar is learned because it makes you a better speaker of your language. There are many areas in which it helps to make things as predictable as possible. Here are some examples.

Speakers need to determine the **part of speech** of new words that they hear, a task that is particularly difficult for children whose knowledge of syntax is still developing. In this connection, Christiansen and Monaghan (2006) and Monaghan, Christiansen and Chater (2007) have used comparative phonotactic analysis to predict part of speech in English. This topic is also pursued in *SPE*; the general question of part-of-speech phonotactics is discussed by Smith (2011).

**Grammatical gender** is surprisingly predictable on phonotactic grounds, though other factors also play a role. Scholars who have successfully performed comparative phonotactic analysis on Spanish and French data include Poplack et al. (1982), Karmiloff-Smith (1979), Lyster (2006), and Glewwe (2014). A comparative phonotactic grammar for gender permits speakers to make better guesses about gender for new words, and to better understand other people’s mistaken or dialectally-varying productions.

As noted already (§4), a comparative phonotactics for lexical strata also helps a speaker command a variety of **styles**.

Lastly, there is evidence that learning the lexical strata assists the process of **speech perception**. In a widely-adopted view, speech perception is guided by a Bayesian “forward model” that assigns prior probabilities to the possible interpretations of the signal, helping to guide the listener to the correct interpretation (see especially Norris and McQueen 2008). Moreton and Amano (1999) demonstrated that Japanese listeners use knowledge of lexical strata when they perceive vowel length. In Japanese, we find that initial [rj] and [hj] do not occur in the Native stratum, whereas long [aː] does not occur in the (learned) Sino-Japanese stratum. In Moreton and Amano’s experiment, they played nonce words like [rjotaː] vs. [potaː] ([p] not confined to Sino-Japanese), smoothly varying the length of the [aː], and calculating from the subjects’ responses the perceptual boundary between [a] and [aː]. They found that when the initial consonants are [rj], a boundary shift occurred: more phonetic length was required for the subjects to perceive phonological [aː]. The implied chain of inference is from the phonotactics of the initial clusters, to the vocabulary stratum of the word being heard, thence to the likelihood that the signal contains a long vowel.

In sum, the effort for speakers to learn various forms of comparative phonotactics would pay off in terms of improved performance in the creation of novel well-formed utterances and in speech perception.

## 6.3 What sort of grammatical architecture could accommodate comparative phonotactics?

I appeal to a distinction between **monolithic** and **atomistic** approaches to phonological grammar. Monolithic approaches achieve generality by using the same devices to cover multiple purposes. An example of work adopting such an approach is Smolensky (1996), who defends the use of a single constraint hierarchy for both production and perception; similarly, classical OT is notable for using the same apparatus to handle both phonotactics and alternations. Atomistic approaches tend to set up separate apparatus, possibly partly redundant, for different purposes. Thus Boersma (1998) proposed separate phonological grammars for production and perception; the “co-phonologies” (roughly, morphology-specific phonologies) proposed by Inkelas and colleagues (e.g., Inkelas and Zoll 2007) are likewise purpose-specific.

Comparative phonotactics, as construed here, follows the atomistic approach. The phonological world is filled with choices and according to the proposal here, separate analyses can be constructed by language learners if appropriate for making each choice.

Resolving the monolithic/atomistic debate is a major task to say the least, but I think the drift of current phonological research is giving us a clue that phonology is likely to be at least somewhat atomistic. I draw here on the strong current trend in the field to investigate phonological knowledge by experiment. This research is yielding many different results, but I think the most significant one so far is that phonologically speaking, children appear to be *virtuosi*: they are skilled extractors of highly detailed phonological patterns from the data they receive. The ability of Hungarian children to detect consonant environments for vowel harmony, mentioned above, is one example; here are two others.

Albright (2002) and Albright and Hayes’s (2003) experiments found “**islands of reliability**” for Italian conjugation class and for English past tenses respectively; i.e. phonological environments where a particular morphological outcome is preferred. These can be synchronically arbitrary and surprisingly detailed; thus for instance Albright and Hayes’s experiments suggest that English-learning children detect that all verbs ending in voiceless fricatives are regular.

Ernestus and Baayen (2003) demonstrated that Dutch speakers have the ability to **undo neutralizations** in the sense that they can use phonological regularities to project base forms from neutralized surface allomorphs. Specifically, they can use place and manner of articulation patterns to guess reliably the base forms corresponding to surface forms neutralized by Final Devoicing.

Although there are isolated and puzzling exceptions (notably Becker, Nevins, and Ketrez 2011), I think the pattern that is emerging from current research is: *when in doubt, bet on the language learner to notice things*. If the things to be noticed are linguistically meaningful and make you a better speaker of the language, our default expectation should be that children will notice them.

Both the island-of-reliability effect and the undoing-of-neutralization effect indicate detailed phonological knowledge that has historically been missed in monolithic approaches, perhaps because they go beyond what is needed simply to describe the existing body of data. To me, the recent findings suggest that children are voracious and opportunistic in their pattern learning; an atomistic approach to theory is made more plausible by the need for theory to recognize such voracity/opportunism. Regarding comparative phonotactics, I've suggested that the binary distinctions that are made by comparative phonotactic grammars are indeed useful, and unlikely to be as effectively learned by other means.

Lastly, this is not a call for recklessly or unnecessarily atomistic grammars. Where monolithicism gets us important results for free (as in OT's derivation of automatic alternation from exceptionless absolute phonotactics) there is no sensible basis for abandoning it.

## References

Albright, Adam. 2002. Islands of reliability for regular morphology: evidence from Italian. *Language* 78.684–709.

Allen, W. Sidney. 1978. *Vox Latina*. Cambridge: Cambridge University Press.

Baayen, Harald, Richard Piepenbrock, & Leon Gulikers. 1995. *The CELEX Lexical Database*. Release 2 (CD-ROM).

Baković, Eric. 2011. Opacity and ordering. In *The Handbook of Phonological* Theory, 2nd ed., ed. by John Goldsmith, Jason Riggle, & Alan C. L. Yu. Oxford: Wiley-Blackwell.

Becker, Michael. 2009. Phonological trends in the lexicon: The role of constraints. Ph.D. dissertation. Amherst: University of Massachusetts.

Becker, Michael & Maria Gouskova. 2012. Source-oriented generalizations as grammar inference in Russian vowel deletion. Ms., SUNY Stony Brook and NYU.

Becker, Michael, Nihan Ketrez, & Andrew Nevins. 2011. The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language* 87.84–125.

Berger, Adam L., Stephen A. Della Pietra, & Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22.39–71.

Boersma, Paul. 1998. *Functional phonology*. The Hague: Holland Academic Graphics.

Bybee, Joan & Carol Lynn Moder. 1983. Morphological classes as natural categories. *Language* 59.251–270.

Cedergren, Henrietta & David Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language*, 50.333–355.

Chomsky, Noam, & Morris Halle. 1965. Some controversial questions in phonological theory. *Journal of Linguistics* 1.97–138.

Chomsky, Noam & Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper and Row.

Christiansen, Morten, & Monaghan, Padraic. 2006. Discovering verbs through multiple cue integration. In *Action Meets Word: How Children Learn Verbs*, ed. by K. Hirsh-Pasek & R. M. Golinkoff, 88–110. New York: Oxford University Press.

Ernestus, Mirjam, & R. Harald Baayen. 2003. Predicting the unpredictable: interpreting neutralized segments in Dutch. *Language* 79.5–38.

Glewwe, Eleanor. 2014. Developing a phonological/morphological model to predict the gender of French nouns. Ms., Dept. of Linguistics, UCLA.

Goldwater, Sharon, & Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, ed. by Jennifer Spenader; Anders Eriksson, & Osten Dahl, 111–120. Stockholm: Stockholm University Department of Linguistics.

Gorman, Kyle & Daniel Ezra Johnson. 2013. Quantitative analysis. *The Oxford Handbook of Sociolinguistics*, ed. by Robert Bayley, Richard Cameron, & Ceil Lucas, 214–240.

Gouskova, Maria & Michael Becker. 2013. Nonce words show that Russian yer alternations are governed by the grammar. *Natural Language and Linguistic Theory* 31.735–765.

Gropen, Jess, Steven Pinker, Michelle Hollander, Richard Goldberg & Ronald Wilson. 1989. The learnability and acquisition of the dative alternation in English. *Language* 65.203–257.

Gussenhoven, Carlos. 1986. English plosive allophones and ambisyllabicity. *Gramma* 10.119–141.

Hayes, Bruce & James White. 2013. Phonological naturalness and phonotactic learning. *Linguistic Inquiry* 44.45–75

Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39.379–440.

Hayes, Bruce, Kie Zuraw, Péter Siptár, & Zsuzsa Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85.822–863.

Inkelas, Sharon & Cheryl Zoll. 2007. Is grammar dependence real? *Linguistics* 45.133–171.

Itô, Junko & Armin Mester. 1995. Japanese phonology. In *The Handbook of Phonological Theory*, ed. by John Goldsmith, 817–838. Oxford: Blackwell.

Jäger, Gerhard & Anette Rosenbach. 2006. The winner takes it all – almost: cumulativity in grammatical variation. *Linguistics* 44.937–971.

Jespersen, Otto. 1909. *A Modern English Grammar on Historical Principles. Part I: Sounds and spellings*. London: George Allen & Unwin.

Kahn, Daniel. 1976. *Syllable-Based Generalizations in English Phonology*. Ph.D. dissertation, MIT. Published 1980, New York: Garland.

Karmiloff-Smith, A. 1979. Production experiments: gender-indicating function of determiners. In *A Functional Approach to Child Language*, 148–169. Cambridge: Cambridge University Press.

Kawahara, Shigeto, Kohei Nishimura, & Hajime Ono. 2005. Unveiling the unmarkedness of Sino-Japanese. In *Japanese/Korean Linguistics* 12, ed. by William McClure. Stanford: CSLI.

Legendre, Géraldine, Yoshiro Miyata, & Paul Smolensky. 1990. Harmonic grammar: A formal multi-level connectionist theory of linguistic well-formedness: an application. In *COGSCI 1990*, 884–891.

Lyster, R. 2006. Predictability in French gender attribution: A corpus analysis. *Journal of French Language Studies*, 16.69–92.

Marchand, Hans. 1969. *The Categories and Types of Present-Day English Word-Formation*. 2nd edition. Munich: Verlag C. H. Beck.

McCawley, James. 1968. *The Phonological Component of a Grammar of Japanese*. The Hague: Mouton.

Monaghan, Padraic, Morten H. Christiansen, & Nick Chater. 2007. The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology* 55.259–305

Moreton, Elliott, & Shigeaki Amano. 1999. Phonotactics in the perception of Japanese vowel length: Evidence for long-distance dependencies. In *Proceedings of the 6th European Conference on Speech Communication and Technology*, Budapest.

Norris, Dennis & James M. McQueen. 2008. Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review* 115.357–395.

*OED* = Oxford English Dictionary, accessed on line at `www.oed.com`.

Plag, Ingo. 2003. *Word-formation in English*. Cambridge: Cambridge University Press.

Poplack, Shana, Alicia Pousada, & David Sankoff. 1982. Competing influences on gender assignment: variable process, stable outcome. *Lingua* 57.1–28.

Prince, Alan & Paul Smolensky. 1993/2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Technical report, Rutgers University Center for Cognitive Science. [Published 2004; Oxford: Blackwell]

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `www.R-project.org`.

Roy, Deb, Rupal Patel, Philip DeCamp, Rony Kubat, Michael Fleischman, Brandon Roy, Nikolaos Mavridis, Stefanie Tellex, Alexia Salata, Jethran Guinness, Michael Levit, Peter Gorniak. 2006. The Human Speechome Project. Paper given at the 28th Annual Conference of the Cognitive Science Society. `http://media.mit.edu/cogmac/publications/ cogsci06.pdf`.

Scholes, Robert J. 1966. *Phonotactic Grammaticality*. The Hague: Mouton.

Smith, Jennifer. 2011. Category-specific effects. In *The Blackwell Companion to Phonology*, ed. by Marc van Oostendorp, Colin Ewen, Beth Hume, & Keren Rice, 2439–2463. Malden, MA: Wiley-Blackwell.

Smolensky, Paul (1986) Information processing in dynamical systems: Foundations of Harmony Theory. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models*, ed. by James L. McClelland, David E. Rumelhart and the PDP Research Group, 390-431. Cambridge, MA: MIT Press..

Smolensky, Paul 1996. On the comprehension/production dilemma in child language. *Linguistic Inquiry* 27.720–731.

Sturtevant, Edgar. 1920. *The Pronunciation of Greek and Latin*. Chicago: University of Chicago Press.

Wells, John. 1982. *Accents of English*. Cambridge: Cambridge University Press.

Wilson, Colin 2006. Learning phonology with substantive bias: an experimental and computational investigation of velar palatalization. *Cognitive Science* 30.945–982