Phonological markedness effects in sentence formation*

Canaan Breiss Bruce Hayes

UCLA

To appear in *Language*

Pre-copying editing version, December 2019

Abstract

Earlier research has found that phonological markedness constraints (for example, against stress clash or sibilant sequences) statistically influence speakers' choices between particular syntactic constructions and between synonymous words. In this study, we test phonological constraints not just in particular cases, but across the board. We employ a novel method that statistically models the distribution of WORD BIGRAMS (consecutive two-word sequences) and how this distribution is influenced by phonological constraints. Our study of multiple corpora indicates that several phonological constraints do indeed play a statistically significant role in English sentence formation. We also show that by examining particular subsets of the corpora we can diagnose the mechanisms whereby phonologically marked sequences come to be underrepresented. We conclude by discussing modes of grammatical organization compatible with our findings.

Keywords: markedness, phrasal phonology, syntax-phonology interface, grammatical architectures

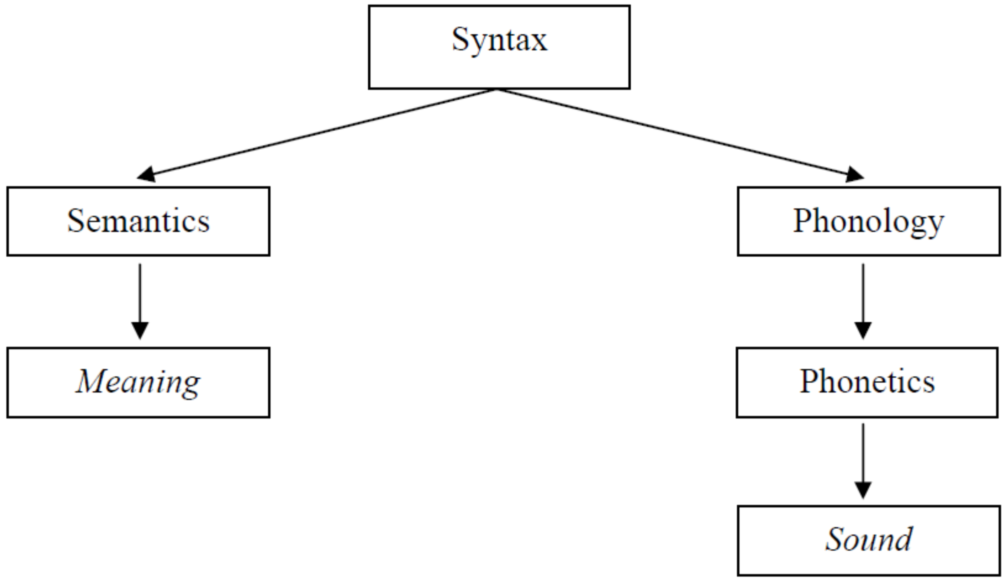
*We would like to thank Joan Bresnan, Tim Hunter, Paul Kiparsky, Omer Preminger, and Stephanie Shih; Siavash Jalal of the UCLA statistical consulting unit; the reviewers for Language; and talk audiences at UCLA, Seoul National University, Cornell University, the Linguistic Society of America, the University of Arizona, and the University of Southern California for their advice and practical assistance. All responsibility for remaining errors rests with us. This work was supported by an NSF Graduate Research Fellowship to Breiss and by the UCLA Academic Senate.

The Supplemental Materials for this article may be obtained from the authors' websites.

1. Goals and setting

The focus of this article is how different domains of linguistic knowledge — here, syntactic, phonological, and lexical — interact in the creation of sentences. An influential proposal for how these domains interact is the componential feed-forward model, laid out in Chomsky (1965:16) and shown in Figure 1. This model places the syntactic component at the core of the grammar, with its output transmitted to two interpretive components, which derive output representations for meaning and sound.

Figure 1: A schematic depiction of the feed-forward model



Zwicky and Pullum (1986:63) offered a way to interpret such diagrams so as to make a clear empirical prediction: "the syntactic component determines the order in which words may be placed in sentences … and the phonological component determines what pronunciations are associated with particular structured sequences of words that the syntax says are well-formed." In other words, syntax is generative, phonology interpretive; the phonology freely accepts whatever the syntax chooses to give it, and forms a pronunciation. It is this interpretation of Figure 1 that we address here.

In the intervening years, Zwicky and Pullum's interpretation has been met with an ever-growing body of empirical challenges. For instance, the syntax of a language often offers more than one way to express a given meaning, and it can be shown that speakers' choices between alternative constructions are sometimes phonologically motivated. Shih and Zuraw (2017, 2018) studied two parallel noun-adjective constructions in Tagalog: {Adj. *linker* Noun} and {Noun *linker* Adj.}, where *linker* is one of two contextually determined allomorphs *na* or -*ng*. Using corpus data, they showed that speakers tend to choose between these options in ways that avoid violating certain phonological markedness constraints: *[+nasal][+nasal], *HIATUS, and *NC̥, all of which are active elsewhere in the language's phonology. Similar instances of phonologically-biased syntactic choice have been detected for the English dative alternation (*give X to* Y vs. *give Y X*; Anttila et al. 2010, Shih and Graffmiller 2011, Shih 2017a), the genitive alternation (*X's Y*, *Y of X*; Shih et al. 2015, Ryan 2018), and the conjunct-order alternation (*X and Y, Y and X*; Benor and Levy 2006, Shih 2017a, Ryan 2018). The Sanskrit Rigveda takes advantage of free word order to avoid hiatus (Gunkel and Ryan 2011).

Other forms of evidence have been put forth. For instance, the linear ordering of clitics is argued to depend in part on phonological factors (Zec and Inkelas 1990, Chung 2003, Erteschik-Shir et al. 2019). The notion of weight, governing Heavy NP Shift, Topicalization, and other aspects of word order, has been argued to be at least partly phonologically defined (Zec and Inkelas, Shih and Grafmiller 2011, Ryan 2018). Along with Zec and Inkelas, Agbayani and Golston have argued for movement operations (for instance, Japanese long-distance scrambling) that apply to phonological, not syntactic, constituents (Agbayani and Golston 2010, 2016; Agbayani et al. 2015).

To accommodate such phenomena, some formal models bifurcate the syntax, establishing a "narrow" core which is phonology-free but performs only a subset of the work of actually arranging words into sentences, leaving other aspects of linearization to separate mechanisms. Such approaches would include the work of Agbayani and colleagues just cited, as well as Embick and Noyer (2001). Our data do not bear, as far as we can tell, on the validity of such models. However, the approach of isolating a phonology-free core of the syntax does have the potential to obscure what is meant by "phonological effects in syntax," ruling out their existence more or less by definition (Anttila 2016:130). To keep our purpose clear, we use the phrase "sentence formation," designating whatever grammatical apparatus is employed, in any framework, to derive complete, observable sentences.

"Sentence formation" would also include word choice, which also appears to be guided by phonological constraints. Schlüter (2005, 2015) has demonstrated that, given a synonym pair, speakers tend to select the word that creates fewer phonological constraint violations in its local syntactic context. Her evidence comes from the distribution of historical English lexical doublets like *worse* and *worser*, which are gradiently deployed to avoid violations of *CLASH (stress on adjacent syllables). Schlüter and Knappe (2018) have obtained similar results for modern synonym pairs such as *glad*/*happy*.

A characteristic of the cases just cited is that the phonology enforces a GRADIENT preference among competing variants, a property we will see below in our own data. Yet the pattern is not always gradient: analysts have suggested cases in which utterances are fully ungrammatical due to violation of a phonological constraint. One such case is provided by Rice and Svenonius (1998) and Rice (2007), who observe that in some varieties of Norwegian, there exist imperative verb forms that end in consonant clusters with reversed sonority, such as *sykl* 'bicycle-IMP.' or *åpn* 'open-IMP.' The phonology of these dialects does not permit reversed-sonority syllable codas, and as a result these verbs cannot be uttered prepausally, nor in a sentence before a consonant-initial word. But before a vowel-initial word in the same sentence, the sonority problem is repaired by resyllabification across the word boundary (that is, *sy.kl#V*), and the imperative becomes useable. A representative contrast (Rice 2007:204) is given in (1).

(1) *Reversed-sonority imperatives in Norwegian*

a. Sykl opp bakken
bike up the.hill
'Bike up the hill!'

b. *Sykl ned bakken
bike down the.hill
'Bike down the hill!'

It would seem that (1b) is a sentence that is ungrammatical for phonological reasons. Further cases of this sort are proposed by Anttila (2016), Zec and Inkelas (1990), Harford and Demuth (1999), and Agbayani et al. (2015).

The literature that critically addresses the feed-forward model appears to be both vast and fragmented among distinct research communities (Anttila 2016:133). The absence of a consensus theoretical framework for such work has surely contributed to this fragmentation. For further access to this literature we recommend the surveys (incorporating specific proposals) by Anttila (2016), Shih (2017a) and Ryan (2018).

We address the question of theoretical framework at the end of this article. Our main purpose, however, is empirical: we seek to expand the set of relevant phenomena and offer a novel research method. In previous research, it has been the norm to choose some specific area to investigate; e.g. a particular choice between competing syntactic constructions or lexical items. This approach is sensible, since it offers closely controlled comparisons. However, we feel it may be appropriate to complement this work with a broader approach. For our project, we have devised a means of testing the entire content of a body of text, examining the effects of multiple phonological constraints at the same time. Our method, outlined below, involves analyzing the complete set of WORD BIGRAMS (sequences of two consecutive words) from a text, using a statistical test that assesses the degree to which a variety of phonological constraints are respected in the creation of these bigrams.

What are the advantages of such a broad approach? First, the simplicity of our method means it is widely applicable, permitting us to test essentially any phonological constraint whose violations can arise across word boundaries. Second, since the method is not tied to particular words or constructions, it can be applied to very large quantities of text, which can provide it with the statistical power needed to detect effects too subtle to be found by other means. Lastly, we demonstrate that our method is sufficiently flexible that it can be used to go beyond merely revealing the existence of phonological effects in sentence formation, and make a start at diagnosing the mechanisms by which these effects arise.

Applying our method, we confirm earlier results in finding pervasive effects of phonological optimization: in both speech and writing, sentences gradiently respect phonological markedness constraints. Our findings also suggest that this pattern is the result of multiple causes. A large fraction of the phonological effects are due, as earlier literature suggests, to choices made between competing syntactic constructions. But we attribute another large portion to a different source: extending ideas of Martin (2011), we suggest there is a preference for lexical listing of phonologically-unmarked fixed phrases. In our final section, we discuss implications of the work, including the forms of grammatical architecture that are compatible with our findings.

2. Constraints studied

As noted above, we treat a text as a sequence of word bigrams; that is, overlapping pairs of consecutive words. For example, in the preceding sentence the first three word bigrams are [*as noted*], [*noted above*], and [*above we*]. When a speaker or writer concatenates two words to form a bigram, there is a possibility that a phonological constraint violation will be created at the juncture; for example, if *book* is concatenated with *concludes*, this creates a violation of the phonological constraint *GEMINATE (see below), as shown: [bʊk kənkludz].

Using English data, we examined violations of nine phonological constraints. These were chosen by two criteria. First, we favored constraints that are only seldom violated within the confines of a word. We also took into account the typological status of constraints, relying on the research literature. We judged that by following these criteria we would have the best chance of locating constraints that would have an effect at the phrasal level. To test for word-internal constraint strength we employed the UCLA Phonotactic Learner (Hayes and Wilson 2008), taking a constraint to be strong if this algorithm assigned it a substantial weight. The weights we obtained in this word-internal modeling may be viewed in the online Supplemental Materials for this article.

2.1 STRESS CLASH CONSTRAINTS. The constraint *CLASH, which forbids adjacent stressed syllables, plays a major role in the analysis of stress. It is well supported for English, where for instance it accounts for the pattern of stress retention in “cyclic” stress patterns. Thus, *assimilation* [əˌsɪməˈleɪʃən] retains a secondary stress on its second syllable inherited from *assimilate* [əˈsɪməleɪt], but *provocation* [ˌpɹɑvəˈkeɪʃən] fails to retain the stress of *provoke* [pɹəˈvoʊk], since it would clash with the main stress on the penult. For a detailed analysis, see Pater (2000). The effects of *CLASH on sentence formation have already been documented in the English corpus study of Temperley (2009), and our results will be seen to confirm his findings.

The more specific constraint *IAMBIC CLASH is violated when a rising sequence of stress within a word[2] immediately precedes a still stronger stress, as in the phrase *maroon sweater* ([məˌrun ˈswɛɾɚ]). Exceptions to *IAMBIC CLASH within words in English are very rare, and the relevant pronunciations are often not shared by all speakers. For instance, one of the authors of this article says *electronic* [əˌlɛkˈtɹɑnɪk] with an iambic clash, thus retaining the base stress pattern of *electron* [əˈlɛkˌtɹɑn]; the other author says *electronic* [ˌilɛkˈtɹɑnɪk], with the iambic clash repaired. Within phrases, the well-known “Rhythm Rule” (Liberman and Prince 1977 et seq.) removes *IAMBIC CLASH violations, as in *unkind* [ʌnˈkaɪnd], but *unkind people* [ˌʌnkaɪnd ˈpipəl]. The research literature on *IAMBIC CLASH has an unusual time depth; see Fijn van Draat (1910), Bolinger (1965). The phrasal effects of *IAMBIC CLASH have also been studied with more modern methods (statistical analysis of a digital corpus) by Hammond (2016); his results largely match what we describe below.

[2] A reviewer points out that iambic clashes also arise in three-word sequences, like *in tàll trées*. This is true, but we save the investigation of these cases for future research; we expect the degree of avoidance for them would be weaker, and we propose to investigate the clearest cases first.

2.2 BANS ON LONG CLUSTERS. Long consonant clusters are well known to be phonologically marked in general; and for English our word-internal modeling indicates they are underrepresented in the lexicon. For ease of assessment, in our constraints we did not attempt to include syllable structure, but simply set up two linear constraints, *TRIPLE CONSONANT CLUSTER as well a more specific version, *TRIPLE OBSTRUENT CLUSTER, that we expected to be the stronger of the two. For our purposes these constraints assess violations only for sequences of the form *C#CC or *CC#C, since *#CCC and *CCC# have no independent bearing on bigram formation.

2.3 *SIBILANT CLASH. Like many languages, English avoids consecutive sibilants within words; there are absolutely no monomorphemes like “*kessha*” *[kɛsʃə]; and even affixed forms like *misshapen* [mɪsʃeɪpən] are rather unusual. Our analysis thus includes the constraint *SIBILANT CLASH, which we state formally as $*[+\text{strident}]\begin{bmatrix}+\text{strident}\\+\text{continuant}\end{bmatrix}$. Note that this constraint is formulated so as not to be violated by sequences in which the second member of the cluster is an affricate; this is appropriate because in fact English has many words of this type, such as *question* [ˈkwɛst͡ʃən].

2.4 *GEMINATE. English strictly avoids geminates (identical consonant sequences) within monomorphemes; for example, to pronounce the Italian word *latte* ‘milk’ with a geminate [tt], as in the original, would be inconceivable for English speakers. Only a few affixed forms, such as *unknown* [ʌnnoʊn], include geminates. Importantly, Martin (2011) demonstrated that geminates are underrepresented even in English compounds; cases like *bookkeeper* do exist, but they are rare with respect to a statistical baseline. Martin’s explanation for this will be adopted and extended below.

2.5 *HIATUS. This constraint is violated by consecutive vowels. *HIATUS (or its near-equivalent *NO ONSET; McCarthy and Prince 1993:34-37) has a substantial pedigree in Optimality Theory and plays a role in the phonology of many languages. Vowel sequences do occur within English words (e.g. in *media* [ˈmidiə]) but nonetheless in the modeling work mentioned above, *HIATUS violations emerge as statistically underrepresented.

2.6 *BAD SONORITY. The Syllable Contact Law (Hooper 1976, Murray and Vennemann 1983), militates against heterosyllabic clusters to the extent that their initial coda consonant has lower sonority then the following onset consonant; in English, any CC cluster formed across word boundaries will consist of a coda followed by an onset and thus fall under the scope of this Law. We set up the constraint *BAD SONORITY, whose violations are computed by subtracting the sonority of the first of two consonants from the sonority of the second, on the scale *obstruent* - *nasal* - *liquid* - *glide* (cf. Clements 1990). Our modeling indicates that violations within words are moderately underrepresented.

2.7 *NC̥. This constraint bans voiceless consonants after nasals. It is not enforced within English words (in our word-internal testing, we found no effect at all), but its typological pedigree (Pater 1999) led us to test its applicability at the phrasal level. In the hope of avoiding statistical confounds from the effect of other constraints, we adopt a very narrowly defined version of *NC̥, banning only homorganic *nasal* + *stop* sequences; that is, [mp, nt, ŋk].

Summing up, (2) gives the full list of constraints we tested.

(2) *List of phonological constraints tested*

a. *CLASH
b. *IAMBIC CLASH
c. *TRIPLE OBSTRUENT CLUSTER
d. *TRIPLE CONSONANT CLUSTER
e. *SIBILANT CLASH
f. *GEMINATE
g. *HIATUS
h. *BAD SONORITY
i. *NC̥

3. Word bigrams as a basis for detecting the effects of phonological constraints

To test these constraints, a method is needed that can digest all the word bigrams of a text corpus and determine whether they collectively underrepresent phonological constraint violations. To start, suppose that the corpus under examination consists of the six canonical novels of Jane Austen. For present purposes, we adopt the (philistine) idealization of Austen as a stochastic device that emits word bigrams; thus we seek to model the frequency with which each bigram is emitted. The Austen corpus is 723,214 words long and so (ignoring some trimming back to be carried out below) there are 723,213 bigrams.

A sensible starting point is to assume that each word is emitted with a probability matching its corpus frequency. For instance, in the corpus, *Elizabeth* occurs 454 times, and hence is emitted with a probability of 454/725,374 = 0.00063. *Bennet* occurs 291 times, thus with a probability of 291/725,374 = 0.0004. If one knows the probabilities of both words in a bigram, then the probability of that bigram can be computed as their product; so that the expected rate for the bigram *Elizabeth Bennet* in the Austen corpus is 454/723,214 × 291/725,374 = 0.00000025, corresponding to a predicted text frequency of 0.18 occurrences.

The observed frequency of a particular bigram will only seldom be identical to the expected value as just computed. For instance, the bigram *Elizabeth Bennet* actually occurs 6 times in the Austen corpus, 33 times the expected frequency. This is hardly surprising, since *Elizabeth Bennet* is the full name of the heroine of Austen's novel *Pride and Prejudice*. But by considering AGGREGATED data, it is possible to abstract away from such factors and hope to find broader principles governing the bigram frequencies. In particular, we focus on possible effects of the phonological constraints given earlier: our estimate of the degree to which Jane Austen respects the phonological constraints when she emits bigrams will be based on the degree of improvement in our ability to predict her bigram frequencies that is obtained when we include the phonological constraints in a statistical model. Below, we discuss how we worked out this scheme in concrete terms.

4. Multinomial logistic regression, a.k.a MaxEnt

Our method of testing is MULTINOMIAL LOGISTIC REGRESSION (Jurafsky and Martin 2019, ch. 5), a statistical method that assigns probabilities to categorical outcomes according to properties possessed by these outcomes. For us, the outcomes in question are the individual bigrams that Jane Austen might emit given her active vocabulary, and the properties in question include the violations by these bigrams of phonological constraints.

Within linguistics, multinomial logistic regression has more than one role and more than one name. As a model of statistics, it (or its binary variant) is widely used in corpus and experimental work. However, under a different name —MAXIMUM ENTROPY (MaxEnt)— it has been employed in recent years as a theory of grammar. More specifically, MaxEnt can be deployed as a probabilistic version of Optimality Theory (Prince and Smolensky 1993/2004).[3] Here, the "outcomes" just mentioned are candidates (taken from GEN), and the "properties" are their constraint violations. Early work using MaxEnt as a theoretical model includes Smolensky (1986), Goldwater and Johnson (2003), Wilson (2006), and Hayes and Wilson (2008).

In this article, MaxEnt will play both of the roles just mentioned. The main part of the article will use MaxEnt in its guise as multinomial logistic regression for statistical testing. For this purpose, we are examining the native speaker's output simply to detect regularities, without aspiring to model the grammatical knowledge that underlies these regularities. Later on (§9.2), we turn to the question of how the observed behaviors relate to the native speaker's internalized knowledge of language, and here MaxEnt as a linguistic theory will play a role.

Terminology: for brevity, we generally use "MaxEnt," not "multinomial logistic regression". The intended application — statistical inference or linguistic theorizing — will be clear from the context. We will also use familiar ideas from Optimality Theory in presenting the math, looking forward to the theoretical exposition in §9.2.

4.1 Summary of MaxEnt

MaxEnt presupposes a set of choices, for instance, the candidate set GEN of Optimality Theory. As a probabilistic model, MaxEnt assigns each candidate a probability, rather than picking a single winner, as in classical OT. Usually, the candidates to which probabilities are assigned are the outputs that could be derived from a particular input. Another possibility, however, is to use MaxEnt as a theory of well-formedness: we equate GEN with a universal set, such as all possible phoneme sequences (as in Hayes and Wilson 2008, studying phonotactics), or all possible iambic pentameter lines (as in Hayes, Wilson, and Shisko 2012, studying metrics). In this latter approach, nothing is "derived from" anything else, and the probability assigned to a candidate is a simply a measure of its well-formedness. We will adopt the universal-set approach here, assigning probabilities to every possible word bigram given the author's vocabulary.

In MaxEnt, constraints are not ranked, as in OT, but given numerical WEIGHTS; constraints with higher weights are, intuitively speaking, stronger. These weights are employed in the core MaxEnt formula (3), which takes as input the set of constraints, constraint weights, candidates, and constraint violations, and outputs a probability Pr (*x*) for each candidate *x*.

[3] Another version of stochastic Optimality Theory, namely Noisy Harmonic Grammar (Pater 2016, Boersma and Pater 2016), might in principle have served our purpose, but we find that at least with our own software MaxEnt gives more accurate results.

(3) *The MaxEnt formula*

$$\Pr(x) = \frac{\exp(-\Sigma_i\, w_i \mathrm{f}_i\,(x))}{Z}\text{, where } Z = \Sigma_j\ \exp(-\Sigma_i\, w_i \mathrm{f}_i\,(x_j))$$

The formula says that to calculate this probability, one must do the things in Table 1 in order.

Table 1: The MaxEnt calculations for a given candidate x

Compute this	*Name of what is computed*	*How and why it is computed*
1. $\Sigma_i\, w_i \mathrm{f}_i\,(x)$	Harmony (Smolensky 1986)	Multiply x's violation counts for each constraint (designated $\mathbf{f}_{\boldsymbol{i}}\,\mathbf{(}\boldsymbol{x_j}\mathbf{)}$) by the weight of the constraint ($\boldsymbol{w_i}$), then add up the results across all constraints ($\boldsymbol{\Sigma_i}$). *All available evidence (i.e. constraint violations) bearing on a candidate is considered, in proportion to the constraint weights.*[4]
2. $\exp(-\Sigma_i\, w_i \mathrm{f}_i\,(x))$	eHarmony (Wilson 2014)[5]	Negate the harmony of x, then compute the function **exp( )** on the result, where $\exp(x)$ is a typographic convenience for e^x, $e \approx 2.72$. *In a series of candidates with ever greater harmony penalties, the probabilities descend not in linear fashion, but instead asymptote to zero (negative exponential curve) — certainty is evidentially expensive.*
3. $\Sigma_j \exp(-\Sigma_i\, w_i \mathrm{f}_i\,(x_j))$	Z, the "normalizing constant"	Compute the eHarmony of every candidate derived from the same input as x (x included), and sum these values.
4. $\dfrac{\exp(-\Sigma_i\, w_i \mathrm{f}_i\,(x))}{Z}$	Probability of x	Divide the eHarmony of x by Z (and similarly for all other candidates). *The probability of a candidate depends inversely on the probability of the candidates with which it competes (probability of all candidates must sum to one).*

[4] This is not so for Optimality Theory, where decisions between candidates are made by the highest-ranking constraint that that distinguishes them, and all the evidence from other constraints is ignored.

[5] Wilson was joking in inventing this name (which also denotes a dating web site), but we feel it is helpful as a mnemonic.

In what follows, the most essential aspect of MaxEnt is that the constraint weights have a consistent and intuitive interpretation: the higher the weight, the lower the probability of candidates that violate it (for the exact relationship, see §8.4 below). In this context, the weight of a constraint is an appropriate measure of its role in determining the speaker's inventory of bigram outputs.

4.2 Computing the weights and statistical testing

The other half of the MaxEnt approach is a method for fitting the constraint weights to match the data accurately, described for instance in Hayes and Wilson (2008:385-389). An attractive characteristic of MaxEnt is that it comes with a guarantee, in the form of mathematical proof (Della Pietra et al. 1997), that appropriate searching will always converge on the best-fitting weights. Employed in their statistical guise as multinomial logistic regression, MaxEnt systems also avoid false (artifactual) results delivered by other methods of data interpretation, as demonstrated by Wilson and Obdeyn (2009) and the Appendix to this article. MaxEnt systems likewise can provide the basis for rigorous statistical testing of the hypotheses embodied in the constraints; this testing is described in §5.4 below. The testing will be crucial to assess our hypothesis that speakers are influenced by phonological constraint violations when they form sentences.

5. Implementation

5.1 Delimiting GEN

The GEN function provides the candidates across which our models will be distributing probabilities. What sort of GEN would be appropriate to our ends? The simplest approach would be to let GEN consist of all possible bigrams that can be formed from the words in the corpus. However, in practice this proves to be difficult: for Jane Austen, who used about 14,000 unique words in the works analyzed here, this would result in a candidate set of about 200 million items, making it computationally infeasible, at least with the resources we command.

However, it is possible to simplify the calculations. The key is to observe that the actual identity of the words in a bigram doesn't matter except insofar as their phonological properties lead to violations of the phonological bigram constraints we are testing. For example, all that is really relevant for us about the word *kiss* [kɪs] is that it ends with a sibilant (and so will form violations of *SIBILANT CLASH when the next word begins with a sibilant), that it bears final stress (and so will be involved in violations of *CLASH when the next word is initially stressed), that it begins with a [k] (and so will form violations of *GEMINATE when the preceding word ends in [k]); and so on. Because of this, it is feasible to pool the roughly 200 million bigram candidates into classes which are defined solely by those phonological characteristics that bear on the constraint inventory in (2). Under this approach, the tableaux will include frequency values, namely the number of actual bigrams in the corpus that fall into each pooled category.

The resulting tableau is far smaller (38,016 candidates[6]), but the weights obtained from it are unchanged.[7]

In order to pool the candidates into classes, we adopt a set of what we will call CONSTRAINTS OF CONVENIENCE (Table 2) which embody the phonological properties of words that determine whether bigrams formed from them will violate the test constraints described earlier in §2.

Table 2*: Constraints of convenience*

Constraint	**Defined on**	**Used in assessing**
FINAL STRESS	Word 1	*CLASH
IAMBIC STRESS	Word 1	*IAMBIC CLASH
FINAL [−son][−son]	Word 1	*TRIPLE OBSTRUENT CLUSTER
FINAL CC	Word 1	*TRIPLE CONSONANT CLUSTER
FINAL VOWEL	Word 1	*HIATUS
{ FINAL C } (21 separate constraints; one for every final consonant)	Word 1	*SIBILANT CLASH, *GEMINATE, *BAD SONORITY, *TRIPLE OBSTRUENT CLUSTER, *TRIPLE CONSONANT CLUSTER, *NC̥
INITIAL STRESS	Word 2	*CLASH, *IAMBIC CLASH
INITIAL [−son][−son]	Word 2	*TRIPLE OBSTRUENT CLUSTER
INITIAL CC	Word 2	*TRIPLE CONSONANT CLUSTER
INITIAL VOWEL	Word 2	*HIATUS
{ INITIAL C } (23 separate constraints; one for every initial consonant)	Word 2	*SIBILANT CLASH, *GEMINATE, *BAD SONORITY, *TRIPLE OBSTRUENT CLUSTER, *TRIPLE CONSONANT CLUSTER, *NC̥

Once the constraints of convenience have received their proper weights in the MaxEnt analysis, they will form a baseline model of the phonological composition of the corpus, reflecting the phonological characteristics of the words available in the lexicon and the overall frequencies with which these words are used.

Here is an example of how our procedure would be applied to the Jane Austen corpus. The corpus includes the bigram *exact plan* [əgˌzækt ˈplæn], which falls under the scope of the constraints of convenience FINAL STRESS, IAMBIC STRESS, FINAL CC, FINAL [−son][−son], FINAL [t] (all for Word 1); and INITIAL STRESS, INITIAL CC, and INITIAL [p] (for Word2). *Exact plan* thus is part of bigram category that, it turns out, includes precisely two other members: *unjust praise* and *distrust providence*. We therefore install in our tableau the frequency value 3 for the

[6] Here are the details: 2 penult stress levels × 2 final stress levels × 3 levels of sonority in the penult C × 22 possible final segments × 24 possible initial segments × 3 levels of peninitial sonority × 2 levels of initial stress. In the calculation of final and initial segments, we collapsed vowels to a single category. A sample spreadsheet for our data analysis may be viewed in the Supplemental Materials.

[7] A proof of this assertion can be obtained from the fact that standard search algorithms for MaxEnt weights upwardly follow the gradient on the hill of log likelihood, and that this gradient is equal to the value *Observed − Expected* for the *pooled* violation counts of each constraint (see e.g. Hayes and Wilson 2008:388-389). A simple example illustrating the feasibility of our pooling procedure appears in the Supplemental Materials.

abstract candidate class that includes these bigrams. We similarly classify all of the bigrams in the corpus, using custom software,[8] which outputs tableaux in the form of a spreadsheet. These tableaux have 38,016 rows of candidates, and there are 53 columns of violations (corresponding to the total number of constraints of convenience) for baseline modeling, plus 9 additional columns (per §2) when we are testing the effect of phonological constraints. A sample spreadsheet file may be viewed in the Supplemental Materials.

5.2 How the system works: an intuitive example

Here is a simplified illustrative example. In our Austen corpus ("Simple" version; §8.4), the constraint of convenience FINAL VOWEL (applicable to Word 1 of a bigram) is violated 198,168 times out of 570,819 bigrams total, a ratio of .347. The constraint of convenience INITIAL VOWEL is violated 138,239 times, or .242. From this, we can informally compute the baseline probability of a *HIATUS violation, namely .347 × .242 = .084, which would correspond to 47,991 *HIATUS violations. In fact, there are only 39,973 such violations; the reduction is by a factor of 39,973/47,991 = .833. This represents the additional penalty, we suggest, arising from hiatus. Moreover, when we look at the weight obtained for *HIATUS in §8.4 below, namely .257, and interpret it with the MaxEnt math (as illustrated in (7)), we find it predicts a reduction in probability of .773, not far off from .833. The difference arises because, unlike the crude calculation given here, the MaxEnt model takes into account all of the constraints in the system, including overlapping ones; see Appendix.

5.3 Finding the weights

Fitting MaxEnt constraint weights to data is a well-studied problem in computer science and many effective algorithms are available. For reasons of speed, we chose the L-BFGS algorithm (Liu and Nocedal 1989) as implemented in software code by Tim Hunter.[9] We checked the outputs of Hunter's program by refitting the weights using the GRG Nonlinear engine in Excel's Solver utility (Fylstra et al. 1998), with very similar results.

5.4 Assessing the results

In our statistical testing, we compare the accuracy of two nested models, one including the phonological bigram constraints being tested, the other not (see, e.g., Wilson and Obdeyn 2009, Morley 2015, Shih 2017b). The key numerical value used for evaluation is the LIKELIHOOD a grammar assigns to the data; this is calculated by multiplying the probabilities assigned by (3) to every datum. Likelihood is then converted, for computational convenience, into log likelihood by taking the natural logarithm of the result. For purposes of statistical testing, two log likelihoods are computed: that of a baseline model that includes just the constraints of convenience, and that of a full model that also incorporates our phonological constraints. We then use the Likelihood Ratio Test (Wasserman 2004:164) to determine whether the phonological constraints, serving as predictors, significantly improve the accuracy of the model. We also test the constraints

[8] A note on procedure: it has been our practice to write all software for the project in two independent versions, each author using a distinct programming language, and checking to make sure that our programs are yielding essentially identical results.

[9] Available at https://github.com/timhunter/loglin.

individually, by comparing for each one the full model against a subset models that leaves out just that constraint.[10,11]

5.5 Corpora employed

We examined fourteen corpora. Eight represent authors from the English literary canon; such authors are beyond any doubt competent native speakers, and their works are out of copyright and available in abundance in carefully prepared electronic editions (we used texts from Project Gutenberg, www.gutenberg.org). The remaining corpora are of spoken language, gathered from various sources, either public or available by subscription. For one of the spoken corpora, we amalgamated five similar sources to obtain a total length similar to that of the other corpora.[12]

Table 3: Corpora examined

Text	*Approx. Length in words*	*Source*
Written texts:		
Six novels by Jane Austen	722,000	Project Gutenberg
Six non-fiction works by Charles Darwin	935,000	Project Gutenberg
Six novels by Charles Dickens	709,000	Project Gutenberg
Six novels by Nathaniel Hawthorne	592,000	Project Gutenberg
Nine novels by Jack London	739,000	Project Gutenberg
Six works by Herman Melville	786,000	Project Gutenberg
Four novels by Anthony Trollope	756,000	Project Gutenberg
Six novels by Mark Twain	568,000	Project Gutenberg
Spoken texts:		
Committee corpus: transcripts of committee hearings from the Corpus of Spoken Professional American English	878,000	www.athel.com/cpsa.html
Fresh Air corpus: transcripts of the “Fresh Air” program on National Public Radio	724,000	www.npr.org/programs/fresh-air/archive, interview transcriptions from 5/15/2018 to 9/12/2018

[10] Our procedure for pooling individual words into categories based on their phonological properties (§5.1) evidently does not affect significance values. In the schematic simulation described there (and in the Supplemental Materials), the log likelihood values obtained using word-by-word data differ from those obtained from pooled data (less detail is available for matching the frequencies), but for the phonological constraints being tested, weights and significance values come out the same.

[11] An alternative test, as pointed out by a reviewer, would be to compare the set of baseline constraints against a superset model adding in one phonological constraints. We adopt the method described above because we judge it more stringent — a constraint must prove its worth even in the presence of other added constraints that may overlap with it to some degree (e.g. *CLASH and *IAMBIC CLASH).

[12] A note on procedure: five of the corpora (Austen, Darwin, Five-Spoken, Twain, Hawthorne) were available to us as we worked out our software and methods; the nine others were examined after the software and analytic methods had been finalized.

Michigan lecture: the lecture portions of the Michigan Corpus of Academic English	912,000	www.lib.umich.edu/database/link/11887
Michigan nonlecture: the remaining portions of the Michigan Corpus of Academic English	709,000	www.lib.umich.edu/database/link/11887
Spoken set, a merged set of several spoken corpora, containing: British Academic Spoken English Corpus Beatles Interview Corpus (Stanton 2016) Buckeye Corpus HCRC Map Task Corpus 2016 Primary Debates Corpus	969,000	 www.warwick.ac.uk/fac/soc/al/research/collections/base/history nyu.app.box.com/s/ku8b32q1orh6twkoank40l3zjc146yog buckeyecorpus.osu.edu/ groups.inf.ed.ac.uk/maptask/maptasknxt.html www.kaggle.com/kinguistics/2016-us-presidential-primary-debates/home
White House corpus: Collected transcripts of live questions and answers from White House press briefings (1994-1997)	758,000	www.athel.com/cpsa.html

5.6 Phonetic transcription

Using software, we isolated the words of each corpus and converted them to phonetic transcription, relying on an augmented version of the dictionary used in Hayes (2012),[13] itself an edited version of the CMU Pronouncing Dictionary (www.speech.cs.cmu.edu/cgi-bin/cmudict). In order to limit the number of bigrams that could not be analyzed because we lacked a dictionary entry for one of their members, we augmented our dictionary to include all words that had a frequency greater than 100 in any of the corpora; we also auto-created inflected forms (plural, past, gerund) of the entries. In the end, our efforts provided phonetic transcriptions for 72.8% of the types and 96.9% of the tokens in the corpora. This implies that the probability that a bigram would go unanalyzed because we lack a dictionary entry for one of its words is about 0.06. One further edit was performed: we removed all bigrams from our bigram sets that began with an allomorph of the indefinite article *a*/*an*. Our purpose was to avoid artificial inflation of the weight of *HIATUS, which would have resulted from including them.

The varieties of English that historically dropped /r/ in syllable codas have different distributions of violations for some of our constraints (*GEMINATE, *HIATUS, *TRIPLE CONSONANT CLUSTER, and *BAD SONORITY, and their affiliated constraints of convenience), and we are not always in a position to know whether our authors spoke rhotic or non-rhotic English. To make sure this did not throw off our results, we repeated all of our calculations using a non-rhotic phonetic dictionary, obtained by applying the relevant historical sound changes to our CMU-based dictionary.[14] The result was that we always obtained stronger effects using the original rhotic version of the dictionary irrespective of our guess of whether the author's dialect

[13] Included in the Supplemental Materials.

[14] Viz.: (1) ɹ → ∅ / ___ {C, #}, (2) ɚ → əɹ / ___ V. There were other changes, too, but they would not affect violations of the constraints we use. For thorough discussion of these correspondences, see Wells (1982).

was rhotic or not.[15] Readers may inspect the results obtained using the non-rhotic dictionary in the Supplemental Materials.

5.7 Editing the bigram sets

In the analysis, each of the 14 text corpora was analyzed in several versions, edited to include different subsets of the full bigram set. Our purpose, discussed in detail below, was to test for the effects of syntactic variation, fixed phrases, and lexical frequency. We will first report on the most heavily edited versions of the corpora, which were shortened in three ways. First, for reasons discussed in §7.2, we discarded bigrams that were separated by a major prosodic break, as diagnosed by the presence of punctuation. Second, for reasons discussed in §8.1, we reduced the text to its HAPAX bigrams; that is, those that occurred only once. Lastly, for reasons discussed in §8.2, we discarded bigrams that contained function words. We will call the form of analysis that employs such bigram sets the Core condition. It emerged from our study that the Core condition was the most stringent test for the existence of phonological effects.

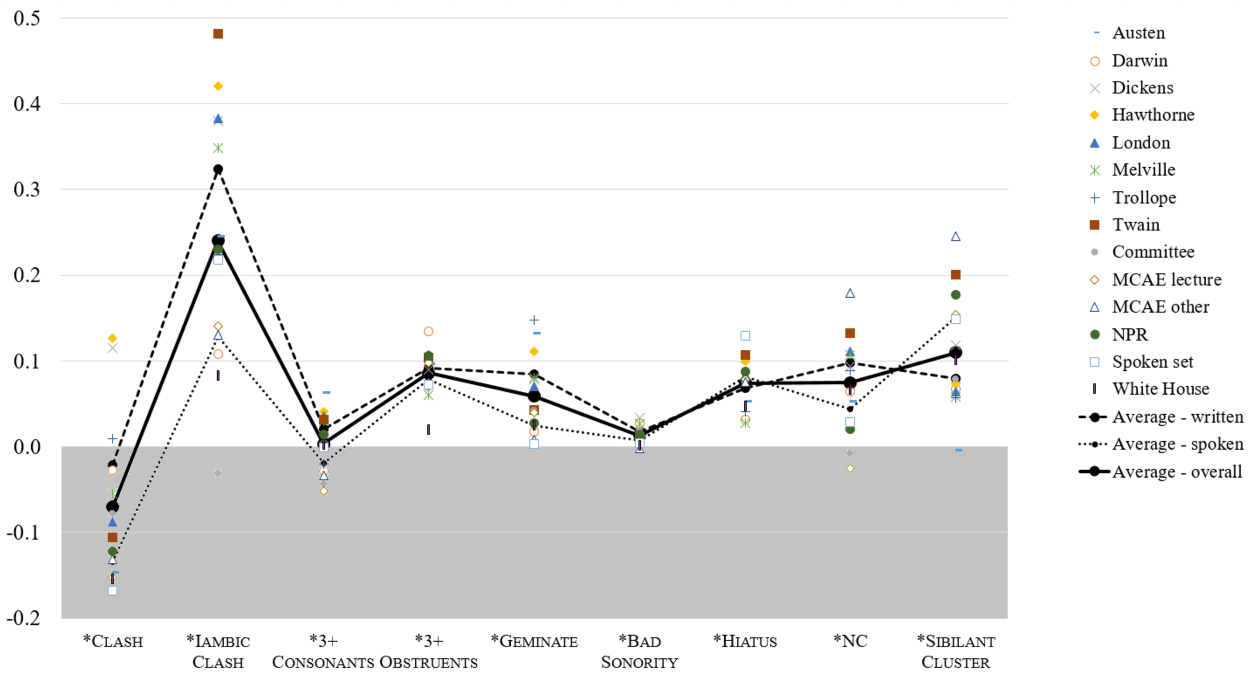
6. Findings for the Core condition

In every corpus, the addition of phonological constraints to the model made a strong positive difference to the model's accuracy. The improvement in the log likelihood of the model ranged from 18.3 for the Committee corpus to 127.6 for the Hawthorne corpus. A likelihood ratio test for the degree of improvement created by adding in the phonological constraints (9 degrees of freedom) yielded significance values ranging from p = 0.00003 for the Committee corpus (worst case) to $p = 7.6 \times 10^{-50}$ for the Hawthorne corpus. The full set of significance results may be viewed in the Supplemental Materials.

The improvement in model accuracy is the work of most, but not always all, of the constraints we tested. Figure 2 shows the weight of each of the constraints as fitted to the Core corpora; we include boldface lines connecting average values for all corpora, as well as dotted lines for all of the written corpora and all of the spoken corpora.

[15] We offer a conjecture for this odd result: nonrhotic dialects generally have r-epenthesis, as in *far away* /fɑː əweɪ/ → [fɑːɹəweɪ] (Wells 1982:§3.2.3). A rhotic dictionary, with /fɑɹ/, is incorrect for these dialects but nevertheless accidentally succeeds in predicting that *far away* would not incur a *HIATUS violation. It was indeed on *HIATUS where the two dictionaries yielded the most distinct results.

Figure 2: Weights obtained for 9 phonological bigram constraints, Core condition.



Plainly, there is variation from constraint to constraint, with frequent negative weights for *Clash — implying it is better to have a stress clash than not. In §8.2 below we offer an explanation of this anomaly, suggesting that speakers actually do respect *Clash when they construct sentences, and that the negative weights observed here are an artifact of the Core condition's bigram exclusion criterion.

In general, the written corpora have higher weights than the spoken ones, plausibly the result of the opportunity writers have to ponder the phonological well-formedness of what they are writing and improve it with edits.[16] Nevertheless, even the spoken corpora show massive statistical effects of phonological markedness. We return to the issue of spoken vs. written language in §8.7.[17]

7. Control studies

Our statistically significant results, found in the stringently-edited Core condition, are encouraging, but invite further scrutiny. In this section, we offer two further tests of our general hypothesis that phonological constraints influence sentence formation across the board; a third test appears in §8.6 below.

[16] For discussion of evidence that writing is genuinely a phonological process, see Schlüter (2005:50-55) and Shih and Zuraw (2017:e320-e321). For evidence that prose authors display different levels of 'metricality' in their prose, see Borgeson et al. (2018).

[17] Beyond the written/spoken difference, there is considerable residual variation among the individual corpora, a pattern that repeats itself for all the analyses and for which we have no explanation.

7.1 Pseudoconstraints

For the first test, we invented an alternative constraint set which was intended (unlike the set in §2) to be utterly arbitrary, so we call them "pseudoconstraints." As far as we can tell, they have no valid typological basis, and our checking with the UCLA Phonotactic Learner indicated that our pseudo-constraints play essentially no role in regulating the sequencing of sounds within words.[18] The set of pseudo-constraints is given in (4).

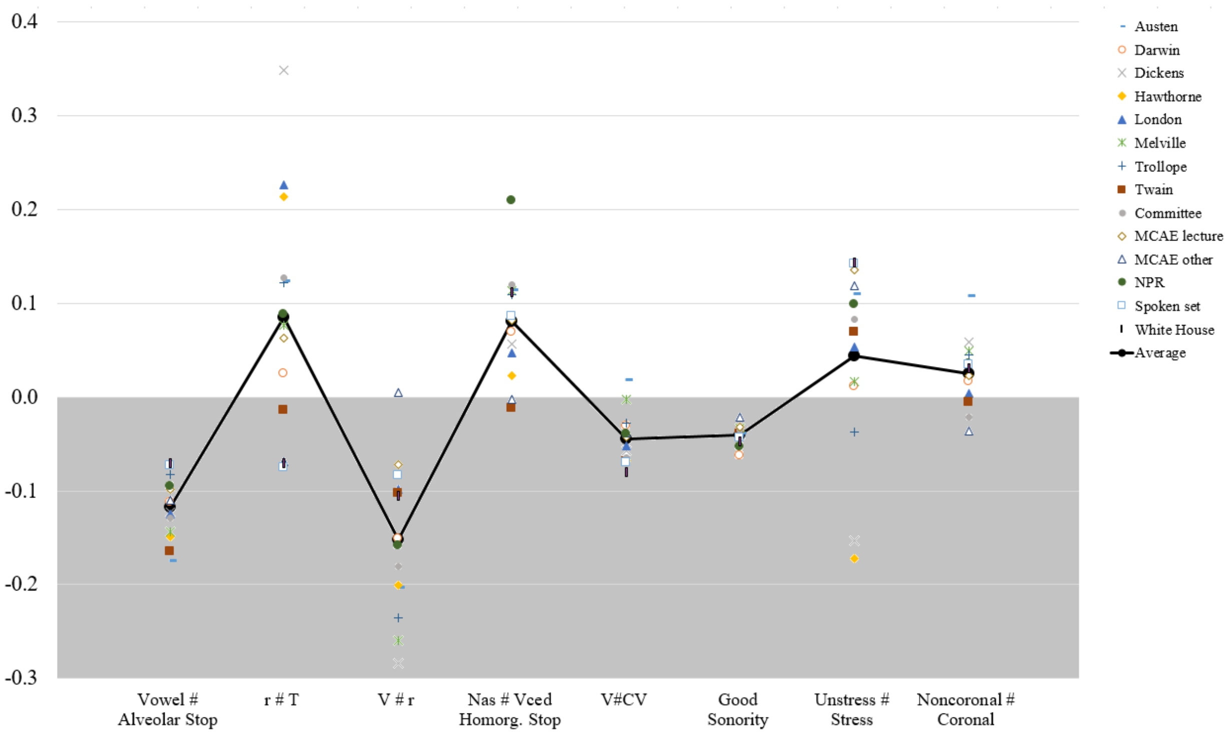
(4) *List of "pseudo-constraints" tested*

	Pseudo-constraint	*Description*
a.	Vowel # Alveolar stop	Open syllable, unmarked consonant
b.	r # Alveolar stop	
c.	Vowel # r	
d.	Nasal # Voiced homorganic stop	Obeys postnasal voicing tendency
e.	V # CV	Maximally unmarked syllabification
f.	In C1 # C2, C1 has more sonority than C2	Obeys the Syllable Contact Law
g.	Unstressed syllable # Stressed syllable	Highly frequent non-clashing configuration
h.	Noncoronal C # Coronal C	See Blust (1979), arguing that this is the unmarked order.

We note up front that we cannot confidently predict ZERO weights for these constraints; our current understanding of phonology or of English is hardly good enough to make such a prediction. Rather, we expect that their weights will fall into no particular pattern; some weights may be positive, some negative (meaning, better to violate than to obey), but no weight especially strong. This is indeed what emerges from the testing, as shown in Figure 3.

[18] Constraint (4h) receives the small weight of 0.25; the others zero.

Figure 3: Weights obtained for 8 pseudo-constraints



Only one pseudo-constraint, *Unstressed syllable # Stressed syllable, tested significant at the .05 level for the majority of the 14 corpora, and below (§8.2) we suggest that even this result is an artifact of our mode of analysis. When we compare the constraint weights in the aggregate for our Core and Pseudoconstraint conditions, we find that the average weight for a “real” constraint (Core condition) is 0.065 and for a pseudo-constraint is −0.14 (i.e., better to violate). A two-tailed unpaired *t*-test comparing the full set of Core condition constraint weights (9 × 14 = 126) with the full set of pseudo-constraint weights (8 × 14 = 112) indicates the difference is highly significant, p < .00001. Full details for all significance tests may be viewed in the Supplemental Materials. In sum, we think the result of this control study is to support the view that the effects found above for our real constraints (§6) result from the fact that they ban configurations that are genuinely disfavored in English phonology, rather than arbitrary configurations.

7.2 Violations across phrasal breaks

A second test for the validity of our results is based on the widely observed pattern that processes of phrasal phonology are blocked across prosodic breaks; see, e.g., Nespor and Vogel (1986/2007), Hayes (1989). To give just one of many empirical examples, Jun (1996:70) demonstrates that the phonological process of Intervocalic Lenis Stop Voicing in Korean is blocked across Accentual Phrase breaks. Thus, in (5), underlying /k/ is converted to [g] when the intervocalic environment is created within an Accentual Phrase, but intervocalic /p/ remains voiceless, since the intervocalic environment arises across a phrase break.

(5) *Korean Intervocalic Lenis Stop Voicing blocked by Accentual Phrase break*

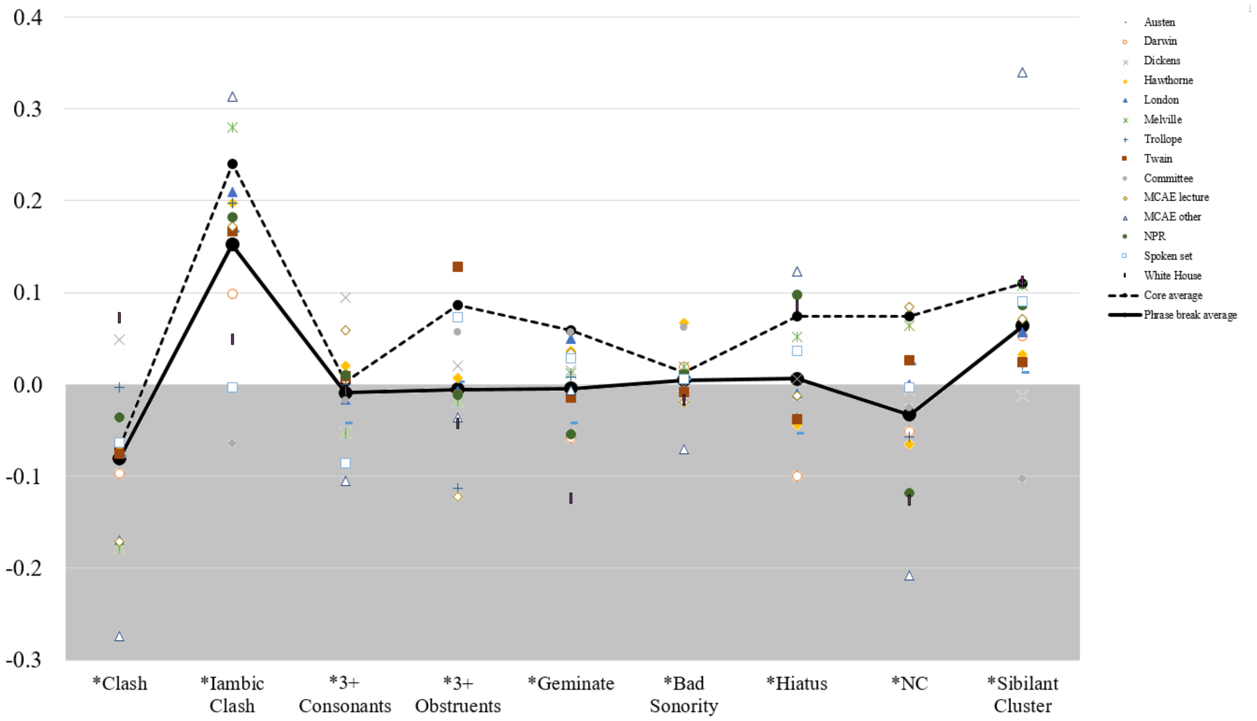
/ (kəmɨn kojaŋi-e)$_A$	(palmok)$_A$ /	phonemic form with Accentual Phrasing
[g]	[p]	phonetic output
black cat-GEN	ankle	gloss: ‘the ankle of the black cat’

This is stated in process terms, but is more appropriately treated for present purposes with constraints: there is a Markedness ban on voiceless intervocalic plain stops that dominates Faithfulness for voicing, but is defined to be applicable only within Accentual Phrases. We expect that similar phrase-bounding is likely to hold true for the markedness constraints for English examined here.

This forms the basis for the test we next describe, which is inspired by Gunkel and Ryan (2011). In our Core condition, we deliberately excluded bigrams whose words are separated by punctuation, which is generally diagnostic of a phrase break.[19] Here, in a contrasting Phrase Break condition, we do the opposite, retaining bigrams only when they are formed across punctuation. The graph in Figure 4 shows the weights of the constraints in the Phrase Break condition; the boldface line represents the average across corpora, and the dotted line shows the comparable averaged results for the Core condition.

[19] As a reviewer pointed out, there are many prosodic breaks, typically weaker ones, that are not marked by punctuation; these occur for instance at subject-predicate boundaries. We suspect these weaker breaks could also be shown to have an effect, but with our current methodology we cannot test them. The fact that our method treats such breaks as phrase internal, despite their possibly exerting a blocking effect, biases AGAINST our hypothesis that phonological markedness skews choices involved in sentence-formation.

Figure 4: Weights obtained for 9 phonological bigram constraints, Phrase Break condition.



As can be seen, the weights for the Phrase Break condition are scattered about zero, meaning that the effects mostly disappear across phrase breaks — following the normal typology of Markedness constraints. The Phrase Break weights are significantly lower on average than the weights obtained in the Core condition (means 0.010 vs. 0.065; *p* = .00002). This difference would hardly be expected if the original effect were just a random occurrence, but it makes sense if what we have found is a true phonological effect.

8. Seeking the causes

Our tentative conclusion, then, is that across-the-board phonological markedness effects do indeed obtain in sentence formation. In particular, the effects occur for constraints that are phonologically plausible (§7.1), and they largely evaporate across phrase breaks, as one would expect in phrasal phonology (§7.2). We next explore versions of our corpora that are edited in other ways, with the goal of learning something about the mechanisms responsible for these markedness effects. We relate our findings to the research literature reviewed in §1.

8.1 Varying bigram frequency to test listed phrases

We first attempt to find in our own data a pattern discovered for compound words by Martin (2011). Using a different statistical technique (see Appendix), Martin found substantial underrepresentation of markedness violations in bigrams consisting of the component elements of compound words. In Navajo, he detected underrepresentation of compounds violating Sibilant Harmony, an important principle of Navajo phonology. In English compounds, as already noted, he found that geminates are disfavored. For Martin, the explanation of these patterns lies in the process whereby newly created compounds propagate through a speech community and become accepted into the lexicon as listed items. Specifically, he suggests that phonotactic markedness is a barrier to such acceptance. Mollin (2012) has found similar evidence that binomials (*X and Y*) are less likely to lexicalize to a fixed order when they violate phonological constraints.

While our own corpora are not limited to compounds, there is nevertheless reason to think that the concept of preferential lexicalization of phonologically-optimal sequences might be applicable. This is because a separate research tradition has found evidence that phrases are also frequently lexicalized; see e.g. Pauley and Syder (1983) and Jackendoff (1997:§7.2). Indeed, a widely expressed view is that the number of memorized phrases in a language is very large, perhaps even greater than the number of memorized words (Mel’čuk 1998).

One other key research result is that lexicalization is dependent on frequency: the more frequent a word sequence is, the more likely it is that it will be lexicalized. This has been established in experimental work; see instance Arnon and Snider (2010) and the body of work cited there. As an intuitive illustration, we list below some word bigrams that appear frequently in the Austen corpus:

(6) *Some familiar-sounding bigrams from the Austen corpus*

a. *Content words*

very well, great deal, young man, very good, few minutes, young ladies, next morning, soon afterwards, same time, young woman, next day, soon after, very soon, only one, last night

b. *With function words*

and then, my dear, the next, I think, I believe, at home, at first, at once, in love, in town, going to, it is, do not, did not, was not[20]

Putting all these elements together, we hypothesize that the Martinian mechanism enforcing phonological unmarkedness in compounds would be expected also to apply to phrasal sequences. Since lexicalized bigrams tend to be frequent, it should be the case that when we look at frequent bigrams, we will find an enhancement of phonological markedness effects.

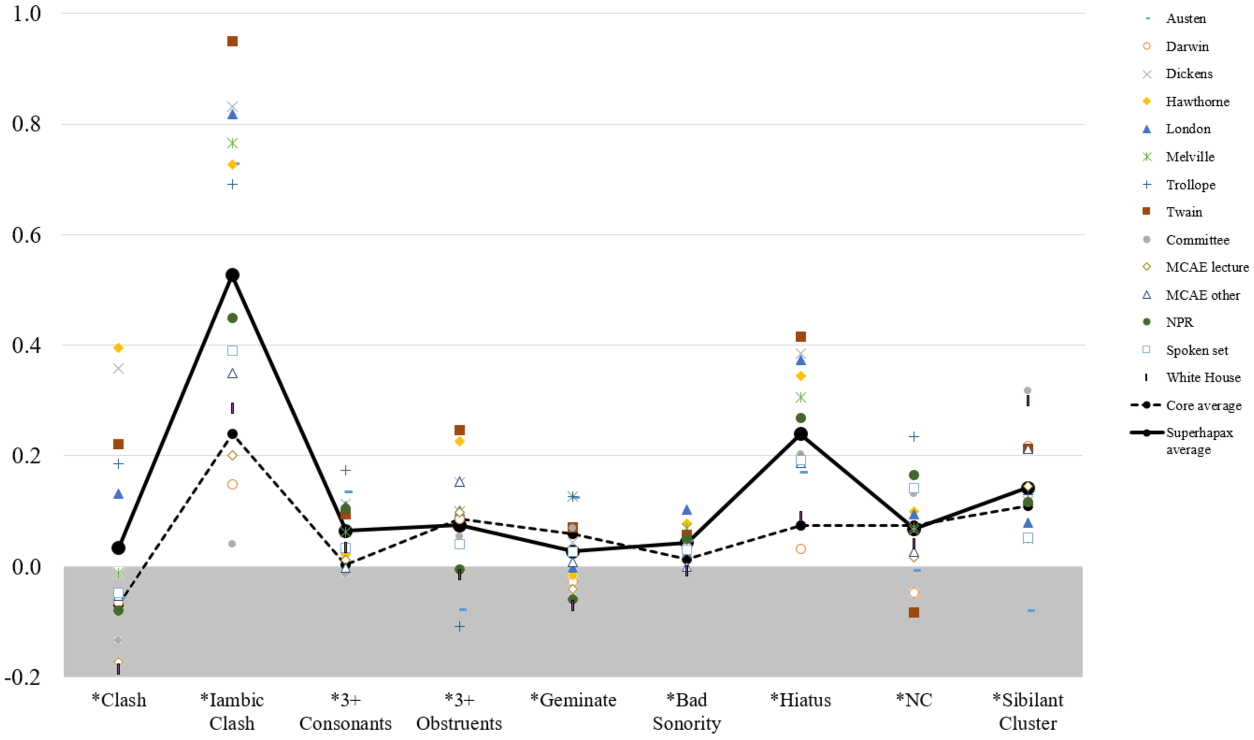
To this end we set up a Superhapax condition, in which the corpora were edited to include only the bigrams of frequency 2 or greater. These edited corpora otherwise matched the criteria for the Core condition above: they excluded function word bigrams and bigrams formed across punctuated phrasal breaks.

Our findings for the Superhapax condition are given in Figure 5, which gives the weights obtained for all 9 phonological constraints across 14 corpora; the boldface line represents the

[20] Observe that the last five items have contracted forms.

constraint weights averaged across corpora, and the dotted line shows the comparable value for the earlier Core condition.[21]

Figure 5: Constraint weights for 14 corpora, Superhapax condition (phrase-internal, function words excluded, superhapax bigrams)



It should be clear from the figure that the superhapax bigrams generally yield stronger phonological markedness effects than the hapax bigrams (means 0.135 vs. 0.065; *p* = .0007). We draw the inference that listed phrases tend to be less phonologically marked, and that the responsible mechanism is plausibly what was outlined above, i.e. that the Martinian principle of preferential lexicalization for phonological unmarked forms carries over to lexically listed phrases.

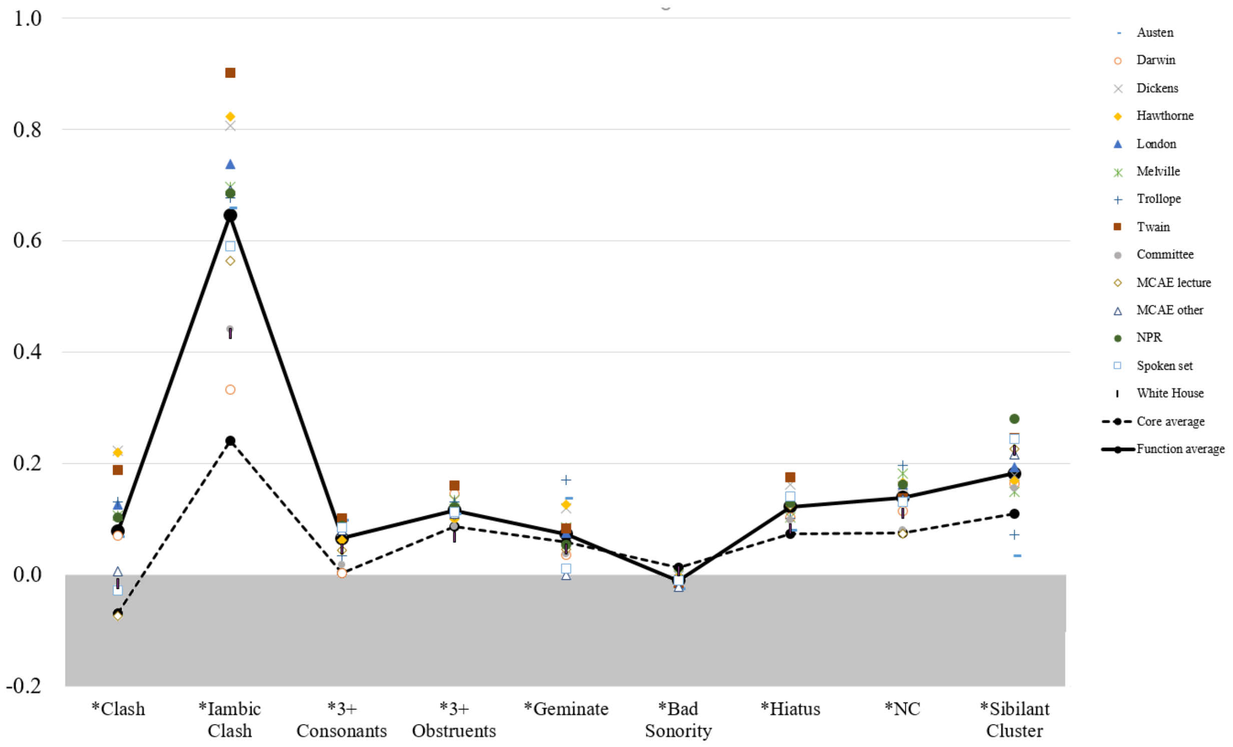
8.2 Function words and syntax

The data described in this section consist of another variant of our bigram sets, again created for each of the 14 text corpora. Here, we diverge from Core in a different way: we include the bigrams that contain function words (e.g. determiners, prepositions, pronouns, auxiliary verbs, complementizers, stressless adverbs). We retain from Core the practice of keeping only hapax bigrams, as we are not interested here in the effects of listed bigrams (such as (6b)) that contain function words. We also retain from Core the restriction that the bigrams must be contained within phrases, as diagnosed by punctuation. We will refer to this as the Function condition.

[21] The vertical scale of this graph, and all subsequent ones, is expanded relative to the scale employed in Figs. 1-4.

Like the Superhapax condition, the Function condition yields higher weights than Core for the phonological markedness constraints (means 0.157 vs. 0.065; *p* = .00001). This can be seen in Figure 6.

Figure 6: Constraint weights for 14 data corpora, Function condition (phrase-internal, function words included, hapax bigrams)



We think the factor most likely to be responsible for this effect is syntax. As already noted (§1), the research literature has adduced multiple instances showing that when the syntax offers a binary choice for expressing the same meaning, speakers tend to pick the phonologically less marked option; recall the examples of Tagalog *Adjective + linker + Noun* vs. *Noun + linker + Adjective*, English *give X to Y* vs. *give Y X*, and English *Y's X* vs. *X of Y*. Typically, at least one of the choices employs a function word. We consider here a taxonomy of possibilities, based on where the function word occurs.

In the first case, both syntactic variants include a function word. For instance, in the Tagalog case, the linker morpheme *na* appears in either word order. As Shih and Zuraw (2017) note, if one flanking word ends in a nasal and the other does not, then one of the two orders will incur a violation of *[+nasal][+nasal]; an example is *áma<u>ng n</u>a túnay* vs. *túnay na ámang* ('real elder/father', p. e326). We expect, given Shih and Zuraw's findings, that Tagalog speakers in forming sentences will particularly favor unmarkedness in this syntactic context, since the grammar gives them a ready opportunity to do so. Consider, then, how such a case would be treated when examined under our Core condition: all of the examples of this syntactic construction would get culled out, because the relevant bigrams contain a function word (*na*). The upshot is that the weight of the markedness constraint *[+nasal][+nasal] will go down in

Core condition relative to the Function condition, since some of the best evidence for it has been discarded.

The second syntax-related pattern occurs where a function word appears in only one of the two syntactic options, namely the one that is phonologically less marked. The canonical instance of this is *CLASH. For instance, if the two syntactic variants are the two forms of the dative construction (*gíve bóoks to Bíll*, *gíve Bíll bóoks*) then the dative function word *to,* being stressless, will often avert a clash (here, the clash between *Bíll* and *bóoks*). If, as Shih (2017a) suggests, variants of the dative construction are indeed deployed to reduce phonological markedness, then the procedure used in forming the Core corpora will create distortion, because the discarded bigrams are clash-free, whereas the retained ones are clashing. The same outcome will occur for other syntactic constructions — see Shih (2017a) on genitives, Wasow et al. (2015) on *to*-dropping, and Jaeger (2006) on *that*-dropping — and thus is probably responsible for the anomaly observed in §6, namely overrepresentation in many corpora under the Core condition of *CLASH violations.[22]

In principle, there should be a third case, one in which the appearance of function words consistently induces, rather than averts, violations of a phonological constraint. In such a case one would expect higher constraint weights in the Core than in the Function condition. We have alerted ourselves to detect such cases empirically but have not yet found any.[23]

In sum, when compared with the Function condition, the Core condition emerges as informative precisely for its distortions, which involve syntax. When we remove function words, we remove many of the examples where syntactic choices permit speakers to avoid phonological markedness, reducing the constraint weights. When the removal actually targets the unmarked cases, we sometimes get negative weights in Core even for markedness constraints which have strong support elsewhere in the language. The result is to provide indirect evidence to support what other scholars have shown directly through scrutiny of particular constructions. Our own comparisons demonstrate how strong the aggregate effect of syntactic choice is likely to be, and how it applies for multiple constraints.

8.3 Rethinking the results for the Core condition

We return to the fact that although the constraint weights found in the Core condition were generally smaller, they were nonetheless generally positive and consistently statistically significant in the aggregate. Why should this be so, given that the Core condition was designed to minimize the influence of syntax and listed phrases? We conjecture that the patterns we found in Core represent residual effects of listed phrases and of syntax that proved impossible to control for completely using our methods; e.g. perhaps some of our hapax bigrams really were listed phrases that by accident happened to be used just once in the corpus; or there are major

[22] The same may hold for the underrepresentation of *UNSTRESSED # STRESSED violations found in §7.1. This pseudo-constraint covers part of the complement set of *CLASH; and our Pseudoconstraint mode likewise excluded function words.

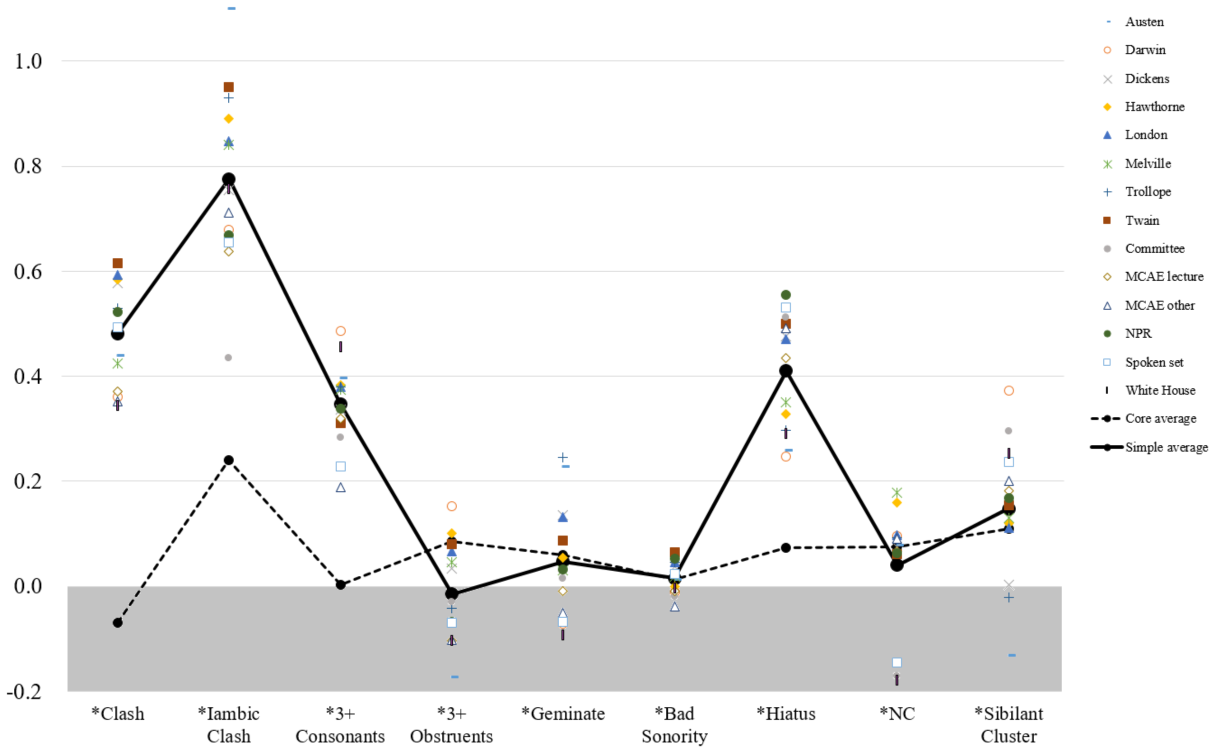
[23] A sensible place to look is *LAPSE (banning adjacent stressless syllables); we cannot check it here because it would require us to expand our search method to an infeasible number of candidates.

effects of syntactic choice that involve no function words. The effects seen in the Core condition may also reflect word choice, as documented by Schlüter and colleagues (§1).

8.4 Including all factors together

Lastly, in Figure 7 below, we give our findings for a set of minimally-curated bigrams, which folds together hapax and the superhapax bigrams, includes function words, and counts tokens rather than types, so that a bigram that appears *n* times is counted *n* times rather than just once. The only bigrams that are excluded from Figure 7 are those occurring across prosodic breaks. Since this condition is not curated in any way (other than the well-motivated phrase break exclusion) we call it the Simple condition.

Figure 7: Constraint weights for 14 data corpora, Simple condition (hapaxes and superhapaxes together, counted by tokens, function words included, phrase-internal only)



As might be expected, the effects here are at their strongest (comparing again to Core: means 0.250 vs. 0.065; *p* < .000001). This is because we have both included superhapax bigrams (where lexical listing encourages obedience to markedness constraints), and function word bigrams (which incorporate the reduced markedness resulting from syntactic choices). The combination also introduces an additional set of lexically-listed bigrams that were absent from the Superhapax condition, namely those that include function words. Lastly, using token counts instead of types would also be expected to increase the effect of phonological markedness, since the highest-frequency bigrams, which are most likely to be listed, receive more influence when token-counting is employed.[24]

A quirk seen in Figure 7 is that *3+ CONSONANTS rises in weight and *3+ OBSTRUENTS falls, relative to the Core, Superhapax, and Function conditions. These two constraints are "ganged" (Jäger and Rosenbach 2006), in the sense that whatever violates *3+ OBSTRUENTS also violates *3+ CONSONANTS; so that a triple obstruent cluster automatically accrues whatever penalty falls on triple consonant clusters. What this implies is that triple obstruent clusters are avoided in the Simple condition, but not any more than any triple consonant clusters are. We observe also that *IAMBIC CLASH is ganged with *CLASH, so that the harmony penalty incurred by any iambic clash is in fact the sum of the weights of *CLASH and *IAMBIC CLASH.

8.5 Interpreting the constraint weights

What do the constraint weights of Figure 7 mean in terms of actual probability of use? The MaxEnt formula (3) provides a concrete answer. Imagine two candidates, identical except that one violates a constraint with weight *w* and the other does not. It is readily deduced from (3) that the probabilities assigned to them will occur in a particular RATIO (odds): the probability of the non-violator is $\boldsymbol{e^{w}}$ times higher than the probability of the violator.

(7) *Probability ratio of Candidate 1 and Candidate 2, differing in only one constraint violation*

$$\frac{\text{P(Candidate 1)}}{\text{P(Candidate 2)}} = \frac{\frac{e^{-H}}{Z}}{\frac{e^{-(H+w)}}{Z}} = \frac{e^{-H}}{e^{-(H+w)}} = \frac{1}{e^{-w}} = e^{w}$$

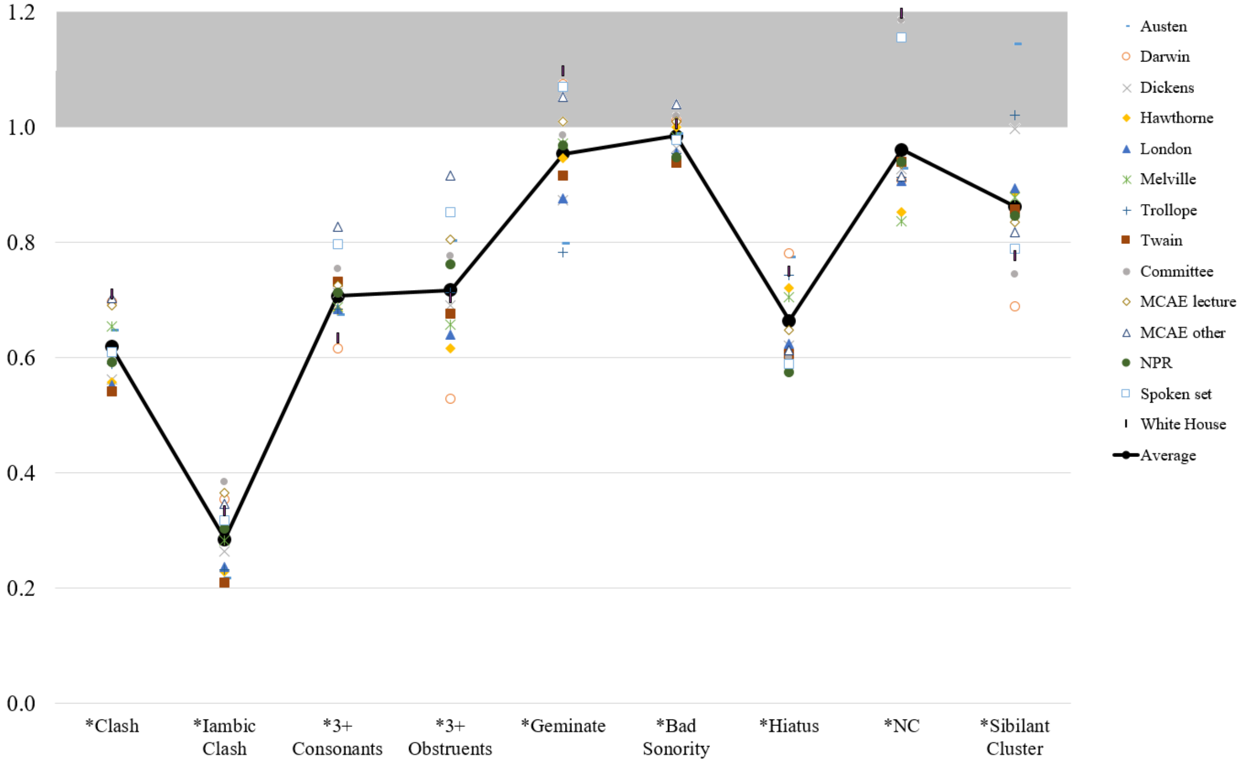
where

H = harmony penalty resulting from shared violations
w = penalty for violating constraint C

We use this formula to replot the data of Figure 7 to display these probability ratios. We have also augmented the weights of *IAMBIC CLASH and *3+ OBSTRUENTS with the more general constraints (*CLASH and *3+ CONSONANTS) that gang with them; this gives a clearer picture of their empirical effect. The result is shown in Figure 8.

[24] We checked the specific contribution of token frequency by creating a further condition just like the Superhapax condition, except that it counted by tokens instead of types. A substantial increase in average constraint weight resulted, namely types 0.135, tokens: 0.186.

Figure 8: Data from Figure 7 replotted as probability ratios; ganging applied to two constraints

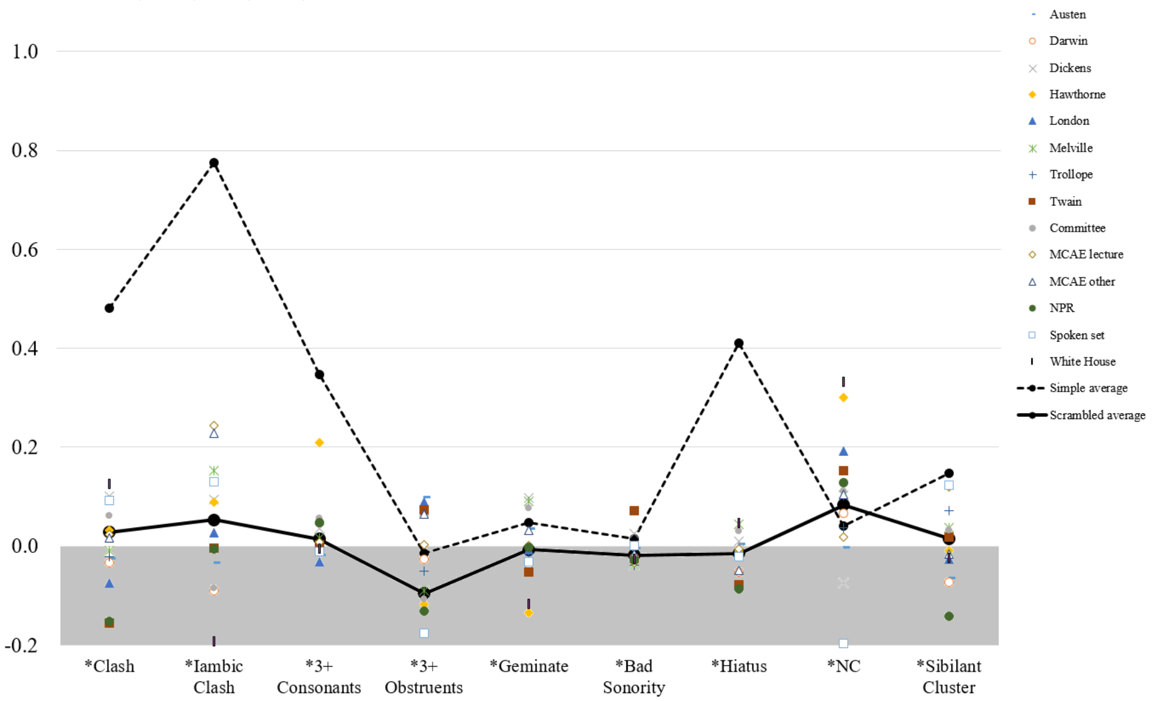


As can be seen, the average reduction in probability (one minus the value shown) ranges from a remarkable 71.5% for *IAMBIC CLASH to just 1.5% for *BAD SONORITY.

8.6 *The Simple condition with random data*

Reviewers asked us what would happen if we employed our method on SCRAMBLED versions of the original texts, with random order. For the Austen corpus, a typical sentence of the scrambled version would be *Anne as had better good Elizabeth chair Fairfax's just dear of a, was, to was they or, always a most weakly the but.*). If the findings laid out above are linguistically-driven, they should evaporate under scrambling, which should remove all forms of systematic linguistic patterning. This procedure is best implemented in the Simple condition, since the reduction of data from types to tokens would be expected to introduce modest degrees of de-randomization (the frequent bigrams have frequent words, which have distinct phonological properties). Figure 9 show the results of this approach; the dotted line shows the average of the Simple condition from Figure 7.

Figure 9: Results for Simple condition with random (scrambled) text

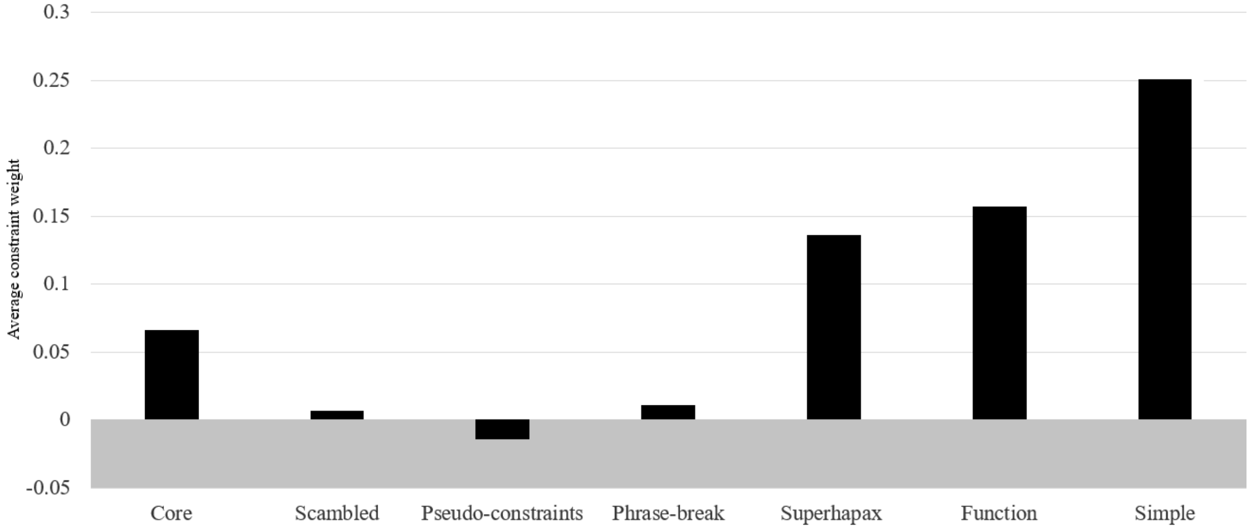


Here, we are interested not in whether the weights are distinct from those found in the Simple condition (they obviously are, *p* < .00001), but simply whether their magnitude is meaningful at all. A one-sample two-sided t-test indicates that the weights obtained in the Scrambled condition do not differ significantly from zero (*p* = .483). This result encourages us in thinking that our method is working reliably.

8.7 Overall summary and statistical testing

As an overview of our findings, Figure 10 gives the average constraint weights for all seven conditions.

Figure 10: Average constraint weights for each condition



The first column, representing our Core condition (§5), shows a relatively weak effect of phonological markedness, but one that nevertheless emerges as highly statistically significant (taking the constraints in the aggregate) in every one of our 14 corpora, including the spoken ones. The next three columns show that the phonological effects largely disappear under the three control conditions we examined: the substitution of pseudo-constraints for true markedness constraints (§7.1), the inspection of bigrams formed across phrasal breaks (§7.2), and the use of scrambled (random) data (§8.6). The fifth and sixth columns suggest that the phonological effects are stronger in listed bigrams (Superhapax condition, §8.1), and that the characteristic ordering of function and content words by the syntax creates the opportunity for speakers to favor phonologically unmarked bigrams (Function condition, §8.2). The final, tallest column (Simple condition, §8.4) illustrates the combination of these effects, augmented by the effect of token frequency.

We can also obtain an overview of how the individual constraints performed, using the results for the Simple condition. To do this, we sorted the constraint/corpus combinations into categories as follows: (1) “Backward” means that the constraint weight actually came out negative (better to violate it). (2) Otherwise, we provide *p*-values for significance testing (likelihood ratio test) at two different alpha levels, $\alpha = .05$ and $\alpha = .001$.[25] We counted how many of the 14 corpora tested as significant at each level; hence the best-performing constraints will have the value 14 in the rightmost column of Table 4, while others will scatter further to the left.

[25] We employed the Bonferroni correction for multiple comparisons ($n = 14$), so the actual cutoff values were .05/14 and .001/14.

Table 4: Assessment of 9 constraints in Simple condition. Entries designate number of corpora in category.

Constraint	*Backward*	*p > .05*	*.001 < p < .05*	*p < .001*
*Clash				14
*Iambic Clash				14
*3+ Consonants				14
*3+ Obstruents	8	1		5
*Geminate	5	3		6
*Bad Sonority	6		1	7
*Hiatus				14
*NC̥	3		2	9
*Sibilant Cluster	2	1		11

It can be seen that the constraints *Clash, *Iambic Clash, *3+ Consonants, and *Hiatus all do very well. *3+ Obstruents performs poorly; the reader will recall that it is ganged with *3+ Consonants, which means simply that sonority in triple clusters does not matter for this domain of analysis.[26] For the remaining constraints the individual results are, for us, suggestive but not conclusive. However, if we take the nine constraints in the aggregate, their effect on bigram frequencies seems unquestionable; the improvement in the log likelihood of the model for the Simple condition are very substantial, ranging from 5000.9 for the MCAE-Nonlecture corpus to 11999.8 for the London corpus. For the full statistical data (by corpus, condition, and constraint) see Supplemental Materials.

Lastly, we return to the difference between written and spoken corpora, already observed in the Core condition. As Table 5 shows, this difference persists in the other test conditions.

Table 5: Average constraint weights for written vs. spoken corpora, all test conditions

	Core	*Function*	*Superhapax*	*Simple*
Written	0.084	0.174	0.176	0.277
Spoken	0.040	0.133	0.082	0.213

9. General discussion

At this point we have demonstrated, we believe, that our bigram/MaxEnt method diagnoses widespread avoidance of phonological Markedness violations in English, and that the mechanisms whereby this happens include syntactic choice, a Martinian tendency to preferentially list phrases that are phonologically unmarked, and probably also lexical choice. We next consider various interpretations and implications of our findings.

[26] In our word internal checking (see section §2), the weight of *3+ Obstruents, ganged with *3+ Consonants, is modest (*3+ Obstruents 1.33, *3+ Consonants 1.94).

9.1 The hypothesis of raw phonetic difficulty

Before making grander claims, we should consider a very modest explanation of our findings. Under this view, the effects we are seeing are not even grammar. Work in phonetically-based phonology (e.g. Hayes 1999, Steriade 2001, Hayes et al. 2004, Wilson 2006) suggests that phonological constraints might be construed as devices that arise as grammatical responses to phonetic difficulty. In this view, stress clashes, adjacent sibilants, and the like are difficult for ALL human speakers, but their distribution in particular languages are regulated by the constraints of the phonological grammar. The idea would be that the effects we are seeing result not from grammar but from phonetic difficulty itself, the essential linking hypothesis being that in every language it is a pragmatic principle of speaking (or writing) to avoid such difficulties to some degree.

However, a crucial finding of Hammond's (2016) work on the Rhythm Rule suggests this hypothesis is untenable. Hammond not only finds (as we and others did) that iambic clashes are statistically avoided in English, but such clashes are avoided even when repaired: a phrase like *ùnkind pérson*, derived from /*unkínd* + *pérson*/, has no clash, but it is still partially avoided. The natural interpretation of this under our assumptions is that what is being avoided is a *Faithfulness* violation, namely of whatever Faithfulness constraint is violated in shifting the underlying stress of *unkínd*. [27] The hypothesis of raw phonetic difficulty cannot explain Hammond's result.[28] Moreover, to the extent that avoidances are language-specific (see Shih and Zuraw 2017 and below), we likewise cannot accept the raw phonetic difficulty theory as valid.

9.2 Architecture of the language faculty

Assuming, then, that the effects we observe are the consequence of grammar, we turn to the question of what sorts of grammatical architecture could explain our findings (and similar findings from other researchers; §1). We think a very simple answer may be possible, and that the ingredients of this answer have already been put forth in the literature.

To begin, researchers have repeatedly proposed parallel architectures for linguistic theory: the various distinct forms of linguistic representation (syntactic, semantic, phonological, etc.), should be dealt with in parallel rather than in an ordered sequence; see Sadock (1991), Jackendoff (1997, 2002, 2010), Bresnan (2000), Anttila (2016), Shih and Zuraw (2018), and Bruening (2019). The key idea is that it is possible to maintain distinct types of representations, each governed by its own set of well-formedness principles, without demarcating separate components among which the direction of information flow is stipulated. Bresnan (1998:67) describes this general approach as follows, referring to

[27] For other constraints we studied, violations are also sometimes repaired in the phrasal phonology; e.g. clusters are simplified through consonant drop, and hiatus resolved by insertion of [ʔ] or [ɹ]. We have no way of assessing when such repairs are taking place (thus incurring Faithfulness violations) and we record them simply as Markedness violations; the essential point is that either way there is a cost in Harmony.

[28] A reviewer suggests an alternative interpretation of the Hammond data; i.e. that stress-shifted allomorphs like *ùnkind* are lexically listed and perhaps not available to all speakers. We think this issue revolves around whether the Rhythm Rule is treated as phonology or phrasal allomorphy.

> “a class of frameworks in which the [grammar] of language is modeled as linked parallel structures, each of a different formal character. The grammar consists of a set of local co-descriptive constraints on partial structures. There are no derivational or transformational operations involved: grammatical structures are defined by constraint satisfaction. Each of the parallel structures of [such theories] models a different dimension of the structure of language.”

A second strand of work concerns the relationship of constraint violations to well-formedness. The work of Keller (2000, 2006) and Featherston (2005, 2019), both of whom have extensively studied syntactic well-formedness judgments experimentally, suggests the following conclusions. (1) Individual syntactic constraints, when violated, contribute PARTICULAR DEGREES of ill-formedness. To capture this fact, both authors advocate some version of Harmonic Grammar, in which each constraint bears a numerical weight. (2) Violations are CUMULATIVE: the ill-formedness contributed by separate syntactic constraint violations, or by multiple violations of the same constraint, must be added together to obtain an accurate model of experimental data. In other words, native well-formedness intuitions, scaled appropriately, match with Harmony scores. Such findings offer empirical support for a Harmony-based theory of well-formedness. Moreover, the very same patterns of constraint-specificity and cumulativity have been shown to hold as well for phonology (Coleman and Pierrehumbert 1997, Hayes and Wilson 2008, and much subsequent work). These findings augur well for the project of blending phonological and syntactic constraints when computing Harmony, and the initial phases of such research might be seen in the work of Clifton et al. (2006) on *CLASH effects and their interaction with syntactic Superiority.

The final step is the one taken by MaxEnt and similar theories, in which further computations based on Harmony yield probability values (as in (3)), opening the way to modeling corpus frequencies and to making use of powerful existing algorithms to model language learning. The MaxEnt version of Harmonic Grammar concretely implements Featherston’s idea that in speaking, individuals select from among the higher-probability candidates, in proportion to their probability. MaxEnt syntax was proposed early in this century by computational linguists (see Manning 2003 and work cited there), and has since been pursued in work such as Velldal and Oepen (2005), Bresnan et al. (2007), Bresnan and Hay (2008), and Irvine and Dredze (2017).

What do these strands of work imply for our results? First, in a model based on Harmony, no constraint can be outright ignored; every constraint has an effect on well-formedness — often small, though sometimes large (as in (1b)). Second, in a parallel architecture, the constraints that participate in assignment of Harmony necessarily include phonological constraints. Lastly, under the Harmony-frequency connection posited under MaxEnt, speakers are predicted to skew their choice of utterances towards the more harmonic possibilities. This means that, at least subtly, this skewing will be based on phonology. We suggest that this article has offered a method sensitive enough to detect this skewing.

Already, empirical work has been done that incorporates all of the elements described above: cross-componentiality (with phonology included), Harmony-based well-formedness, and

MaxEnt-based predictions about corpus frequency; see in particular work such as Bresnan et al. (2007), Shih et al. (2015), Shih and Zuraw (2017), and Szmrecsanyi et al. (2017).[29]

9.3 Restrictiveness

Zwicky and Pullum (1986) long ago argued for a pure feed-forward model on grounds of restrictiveness: the feed-forward architecture automatically rules out bizarre patterns like "a movement transformation that obligatorily moves ... [a] constituent that begins phonetically with a bilabial consonant" (p. 75). We agree with Shih and Zuraw (2018:4) and Bruening (2019) that the right explanation for such absence is not feed-forward architecture but rather the specific character of syntactic and phonological constraints (§9.2). Assuming that there exists syntactic movement into (for instance) Spec-CP, we assert it unlikely that the phonology would ever have a constraint specifically penalizing the absence of labials in Spec-CP. The empirical effects seen so far, both here and in the literature cited in §1, are compatible with the mechanism we propose, namely candidate competition regulated by competing syntactic and phonological constraints.[30]

9.4 Choice of phonological frameworks

Rule-based phonology (Chomsky and Halle 1968), which still has many adherents, rejects a key idea of Optimality Theory and other constraint-based theories, namely that putative "rules" like A → B / C __ D are actually composite entities; there is a Markedness constraint, *CAD, which outranks all Faithfulness constraints that militate against the change A → B. A key argument in favor of OT is that *CAD often has its own independent existence. That this is so have been argued for on the basis of conspiracies (Kisseberth 1970 et seq.), phonotactics (Kenstowicz and Kisseberth 1977), optionality patterns (Anttila 1997), and speech errors (Goldrick and Daland 2009).

We submit that our results add to this list of arguments. Our findings do not concern phonological alternations (the focus of rules), but probability distributions in output forms generated by principles found throughout the entire grammar. Rule-based phonology has nothing to say about such cases, but they are a natural consequence of including phonological constraints in a parallelist MaxEnt system.

[29] Any effort to work out a parallelist approach along the lines of this section must come to terms with the arguments given by Agbayani and colleagues (Agbayani and Golston 2010, 2016; Agbayani et al. 2015) for a component-ordered system in which movement in syntax strictly precedes movement in phonology. While parallelist reanalysis of this body of work is a task that goes far beyond the scope of this article, we note that it is encouraging that the ordering arguments these authors provide are specifically of the *bleeding* type (in the taxonomy of Kiparsky 1968); this is precisely the ordering known to be most readily reanalyzed in parallelist constraint-based frameworks.

[30] Although we cite Zwicky and Pullum in §1 above for an informal characterization of the syntax-phonology connection that we and many others consider to be wrong, the more explicit Principle of Phonology-Free Syntax that they end up proposing (p. 71) is actually very close to what Shih, Zuraw, Bruening, and we ourselves think: "No syntactic rule can be subject to language-particular phonological conditions or constraints."

10. For future work

10.1 A puzzle from Hungarian

We sought to generalize our results by examining the phrasal patterning of three languages well known for their vowel harmony: Turkish (e.g., Clements and Sezer 1982), Finnish (Kiparsky 1973), and Hungarian (Siptár and Törkenczy 2000). Our tentative results indicate that both Turkish and Finnish texts show a modest tendency to avoid phrasal bigrams that violate their respective vowel harmony principles (backness harmony and rounding harmony for Turkish; just backness harmony for Finnish). However, bafflingly (from the viewpoint of our research experience), in Hungarian there is a statistically significant tendency to favor bigrams that actually violate the backness harmony found in the word-level phonology of language.

The Hungarian pattern finds least a modest rationalization in the principle, dating from Trubetzkoy (1939), that phonology provides *Grenzsignale*, boundary signals that assist the listener in parsing the incoming speech stream into words (see, e.g. Cutler and Norris 1988 et seq.). Thus, when there is phrasal disharmony, a shift between harmonic categories of two vowels in sequence will assist listeners by informing them of a greater probability that a word boundary is present. But why vowel harmony should be a simple Markedness effect in Turkish and Finnish, but a *Grenzsignal* in Hungarian, is a mystery to us.

10.2 Learnability

We argued in §9.2 for why, under MaxEnt, we expect that relatively weak phrasal phonological constraints should make their presence felt, albeit subtly. Yet we did not address why such constraints should occur in the grammar in the first place, and how they are related to the word-internal phonology. A plausible mechanism for this comes from Martin's (2011) concept of grammatical "leakage," a form of overgeneralization. His idea is that when children learn the phonotactic restrictions active within words, they weakly overgeneralize, expressing the same constraints in non-word-bounded versions that end up influencing higher-level constructions. He also shows with learning simulations how such overgeneralizations obtain by default under a specific, conservative strategy of language acquisition. If Martinian overgeneralization is correct, it directly follows that the constraint *NC̥ (§2.7) should generally have yielded no effects in our analyses; it is ineffective within English words and thus cannot be overgeneralized to the phrasal context.

An additional possibility, which seems more radical to us, is that patterns of bigram avoidance are outright learned by children as part of the grammar of their language. This is the obvious explanation to be applied to our Hungarian findings, which remain tentative. The clinching evidence for this (as with Bresnan and Ford's (2010) syntactic work) would be the demonstration of consistent dialect-specific effects in weights assigned to the constraints; and some tentative evidence for this in the case of *SIBILANT CLASH has been offered by Szmrecsanyi et al. (2017).

10.3 Complete grammars

In our study we tried to control for syntactic effects by making a comparison between our Core condition and the Function condition (§8.2), but clearly one could do more: ideally, one would adopt one single probabilistic grammar, along the lines given in §9.2, containing a complete set of all the constraints needed for both syntax and phonology. With such a grammar, we could test statistically if the phonological constraints are significantly impacting sentence formation, in a way that could control more carefully for syntactic effects.[31] Currently, the kind of syntactic grammars that could be adapted to this purpose — by which we mean, computationally-implemented, probabilistic, and comprehensive — are sparse on the ground, but we anticipate that progress on such grammars is likely to be rapid in the future. Such grammars would obviously permit greater rigor in the work that is described here.[32] We also see the pursuit of multi-component grammars as an beneficial counterweight to the increasing separation of subdisciplines in our field, and that such study would help us to share our thinking concerning issues common to both the “S-side” and “P-side” of grammar. [33]

Appendix: Methodology

We defend here our decision to use MaxEnt as the basis for modeling, as opposed to simpler forms of statistical reasoning.

Our project was originally inspired by Martin’s (2011) study of statistical underrepresentation in compounds, which demonstrated that English compounds are formed in lesser numbers when a geminate would be created, as in *bookkeeper*. For some time we actually used Martin’s method, which is based on calculating the expected statistics of two-word sequences (for Martin, compounds) if the choice of Word 1 and Word 2 is independent. Martin does this with a “shuffling” procedure: each Word 2 is re-paired with a randomly-chosen Word 1, resulting in a set of shuffled pairs that respects the statistics of the Word 1 and Word 2 populations. The shuffle is evaluated for violations, then the whole procedure is repeated several thousand times, yielding a probability distribution for violation counts. The counts of the real text are then compared with this distribution, yielding a statistical significance value.

We abandoned this method when we realized that it cannot be trusted to handle cases where constraints apply to overlapping sets of forms, as in our work. To see this, imagine the following language: every word takes the form CVC, where C = one of [p b t d s ʃ z ʒ] and V = one of [i e a o u]. Thus there are 8 × 5 × 8 = 320 possible words — all of which are assumed to exist. We construct a synthetic set of word bigrams by first assembling every possible bigram (320 × 320 =

[31] Such an approach might also be able to handle nonlocal phonological effects (e.g. prosodic phrasing, tone spread), which are not treatable with simple bigrams.

[32] For that matter, an approach of the type we suggest could provide greater rigor in dealing with syntactic data: in light of the forcefulness of *IAMBIC CLASH in lowering the probability of a sentence, we think it would be a mistake, for example, to test with consultants a syntactic minimal pair in which only one of the two options included an *IAMBIC CLASH violation.

[33] We can think of: behavior governed by individual lexical items, productivity and exceptionality, diachronic vs. synchronic explanation, the role of listed sequences, free variation, and the relationship of computed probability to well-formedness judgments.

102,400), then removing precisely one half of the bigrams that contain a *SIBILANT CLASH violation, leaving 89,600 bigrams total. Clearly, the right conclusion to draw for such a text would be that *SIBILANT CLASH is active, and that no other constraint is active, in the phrasal domain — the extreme symmetry of the bigram set is meant to guarantee this.

We tested Martin's shuffling method on our imaginary language, with the constraints *SIBILANT CLASH and *GEMINATE. Unsurprisingly, the method found a strong effect of *SIBILANT CLASH (the real count was 0.78 times the expected value from the shuffles, and the effect size was 50 standard deviations). However, the shuffling method also found a strong effect for *GEMINATE (0.82 times expected value, effect size 24 standard deviations) which as noted above is a wrong diagnosis. The reason for the error is that many of the *SIBILANT CLASH violations ([ss, ʃʃ, zz, ʒʒ]) also happen to be *GEMINATE violations.

Analyzing the same language with MaxEnt gives a very different outcome: *SIBILANT CLASH receives a weight of 0.693, which corresponds (§8.5) to 50% underrepresentation, the correct value. The weight assigned to *GEMINATE is zero, again correct.

Why the difference in performance? The Martinian shuffling procedure has no basis for attributing effects to particular constraints when they overlap in their violation patterns. In contrast, MaxEnt invokes a highly effective procedure intended to predict the data as a whole as accurately as possible; this forces the constraints to do the jobs for which they are best suited — MaxEnt (as its name implies) penalizes all empirically unjustified deviations from randomness, in this case, any nonzero weight for *GEMINATE.

In sum, MaxEnt, unlike the shuffling method, is capable of attributing underrepresentation to the appropriate constraint when constraints overlap in coverage, and thus is to be preferred for investigations of the kind we are conducting.

The discussion in this appendix is based on the parallel example given in Wilson and Obdeyn (2009), who address the well-known Observed/Expected statistic (Pierrehumbert 1993) of which Martin's shuffling system is a variant; see also Jurafsky and Martin (2019:ch5) for discussion of the ability of MaxEnt to disentangle correlated factors.

References

Agbayani, Brian, and Chris Golston. 2010. Phonological movement in classical Greek. *Language* 86.133–167.

Agbayani, Brian, and Chris Golston. 2016. Phonological constituents and their movement in Latin. *Phonology* 33.1–42.

Agbayani, Brian, Chris Golston, and Toru Ishii. 2015. Syntactic and prosodic scrambling in Japanese. *Natural Language and Linguistic Theory* 33.47–77.

Anttila, Arto. 1997. Deriving variation from grammar: A study of Finnish genitives. *Variation, change and phonological theory*, ed. by Frans Hinskens, Roeland van Hout, and Leo Wetzels, 35–68. Amsterdam: John Benjamins.

Anttila, Arto. 2016. Phonological effects on syntactic variation. *Annual Review of Linguistics* 2.115–137.

Anttila, Arto, Matthew Adams, and Michael Speriosu. 2010. The role of prosody in the English dative alternation. *Language and Cognitive Processes* 25.946–981.

Arnon, Inbal and Neal Snider. 2010. More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language* 62.67–82.

Benor, Sarah Bunin, and Roger Levy. 2006. The chicken or the egg? A probabilistic analysis of English binomials. *Language* 82.233–278.

Blust, Robert. 1979. Coronal-noncoronal consonant clusters: New evidence for markedness. *Lingua* 47.101–117.

Boersma, Paul, and Joe Pater. 2016. Convergence properties of a gradual learning algorithm for Harmonic Grammar. In McCarthy and Pater, 389–434.

Bolinger, Dwight. 1965. Pitch accent and sentence rhythm. *Forms of English: Accent, morpheme, order*, 139–180. Cambridge, MA: Harvard University Press.

Borgeson, Scott, Arto Anttila, Ryan Heuser, and Paul Kiparsky. 2018. The rise and fall of antimetricality. Ms., Department of Linguistics, Stanford University.

Bresnan, Joan. 1998. Morphology competes with syntax: Explaining typological variation in weak crossover effects. *Is the best good enough? Optimality and competition in syntax*, ed. by Pilar Barbosa, Danny Fox, Paul Hagstrom, Martha McGinnis, and David Pesetsky, 59–92. Cambridge, MA: MIT Press and MIT Working Papers in Linguistics.

Bresnan, Joan. 2000. Optimal syntax. *Optimality theory: Phonology, syntax and acquisition*, ed. by Joost Dekkers, Frank van der Leeuw and Jeroen van de Weijer, 334–385. Oxford: Oxford University Press.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the dative alternation. *Cognitive foundations of interpretation*, ed. by G. Boume, I. Krämer, and J. Zwarts, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, Joan, and Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86.168–213.

Bresnan, Joan and Jennifer Hay. 2008. Gradient grammar: An effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua* 118.245–259.

Bruening, Benjamin. 2019. On the motivation for phonology-free syntax and multiple levels in a derivational grammar. Ms., University of Delaware. https://udel.edu/~bruening/Downloads/PhonFreeSyntax1.pdf

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

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

Chung, Sandra. 2003. The syntax and prosody of weak pronouns in Chamorro. *Linguistic Inquiry* 34.547–599.

Clements, George N. 1990. The role of the sonority cycle in core syllabification. *Papers in laboratory phonology 1*, ed. by John Kingston and Mary E. Beckman, 283–333. Cambridge: Cambridge University Press.

Clements, George N., and Engin Sezer. 1982. Vowel and consonant disharmony in Turkish. *The structure of phonological representations (Part II)*, ed. by Harry van der Hulst and Norval Smith. Dordrecht: Foris.

Clifton, Charles, Gisbert Fanselow, and Lyn Frazier. 2006. Amnestying Superiority violations: Processing multiple-Wh questions. *Linguistic inquiry* 37.51–68.

Coleman, John, and Janet Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. In *Computational Phonology, Third Meeting of the ACL Special Interest Group in Computational Phonology*, 49–56. Somerset, N.J.: Association for Computational Linguistics.

Cutler, Anne, and Dennis Norris. 1988. The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance* 14.113–121.

Della Pietra, Stephen, Vincent J. Della Pietra, and John D. Lafferty. 1997. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19.380–393.

Embick, David, and Rolf Noyer. 2001. Movement operations after syntax. *Linguistic Inquiry* 32.555–595.

Erteschik-Shir, Nomi, Gunlög Josefsson, and Björn Köhnlein. 2019. Variation in Mainland Scandinavian Object Shift and prosodic repair. lingbuzz/003688.

Featherston, Sam. 2005. The decathlon model of empirical syntax. *Linguistic evidence: Empirical, theoretical, and computational perspectives*, ed. by Stephan Kepser and Marga Reis, pp. 187–208.

Featherston, Sam. 2019. The Decathlon Model. *Current Approaches to Syntax: A Comparative Handbook*, ed. by Andras Kertesz, Edith Moravcsik, and Csilla Rakosi, pp. 155–186.

Fijn van Draat, Pieter. 1910. Rhythm in English prose. *Anglistische Forschungen* 29. Heidelberg: C. Winter.

Fylstra, Daniel, Leon Lasdon, John Watson, and Allan Waren. 1998. Design and use of the Microsoft Excel Solver. *Interfaces* 28.29–55.

Goldrick, Matthew, and Robert Daland. 2009. Linking speech errors and phonological grammars: Insights from Harmonic Grammar networks. *Phonology* 26.147–185.

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. *Proceedings of the Stockholm workshop on variation within optimality theory*, ed. by Jennifer Spenader, Anders Eriksson and Östen Dahl, 111–120. Stockholm: Stockholm University.

Gunkel, Dieter and Kevin M. Ryan. 2011. Hiatus avoidance and metrification in the Rigveda. In *Proceedings of the 22nd Annual UCLA Indo-European Conference*, ed. by Stephanie W. Jamison, H. Craig Melchert, and Brent Vine, 53–68. Bremen: Hempen.

Hammond, Michael. 2016. Input optimization: phonology and morphology. *Phonology* 33.459.

Harford, Carolyn, and Katherine Demuth. 1999. Prosody outranks syntax: An Optimality approach to subject inversion in Bantu relatives. *Linguistic Analysis* 29.47–68.

Hayes, Bruce. 1989. The Prosodic Hierarchy in meter. *Rhythm and meter*, ed. by Paul Kiparsky and Gilbert Youmans, 201–260. Orlando, FL: Academic Press.

Hayes, Bruce. 1999. Phonetically-driven phonology: the role of optimality theory and inductive grounding. *Functionalism and formalism in linguistics, Volume I*, ed. by Michael Darnell, Edith Moravcsik, Michael Noonan, Frederick Newmeyer, and Kathleen Wheatly, 243–285. Amsterdam: John Benjamins.

Hayes, Bruce. 2012. The role of computational modeling in the study of sound structure. Paper presented at the Conference on Laboratory Phonology, Stuttgart.

Hayes, Bruce, Robert Kirchner, and Donca Steriade (eds.) 2004. *Phonetically based phonology*. Cambridge: Cambridge University Press.

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

Hayes, Bruce, Colin Wilson, and Anne Shisko. 2012. Maxent grammars for the metrics of Shakespeare and Milton. *Language* 88.691-731.

Hooper, Joan B. 1976. *An introduction to natural generative phonology*. New York: Academic Press.

Irvine, Ann and Mark Dredze. 2017. Harmonic Grammar, Optimality Theory, and syntax learnability: An empirical exploration of Czech word order. arXiv preprint arXiv:1702.05793.

Jackendoff, Ray. 1997. *The architecture of the language faculty*. Cambridge, MA: MIT Press.

Jackendoff, Ray. 2002. *Foundations of language*. New York: Oxford University Press.

Jackendoff, Ray. 2010. The parallel architecture and its place in cognitive science. *The Oxford handbook of linguistic analysis*, ed. by B. Heine and H. Narrog, 583–605. Oxford: Oxford University Press.

Jaeger, Florian. 2006. Phonological optimization and syntactic variation: The case of optional 'that'. *Proceedings of the Berkeley Linguistics Society* 32.175–187.

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

Jun, Sun-Ah. 1996. *The phonetics and phonology of Korean prosody: intonational phonology and prosodic structure*. New York: Garland Publishing.

Jurafsky, Daniel, and James H. Martin. 2019. *Speech and Language Processing* (3rd ed. draft). https://web.stanford.edu/~jurafsky/slp3/.

Keller, Frank. 2000. Gradience in grammar: Experimental and computational aspects of degrees of grammaticality. Ph.D. dissertation, University of Edinburgh.

Keller, Frank. 2006. Linear Optimality Theory as a model of gradience in grammar. *Gradience in grammar: Generative perspectives*, ed. by Gisbert Fanselow, Caroline Féry, Matthias Schlesewsky, and Ralf Vogel, 270-287.

Kenstowicz, Michael, and Charles Kisseberth. 1977. *Topics in phonological theory*. New York: Academic Press.

Kiparsky, Paul. 1968. Linguistic universals and linguistic change. Universals in linguistic theory, ed. by Emmon Bach and Robert T. Harms, 170-202. New York: Holt, Rinehart & Winston.

Kiparsky, Paul. 1973. Phonological representations. *Three dimensions of linguistic theory*, ed. by Osamu Fujimura, 5–135. Tokyo: Tokyo Institute for Advanced Studies of Language.

Kisseberth, Charles W. 1970. On the functional unity of phonological rules. *Linguistic Inquiry* 1.291–306.

Liberman, Mark, and Alan Prince. 1977. On stress and linguistic rhythm. *Linguistic Inquiry* 8.249–336.

Liu, Dong C., and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming* 45.503–528.

Manning, Christopher D. 2003. Probabilistic syntax. *Probabilistic linguistics*, ed. by Rens Bod, Jennifer Hay, and Stefanie Jannedy, pp. 289–341.

Martin, Andrew. 2011. Grammars leak: Modeling how phonotactic generalizations interact within the grammar. *Language* 87.751–770

McCarthy, John and Joe Pater (eds.) 2016. *Harmonic grammar and harmonic serialism*. London: Equinox Press.

McCarthy, John, and Prince, Alan. 1993. *Prosodic morphology: Constraint interaction and satisfaction*. Linguistics Department Faculty Publication Series 14. Amherst, MA: Department of Linguistics, University of Massachusetts.

Mel'čuk, Igor. 1998. Collocations and lexical functions. *Phraseology: Theory, analysis and applications*, ed. by A. P. Cowie, 23–53. Oxford: Clarendon Press.

Mollin, Sandra. 2012. Revisiting binomial order in English: ordering constraints and reversibility. *English Language and Linguistics* 16.81–103.

Morley, Rebecca L. 2015. Can phonological universals be emergent?: Modeling the space of sound change, lexical distribution, and hypothesis selection. *Language* 91.e40–e70.

Murray, Robert W. and Theo Vennemann. 1983. Sound change and syllable structure in Germanic phonology. *Language* 59.514–528.

Nespor, Marina, and Irene Vogel. 1986/2007. *Prosodic phonology.* Dordrecht: Foris. 2nd. ed. Berlin: Walter de Gruyer.

Pater, Joe 1999. Austronesian nasal substitution and other NÇ̥ effects. *The prosody-morphology interface*, ed. by René Kager, 310–343. Cambridge: Cambridge University Press.

Pater, Joe. 2000. Non-uniformity in English secondary stress : the role of ranked and lexically specific constraints. *Phonology* 17.237–274.

Pater, Joe. 2016. Universal grammar with weighted constraints. In McCarthy and Pater, 1–46.

Pauley, A. and F. H. Syder. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. *Language and Communication*, ed. by J. C. Richards and R. W. Schimdt, 191–226. London: Longman.

Pierrehumbert, Janet B. 1993. Dissimilarity in the Arabic verbal roots. *Proceedings of the North East Linguistic Society* 23.367–381.

Prince, Alan and Paul Smolensky. 1993/2004. *Optimality theory: Constraint interaction in generative grammar*. Technical report, Rutgers University Center for Cognitive Science. [Published 2004; Oxford: Blackwell]

Rice, Curt. 2007. Gaps and repairs at the phonology–morphology interface. *Journal of Linguistics* 43.197–221.

Rice, Curt, and Peter Svenonius. 1998. Prosodic V2 in Northern Norwegian. Ms., University of Tromsø.

Ryan, Kevin. 2019. Prosodic end-weight reflects phrasal stress. *Natural Language and Linguistic Theory* 37.315–356.

Sadock, Jerrold M. 1991. *Autolexical syntax: A theory of parallel grammatical representations*. Chicago: University of Chicago Press.

Schlüter, Julia. 2005. *Rhythmic grammar: The influence of rhythm on grammatical variation and change in English.* Berlin: Mouton de Gruyter.

Schlüter, Julia. 2015. Rhythmic influence on grammar: Scope and limitations. In Vogel and Ruben van de Vijver, 179–205.

Schlüter, Julia, and Gabriele Knappe. 2018. Synonym selection as a strategy of stress clash avoidance. *Corpora and lexis*, ed. by Sebastian Hoffmann, Andrea Sand, Sabine Arndt-Lappe, and Lisa Marie Dillman, 69–105. Leiden: Brill.

Shih, Stephanie S. 2017a. Phonological influences in syntactic choice. *The morphosyntax-phonology connection: locality and directionality at the interface*, ed. by Vera Gribanova and Stephanie S. Shih, 223–252. Oxford: Oxford University Press.

Shih, Stephanie S. 2017b. Constraint conjunction in weighted probabilistic grammar. *Phonology* 34.243–268.

Shih, Stephanie S., and Jason Grafmiller. 2011. Weighing in on end weight. Paper presented at the 85th Annual Meeting of the Linguistic Society of America, Pittsburgh, Pennsylvania.

Shih, Stephanie S., Jason Grafmiller, Richard Futrell, and Joan Bresnan. 2015. Rhythm's role in predicting genitive alternation choice in spoken English. In Vogel and van de Vijver, 207–234.

Shih, Stephanie S., and Kie Zuraw. 2017. Phonological conditions on variable adjective-noun word order in Tagalog. *Language: Phonological analysis* 93.e317–e352.

Shih, Stephanie S., and Kie Zuraw. 2018. The nature of the phonology-syntax interface, from variable adjective and noun word order in Tagalog. Unpublished manuscript, USC and UCLA. lingbuzz/004296.

Siptár, Péter, and Miklós Törkenczy. 2000. *The phonology of Hungarian*. Oxford: Oxford University Press.

Smolensky, Paul. 1986. Information processing in dynamical systems: Foundations of harmony theory. *Parallel distributed processing, 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.

Stanton, Juliet. 2016. Learnability shapes typology: the case of the midpoint pathology. *Language* 92.753–791.

Steriade, Donca. 2001. The phonology of perceptibility effects: the P-map and its consequences for constraint organization. Ms, University of California, Los Angeles.

Szmrecsanyi, Benedikt, Jason Grafmiller, Joan Bresnan, Anette Rosenbach, Sali Tagliamonte, and Simon Todd. 2017. Spoken syntax in a comparative perspective: The dative and genitive alternation in varieties of English. *Glossa* 2.1–27.

Temperley, David. 2009. Distributional stress regularity: A corpus study. *Journal of Psycholinguistic Research* 38.75–92.

Trubetzkoy, Nikolai S. 1939. *Grundzüge der Phonologie.* Göttingen: Vandenhoeck and Ruprecht. [Translated 1969 by Christiane A. M. Baltaxe as *Principles of phonology*. Berkeley: University of California Press.]

Velldal, Erik & Oepen, Stephan. 2005. Maximum entropy models for realization ranking. Proceedings of the 10th Machine Translation Summit, ed. by Jun-ichi Tsujii. Asia-Pacific Association for Machine Translation.

Vogel, Irene and Ruben van de Vijver (eds). 2015. *Rhythm in cognition and grammar: A Germanic perspective*. Berlin: De Gruyter.

Wasow, Thomas, Roger Levy, Robin Melnick, Hanzhi Zhu, and Tom Juzek. 2015. Processing, prosody, and optional *to*. *Explicit and implicit prosody in sentence processing*, ed. by Lyn Frazier and Edward Gibson, 133–158. Heidelberg: Springer.

Wasserman, Larry. 2004. *All of statistics: A concise course in statistical inference*. New York: Springer.

Wells, John C. 1982. *Accents of English: An introduction*. 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.

Wilson, Colin. 2014. Maximum Entropy models (tutorial presentation). Department of Linguistics, MIT.

Wilson, Colin, and Marieke Obdeyn. 2009. Simplifying subsidiary theory: statistical evidence from Arabic, Muna, Shona, and Wargamay. Ms., Johns Hopkins University.

Zec, Draga, and Inkelas, Sharon. 1990. Prosodically constrained syntax. *The phonology-syntax connection*, ed. by Sharon Inkelas and Draga Zec, 365–378. Chicago: University of Chicago Press.

Zwicky, Arnold, and Geoffrey K. Pullum. 1986. The principle of phonology-free syntax: introductory remarks. *Ohio State University Working Papers in Linguistics* 32.63–91.