Deconstructing Wordlikeness Judgments

1 Introduction

It has long been observed that native speakers have strong intuitions about "possible words" in their language. For example, Halle (1962) and Chomsky and Halle (1965, 1968) claim certain nonce words—like [bɪk] or [blɪk]—are judged by native speakers to be possible words of English, whereas others—[bnɪk] or [vnig]—are judged impossible. There have been many experimental studies attempting to probe the source of such "wordlikeness" judgments, i.e., acceptability judgments of nonce words, and such studies now often include sophisticated computational modeling of the human judgments obtained (Albright 2009; Albright and Hayes 2003; Bailey and Hahn 2001; Coetzee and Pater 2008; Coleman and Pierrehumbert 1997; Daland et al. 2011; Frisch, Pierrehumbert, and Broe 2004; Gorman 2013; Greenberg and Jenkins 1964; Hayes and Wilson 2008; Ohala and Ohala 1978, 1986; Pierrehumbert 2001, 2003; Pizzo 2015; Scholes 1966; Vitevitch and Luce 1998, 1999; Vitevitch et al. 1997; Wilson and Gallagher 2018, amongst others).

One key observation about wordlikeness judgments, going back to some of the earliest work on the topic (Chomsky and Halle 1968; Greenberg and Jenkins 1964; Scholes 1966), is that they are gradient in the sense that nonce words tend to form a cline of acceptability. For example, Scholes (1966) asked American seventh grade students to judge whether a series of nonce words were possible or impossible words of English; by aggregating the judgments across students, he discerned a cline of acceptability. Many researchers presume that such observed gradience ultimately stems from a gradient phonotactic grammar independent of the lexicon (Albright 2009; Albright and Hayes 2003; Coetzee and Pater 2008; Coleman and Pierrehumbert 1997; Daland et al. 2011; Frisch, Pierrehumbert, and Broe 2004; Hayes 2000; Hayes and Wilson 2008; Pierrehumbert 2001, 2003; Pizzo 2015; Shademan 2006; Wilson and Gallagher 2018, among others). Alternatively, other researchers have proposed that observed gradience reflects similarity to existing words, perhaps because the judgment task makes use of a lexical search procedure (Bailey and Hahn 2001; Ohala and Ohala 1978, 1986; Vitevitch and Luce 1998, 1999; Vitevitch et al. 1997).

In recent work, Gorman (2013) found that models of lexical similarity are often better models of human wordlikeness judgments than computational models implementing a gradient phonotactic grammar. In that study, a simple non-gradient baseline phonotactic grammar which tracks whether or not the given nonce words are comprised solely of licit (i.e., attested) onsets and rimes often outperforms gradient computational models. These results suggest that it is possible to maintain a categorical phonotactic grammar and still account for the observed gradient wordlikeness judgments, consistent with traditional categorical grammatical architectures (Chomsky 1965; Schütze 1996, 2011; Sprouse et al. 2018; Valian 1982). However, it is important to point out that Gorman's (2013) results are limited to previously-collected judgments, aggregated across participants, from three small-scale English experiments. Thus there is a need to reproduce these results without the use of aggregated data, and more generally, to probe gradience in wordlikeness judgments.

Another question, closely related to the question of gradience, concerns the notion of cumulativity of phonotactic violations. Some wordlikeness judgment tasks have found that nonce words with multiple violations are rated worse than those with fewer violations (Ohala and Ohala 1986; Pizzo 2015), and similar observations have been made in artificial language learning tasks (Breiss 2020; Durvasula and Liter 2020). It is not clear which computational models effectively capture these effects. Relatedly, Albright (2009) finds that computational models trained on whole words fit human wordlikeness judgments more poorly than those trained only on onsets. Albright suggests these results are artifacts of stimulus selection, but they show the need for systematic study of whole-word computational models.

Beyond knowledge of phonotaxis and lexical similarity, there are likely many other interacting factors that affect wordlikeness judgments. Indeed, linguists and psychologists have long recognized that all linguistic performance is a complex, multifactorial phenomenon (Chomsky 1965; Schütze 1996, 2011; Valian 1982). One such additional factor on wordlikeness judgments is the perceptual system itself. There is some reason to suspect that the perceptual system produces categorical probability distributions over possible representations (Norris and McQueen 2008; Norris, McQueen, and Cutler 2003), and some of the observed gradience in wordlikeness judgments may reflect probabilistic perception. Furthermore, stimuli which violate native language phonotactic generalizations may induce perceptual illusions (Dupoux et al. 1999; Kabak and Idsardi 2007). For instance, Japanese speakers perceive /ebɯzo/ when [ebzo] is presented, likely because consonant clusters like [bz] are illicit in Japanese. The preceding observations raise a question relevant to wordlikeness judgments: just how do speakers make judgments of unacceptability, if unacceptable stimuli are those which induce perceptual illusions, i.e., those which are unfaithfully perceived and perceptually "repaired"? Two possibilities suggest themselves: either (a) acceptability judgments are made before perceptual repair, making them independent of repair, or (b) the acceptability judgment is made after perceptual repair, making them contingent upon the post-repair percept. The second possibility would pose serious problems for the study of phonotactic knowledge via wordlikeness judgments. This also raises a question as to which level(s) of representation are implicated in wordlikeness judgments. While it is often assumed that wordlikeness judgments are made on the basis of surface phonological structure, some have argued that the perceptual system outputs the best estimates of the intended underlying representations (Daland, Oh, and Davidson 2019; Durvasula and Kahng 2015)—if so, wordlikeness judgments are likely made based on underlying representations. This view is consistent with early generativist theories of wordlikeness (Stanley 1967), the "reverse inference" tradition in speech perception (Halle and Stevens 1962), and accounts for perceptual compensations to assimilation (Gaskell and Marslen-Wilson 1996; Gow 2003; Mitterer, Kim, and Cho 2013).

Another key factor affecting wordlikeness judgments is the nature of the task itself. This topic has not received any systematic study: virtually all studies other than Scholes' early work use Likert scale responses aggregated across subjects. In a foundational study on the bias introduced by the use of ratings scales, Armstrong, Gleitman, and Gleitman (1983) observe that participants give gradient judgments of membership even for categorical, "definitional" categories like "even number". From this Armstrong et al. argue that use of Likert scales influences the task model adopted by participants. Relatedly, Ohala and Ohala (1986, p. 247), point out that Likert scales "permit a great deal of 'noise,' from, among other things, drift in subjects' frame of reference when making their judgments." In other words, there can be no experiments totally devoid of confounds (Valian 1982). Therefore, it is necessary to employ different stimulus presentation and response modalities to fully understand the task models speakers use to judge wordlikeness. Of course, there are many other possible factors—prosodic factors like minimal word constraints (McCarthy and Prince 1986; Nespor and Vogel 1986), or morphological parsing (Coleman and Pierrehumbert 1997; Hay, Pierrehumbert, and Beckman 2004), for instance—that could influence wordlikeness judgments, but we believe it is essential to sift through the aforementioned factors systematically, and at greater scale than before, before addressing more subtle performance factors.

In this grant we ask a set of interconnected experiments which target these core issues while also building in conceptual replications of many of our results, both within a single language and across two different languages. By employing this strategy, we will be able to increase the depth of our understanding of phonotactic knowledge while simultaneously reducing the uncertainty about inferences drawn from wordlikeness experiments. Specifically, we address the issues listed below, some of which will be answered through computational model comparison (questions 1–2), and others through statistical analysis (question 3):

1. To what degree do different computational models predict wordlikeness judgments in English and Korean?
2. To what degree is gradience in wordlikeness judgments related to probabilistic perception?
3. How do stimulus presentation and response modalities affect such judgments?

In our previously-submitted proposal, reviewers requested a narrower set of research questions, clearer motivations for our experimental design, and a clearer discussion of how hypotheses will be tested. We have attended to all of these concerns. We note that our proposed research is part of a longer arc that we plan to undertake; in this proposal, we focus on a connected set of questions we believe to both be coherent and essential for near-future progress in phonological theory.

2 Proposed project

The project has two informal phases. In the first phase, we will perform 6 carefully-controlled experiments to deepen our understanding of the methodological issues underlying wordlikeness judgment tasks. In the second phase, we will explore the computational underpinnings of wordlikeness judgments through model comparison. These model comparisons will allow us to probe the efficacy of different computational models that embody claims about the nature of the underlying grammar and the relationship between gradience in perception and wordlikeness judgments.

Our experiments target two languages: English and Korean. This choice is motivated not only by a general scientific desire to evaluate computational models on multiple languages (Bender 2009) but also because the phonotactic patterns of the two languages have important affordances for our experiments, as described below.

Mixed effects linear regression analyses will be used to test for effect of stimulus presentation and response modalities. A variety of computational models, described below, will be used to predict wordlikeness judgments, and associations between models' predicted acceptability and participant judgments will be measured using standard measures (discussed below).

2.1 Exp. 1a: Auditory Likert-scale acceptability (American English)

This experiment allows us to study the wordlikeness judgments of a large set of nonce words. It will also allow us to look at the performance of whole-word computational models when multiple phonotactic violations are present. Native American English speakers will be presented auditorily with a series of monosyllabic and bisyllabic nonce words, varying in the initial onset and the syllable contact (i.e., word-medial coda-onset) cluster. For each stimulus presented, we will collect Likert-scale wordlikeness judgments from native English speakers with the following prompt: "On a scale of 1–7, how good an English word do you think this is?” The results will be compared against the predictions of the different computational models (see subsection 2.8 for the models we consider). While this study is in some aspects a replication of prior research, it will also serve as the basis of comparison for Experiments 1b, 2, and 4 proposed below.

2.2 Exp. 1b: Auditory binary-choice acceptability (American English)

This experiment allows us to compare the binary response modality (Scholes 1966) to those obtained with a Likert scale. In this experiment, we will collect binary yes-no judgments from native English speakers using the same stimuli as Exp. 1a. Participants will be prompted: “Is this a possible word of English?” The results of this experiment will be compared with those of Experiment 1a to determine whether response modality biases wordlikeness judgments. As with Exp. 1a, these results will support computational modeling, including per-subject modeling.

2.3 Exp. 2: Auditory Likert-scale acceptability & transcription (American English)

This experiment allows us to determine whether gradience in perception affects wordlikeness judgments. Native American English speakers will be presented auditorily with monosyllabic nonce words that vary in the initial onset clusters. For each stimulus presented, there will be a Likert-scale rating task identical to Exp. 1a. Immediately after the Likert-scale judgment, the participants will be asked to transcribe the stimulus using English orthography. For analysis, we will convert these orthographic transcriptions to phonemic transcriptions using a neural network-based grapheme-to-phoneme conversion system (Ashby et al. 2021; Gorman et al. 2020) followed by manual correction of the resulting transcriptions. We will then test whether computational model predictions derived from the veridical (i.e., actual) or perceived (i.e., transcribed) stimulus are more predictive of participants’ judgments. Furthermore, we will be able to determine whether gradience in the wordlikeness responses stems in part from gradience in perceptual identification of the stimulus; we will measure whether acceptability ratings track rates of perceptual repair.

2.4 Exp. 3a: Auditory Likert-scale acceptability & transcription (Korean)

Exps. 3a–b will allow us to test the robustness of computational models across different languages. We specifically choose Korean because it has a different set of challenges for computational models than English. For example, unlike English, Korean does not have any onset clusters other than consonant-glide clusters. This therefore allows us to test the performance of computational models faced with fewer positive exemplars in the training data. Relatedly, Korean provides speakers with minimal evidence for the Sonority Sequencing Principle (SSP), a putative linguistic universal representing an implicational hierarchy for within-syllable structure (e.g., [tm] is a better onset than [mt]). Berent et al. (2008, 2009, 2007) claim native Korean speakers have knowledge about the SSP since they exhibit greater perceptual accuracy with onset clusters that satisfy the SSP (e.g., [bl, gl, …]) than those that violate the SSP (e.g., [lb, lg, …]). However, there is debate about the source of this effect, and Peperkamp (2007) and Yun (2016) argue the original perceptual results are marred by experimental confounds. Furthermore, Daland et al. (2011) present evidence that English speakers' wordlikeness judgments reflect knowledge of the SSP, and argue that such knowledge could indeed have arisen from exposure to English syllables, which largely obey the SSP. Our use of a transcription task will allow us to probe the acceptability of veridically perceived stimuli, and thus to test whether Korean speakers have knowledge about the SSP in the absence of sufficient evidence from their linguistic experience. This situation is quite different from English, in which Daland et al. (2011) argue that there is sufficient evidence in the lexicon for a learner to infer an SSP-like generalization. Finally, Korean has systematic restrictions on onset-vowel sequences (e.g., [ʃ] only occurs before [i]), and these experiments will allow us to test how well the relevant computational models perform in the face of such CV phonotactic interactions.

This experiment allows us to test (a) the influence of innate knowledge of sonority contributes to wordlikeness judgments of non-native clusters, (b) the influence of onset-vowel phonotactic violations on wordlikeness judgments, and (c) the influence of gradience in perceptual identification on gradience in wordlikeness judgments. As in Exps. 1–2, native Korean speakers will be presented auditorily with a series of monosyllabic and bisyllabic nonce words, varying in the initial onset and the syllable contact (i.e., word-medial coda-onset) cluster. For each stimulus presented, there will be a Likert-scale rating task prompting participants to do as follows: "1-7점 중에 들은 소리는 한국어 단어로 얼마나 적합합니까?" ('On a scale of 1–7, how good a Korean word do you think this is?') Immediately after the Likert-scale judgment, the participants will be asked to transcribe the stimulus using Hangul linearly using *pureo-sseugi* (lit. 'unassembled writing'), in which Hangul characters are written individually in a sequence such as in ㄱㅗㅁ instead of grouped by syllable as in 곰. This will allow the participants to transcribe ill-formed consonant clusters if they perceive them.

2.5 Exp. 3b: Auditory binary-choice acceptability & transcription (Korean)

This experiment allows us to compare the binary response modality to those obtained with a Likert scale, and will allow us to test whether the results of a similar methodological comparison with English speakers (Exps. 1a–1b) generalize to Korean. The stimuli will be the same as in Exp. 3a. Native Korean speakers will rate the wellformedness of each stimuli using a binary yes-no judgment. Participants will be prompted: "이 소리는 한국어로 가능한 단어입니까?" ('Is this a possible word of Korean?') They will then perform a transcription task following the procedures outlined in Exp. 3a to ensure that the only difference between Exp. 3a and this experiment is the response modality for the wordlikeness judgment.

2.6 Exp. 4: Orthographic Likert-scale acceptability (American English)

Orthographic presentations are sometimes seen as directly informing the participant of the representation to be judged (Daland et al. 2011; Pizzo 2015), bypassing the issue of probabilistic perception raised above. While this account is plausible, it is important to establish it experimentally, since it is alternatively possible that participants depend on orthotactic knowledge—i.e., knowledge of spelling rules—rather than purely phonotactic knowledge. In this final experiment, we will run a Likert-scale judgment task similar to Exp. 1a, except that the stimuli will be presented orthographically. The proposed experiment determines whether participants judge the input orthographic representation or if they use this representation to project an auditory percept. If the latter is what participants do, then the wordlikeness judgments should parallel those in Exp. 1a. Such a result would suggest that orthographic presentation is a reliable enough proxy of auditory presentation and can be used in future studies, particularly those conducted online.

2.7 General design and methods

For crosslinguistic comparison, we form stimuli from segments which (ignoring subtle subphonemic details of pronunciation) are found in both American English and Korean. In American English experiments, the stimulus set consists of two types of nonce words: (a) monosyllabic stimuli, manipulating the wellformedness of onsets, (b) bisyllabic stimuli, manipulating the wellformedness of the initial onset sequences and the wellformedness of word-internal syllable contact consonant sequences. In both languages, nonce words in which the onset and coda contain the same consonants (e.g., words of the form [sC_1VC_1]) will be removed; while words of this shape are found in English, Coetzee (2008) claims they are marked. Finally, any attested words will also be removed.

Monosyllabic stimuli will be of the form C(C)VC and will consist of an (singleton or cluster) onset followed by a vowel and an [n, s] coda. The inventory of onsets will include a wide variety of sonority contours, including falling sonority contours not attested in either language. Word-initial onsets for all stimuli will include singleton consonants—stops and fricatives, [s]-consonant clusters, obstruent-liquid clusters, stop-[w] clusters, stop-nasal clusters, and nasal-stop clusters. Of these, clusters consisting of a coronal obstruent followed by [l], and both stop-nasal and nasal-stop clusters, are all unattested in English, and some stop-[w] clusters have been labeled "marginal" (Daland et al. 2011, p. 204). In Korean, all but stop-[w] onset clusters are bad. This stimulus set will be used for both American English and Korean.

Bisyllabic stimuli will be of the form C(C)VC.CVC, with initial stress. They will consist of the preceding inventory of initial onsets followed by a vowel, a syllable contact cluster consisting of an [m, s] coda, a singleton stop onset, a vowel, and finally an [n, s] coda. These stimuli provide a chance to test the effect of cumulativity: in addition to the potential illformedness of the initial onset, nasal-obstruent syllable contact clusters in which the nasal is not homorganic with the obstruent—like […m.t…]—and hetero-voiced obstruent clusters—like […s.b…]— are both quite rare in English (Gorman 2013; Pierrehumbert 1994). Sample stimuli for all the experiments are presented in Table 1. No separate set of fillers are necessary, as a large portion of the nonce words—including all monosyllables with singleton onsets—are well-formed in both languages.

Singleton onsets	[p, t, …]
/s/-consonant onsets	[sk, sp, …]
obstruent-liquid onsets	[dl, tl, …]
stop-/w/ onsets	[kw, pw, …]
stop-nasal onsets	[tm, tn, …]
nasal-stop onsets	[mt, nt, …]

Table 1: Sample initial onsets for stimulus sets in the proposed experiments.

All the auditory stimuli will be recorded by a trained phonetician. For both English and Korean, we will use stratified sampling to select 200 stimuli balanced across all conditions (Table 1). We will use an American English speaker to record the stimuli for the English experiments, and a Korean speaker to record stimuli for the Korean experiments. We choose this approach because it is important to ensure that the participants genuinely believe that the stimuli are produced by a speaker of their native language, or they may adopt a different task model than intended (Schütze 2005). Consonant clusters will be produced with a medial vowel during the recording and the medial vowel will be spliced out in order to keep the stimuli as natural as possible, while controlling for any confounds resulting from producing illicit sequences of clusters. This allow us to create stimuli that are speakers cannot naturally produce and will allow us to maintain a degree of consistency across stimuli and languages.

Orthographic stimuli will be generated by converting phonemic transcriptions of the English auditory stimuli to orthographic form using a neural network phoneme-to-grapheme system, and then manually inspecting the resulting nonce words. The resulting stimuli will also be "back-converted" to phonemic form using a state-of-the-art grapheme-to-phoneme engine (Ashby et al. 2021; Gorman et al. 2020) to detect ambiguity in pronunciation.

Each experiment will recruit 120 participants. According to a heuristic proposed by statisticians (Lakens 2021; Simonsohn 2015), a sample size 2.5 times that of the original study should achieve 80% power. Daland et al. (2011), one of the larger prior studies and an obvious point of comparison for the proposed project, had 48 participants in their largest experiment, and thus $n = 2.5 \times 48 = 120$. We will also use more than twice as many stimuli per participant as Daland et al. (2011).

Therefore we are confident the proposed experiments are sufficiently sensitive. Note that the large data sets will support detailed hyperparameter search along the lines as Daland et al.'s experiments varying the UCLA Phonotactic Learner's grammar size (i.e., number of constraints posited).

We acknowledge that getting perfectly monolingual English and Korean speakers might be quite difficult even though recruiting will be done in the U.S. and Korea, respectively. We address this concern by administering instructions and prompts in the target language and ensuring that the experimenters do not in anyway indicate that the experiment is about another language. Prior perception experiments run with this general strategy produced consistent and interpretable results in which Korean and Mandarin speakers had systematically different from responses from English despite their exposure to English (Durvasula and Kahng 2015, 2016; Durvasula et al. 2018). In addition, we will also test and control for this issue by including standard measures of second language exposure as covariates in our statistical models. We further acknowledge that there are many alternative ways to construct stimuli and perform the experiments. First, we could have recorded stimuli using a speaker whose language allows all the relevant clusters. However, we believe that, in the absence of a careful study of the articulatory patterns of all the relevant sequences in the language and for that speaker, this would produce less controlled stimuli, as there could be differences in the co-ordination patterns of the consonantal sequences (Hermes, Mücke, and Grice 2013; Mücke, Hermes, and Tilsen 2020). Second, there are other experimental paradigms such as pairwise comparison or magnitude estimation that could be used here. However, we believe that the tasks we propose will produce results similar to those obtained with other paradigms, but will be more easily compared to prior work, will have greater sensitivity (Marty, Chemla, and Sprouse 2020; Sprouse 2011), and will be easier to statistically model (Daland et al. 2011). Third, we could use a nonce word repetition task instead of a transcription task, but this introduces a problem of coding the responses. Furthermore, it is possible that participants veridically perceive the stimuli but are unable to produce them faithfully. While we expect substantial variation even in orthographic transcription we have technical means to map variable orthography onto phonemic representations (Ashby et al. 2021; Gorman et al. 2020). Finally, while it is logically possible that different levels of lexical knowledge might affect word-likeness judgments, we known of no compelling evidence that variation in degree of lexical knowledge affects wordlikeness judgments made by adult native speakers. Indeed, some evidence suggests that a very small lexicon is sufficient to achieve native-like performance on wordlikeness judgments (Oh et al. 2020). We consequently do not control for this factor in the current series of experiments.

2.8 Computational modeling

All experiments will support computational modeling efforts using both aggregated and per-subject data. In what follows, we present the models we will use, and categorize them according to the type of underlying source of the wordlikeness judgment they embody:

- Gradient phonotactic grammars:
    1. The UCLA Phonotactic Learner (Hayes and Wilson 2008; Wilson and Gallagher 2018)
    2. Segmental *n*-gram language models (Albright 2009), implemented using OpenGrm-NGram (Roark et al. 2012)
    3. Segmental recurrent neural network-based language models (Mayer and Nelson 2020), implemented using Fairseq (Ott et al. 2019)
- Lexical similarity models:
    1. Coltheart's *N* (Coltheart et al. 1977), implemented using Pynini (Gorman 2016)
    2. PLD20 (Suárez et al. 2011), implemented using Pynini (Gorman 2016)
    3. The Generalized Context Model (Bailey and Hahn 2001)
- Categorical phonotactic models:
    1. The gross categorical phonotactic model (Gorman 2013)
    2. The simple categorical phonotactic model (Durvasula 2020)

We select these models so as to cover a wide variety of modeling techniques previously used for phonotactic modeling, though we can imagine an enormous number of other models drawn from the machine learning (e.g., regression trees) or computational cognitive modeling (e.g., naive discriminative learners) literature. The UCLA Phonotactic Learner and recurrent neural networks are sophisticated discriminative models that have been widely used for phonotactic modeling. The segmental *n*-gram model is a simple and effective grammatical model that tracks the product of the probability of subsequences of segments in the stimuli. The Generalized Context Model, Coltheart's *N*, and PLD20 all measure similarity to existing lexical items, operationalizing the hypothesis that gradience in judgments is related to lexical search. Gorman's (2013) categorical phonotactic model dichotomously partitions stimuli into those which only have attested onsets and rimes and those which do not; it is categorical because it is insensitive to type or token frequency. Durvasula's (2020) categorical phonotactic model similarly dichotomizes stimuli according to whether all featural and segmental *n*-grams are attested.

Digital pronunciation dictionaries will be used to train the models, including the CMU Pronouncing Dictionary (Weide 1994) and the CALLHOME Lexicon (Kingsbury et al. 1997) for American English, and WikiPron pronunciation dictionaries (Lee et al. 2020) for both languages.

We adjudicate between the models in three ways: (a) a gross fit for the overall data for each model using model fit measures such as the Spearman ρ rank correlation coefficient and the Akaike Information Criterion (Burnham and Anderson 2002)—this allows us to avoid issues of multicollinearity; (b) a more nuanced set of separate model comparisons for word-onsets acceptabilities and syllable contact acceptabilities; (c) a test of cross-linguistic validity of the models by separating modeling two different languages (Korean, and English). These three ways allow us to test the models both in terms of depth and breadth, lending robustness to the inferences.

3 Preliminary work

We present some preliminary work that suggests (a) that baseline computational models of the lexicon or categorical phonotactic grammars match human judgments at least as well as those that incorporate gradient phonotactic grammars, (b) that there is a need to study perceptual patterns more carefully in modeling and understanding gradience in wordlikeness judgments (subsection 3.2).

3.1 Modeling acceptability judgments with computational grammars

As mentioned above, Gorman (2013) reports that simple baseline models of wordlikeness were roughly as good as more complex gradient phonotactic models at predicting speakers' nonce word wordlikeness judgments. To replicate these results, we attempt to model wordlikeness judgment data, aggregated across participants, collected by four previous studies of English; these are the only studies of which we are aware for which stimulus-level rating data were made publicly available. In three of these studies (Albright 2007; Daland et al. 2011; Scholes 1966), the primary manipulation is the wellformedness of the initial onset; in the fourth (Albright and Hayes 2003), all words are expected to be "possible" words of English and have attested English onsets. Summary information for these data sets is presented in Table 2.

	# subjects	# items	response	presentation
Scholes (1966)	33	66	yes-no	auditory
Albright and Hayes (2003)	24	60	7-point scale	auditory
Albright (2007)	30	30	7-point scale	auditory
Daland et al. (2011)	48	96	6-point scale	orthographic

Table 2: Design of wordlikeness experiments used to test computational models.

We test four models, a subset of the computational models proposed in subsection 2.8.

- A featural maximum entropy ("UCLA Phonotactic learner") model, following the procedure laid out in Daland et al. (2011).
- A segmental trigram model, with Witten-Bell smoothing
- PLD20
- The gross categorical model (GC)

All models are trained on pronunciations from the CALLHOME American English lexicon (Kingsbury et al. 1997). For the maximum entropy model, we train on initial onsets only, and use the hyperparameters proposed by Hayes and Wilson (2008, §5). Models are scored by comparing their scores to the Spearman's ρ, a non-parametric measure of association; note that this choice makes the "temperature" hyperparameter introduced by Hayes & Wilson otiose. Higher scores indicate a greater association between human wordlikeness judgments and the model's score. The results are shown in Table 3. A few comments are in order. All four models are quite effective overall, and it is difficult to identify a clear "winner" on the basis of these results. The categorical model is inapplicable for the Albright and Hayes (2003) data, since it contains no gross phonotactic violations in the form of unattested onsets, and the UCLA Phonotactic learner also performs quite poorly on this data set. The PLD20 and trigram models are unexpectedly strong models across all four data sets.

	UCLA Phonotactic learner	Trigram	PLD20	GC
Scholes (1966)	.76	.79	.76	.77
Albright and Hayes (2003)	.12	.72	.71	n.a.
Albright (2007)	.84	.81	.62	.84
Daland et al. (2011)	.85	.80	.69	.74

Table 3: A comparison of four computational models across four data sets suggests that baseline models do as well as the UCLA Phonotactic learner, which represents a gradient phonotactic model. Values represent Spearman's ρ.

The fact that lexical and categorical models—as well as a simple trigram model—are competitive with the sophisticated UCLA Phonotactic learner casts further doubt on the utility of gradient phonotactic grammars. However, there are substantial limitations to this pilot experiment. First, the number of stimuli and subjects of each study are substantially smaller than the proposed experiments. Secondly, the data is only available in aggregated form, making per-subject modeling impossible. Third, though it is proposed as part of this project, results from recurrent neural network-based language models are not yet available, as such models require extensive hyperparameter tuning to achieve optimal performance. Fourth, the various studies used various presentation and response modalities. For example, Daland et al. (2011) used othographic stimulus presentation, and Scholes (1966) used a binary response modality. The proposed experiments will allow us to probe these effects in much greater detail than is currently possible, while avoiding the limitations mentioned above.

3.2 Studying acceptability and perception

In some preliminary work, we sought to study how the gradience in perception affects the wordlikeness judgments of nonce words. We conducted two experiments. Experiment 1 had 29 native Korean speakers who were in Daegu, South Korea, and 21 native English speakers in East Lansing, Michigan. The test items were nonce words of the form [$V_1C_1V_2$ma] where V_1 = [a, i, u], C_1 = [c, b], and V_2 = [i, ɯ, ∅]). A fully randomized list consisting of two repetitions of each test item was auditorily presented to the participants. We specifically looked at these [VCCV] sequences given they are phonotactically illicit in Korean (Sohn 1999), and have been observed to trigger illusory vowels (Durvasula and Kahng 2015, 2016; Kabak and Idsardi 2007).

In Experiment 1, for each stimulus, participants had to first identify the vowel between the two consonants, and then judge how good a word of their language it was on a 5-point scale. The instructions were given to the participants in their respective native languages. Note, the English participants were primarily included as controls for the identification task (results presented here), though they performed an equivalent acceptability task (results not presented here). In Experiment 2, we presented another 20 native Korean speakers in Daegu, South Korea with the same stimuli, but without the identification task, to test whether the identification task biased the wordlikeness judgments. The experiments were conducted using PsychoPy (Peirce et al. 2019).

The results for the identification task in Experiment 1, presented in Figure 1, were consistent with previous findings (Durvasula and Kahng 2015). A logistic mixed-effects model was fitted to the vowel identification of just the [VCma] nonce words, with random intercepts of participant and stimulus; the relevant statistical comparisons from this model are incorporated into the figure as pairwise comparisons. In post-labial contexts (i.e., C_1 = [b]), there was no clear evidence of the Korean speakers experienced more illusory [i] than English speakers, but they experienced more illusory [ɯ] (see Figure 1, left facet). In contrast, in post-palatal contexts (i.e., C_1 = [c]), Korean speakers experienced more illusory [i] and illusory [ɯ] than English speakers (see Figure 1, right facet). The results suggest that Koreans hear both illusory [i] and illusory [ɯ] after illicit palatal obstruents, but only illusory [ɯ] after illicit labial obstruents. These findings therefore replicate the results of Durvasula and Kahng (2015) and show that the illusory vowel phenomena reveal a role of native language phonological patterns in speech perception. More importantly for current purposes is the observation that, when the stimuli contained no medial vowels (i.e., in [VCma] stimuli), the Korean speakers overall perceived more illusory medial vowels after palatal contexts than after labial contexts ($\text{MeanDiff}_{\text{Pal}-\text{Lab}} = 49\%, p < .0001$).

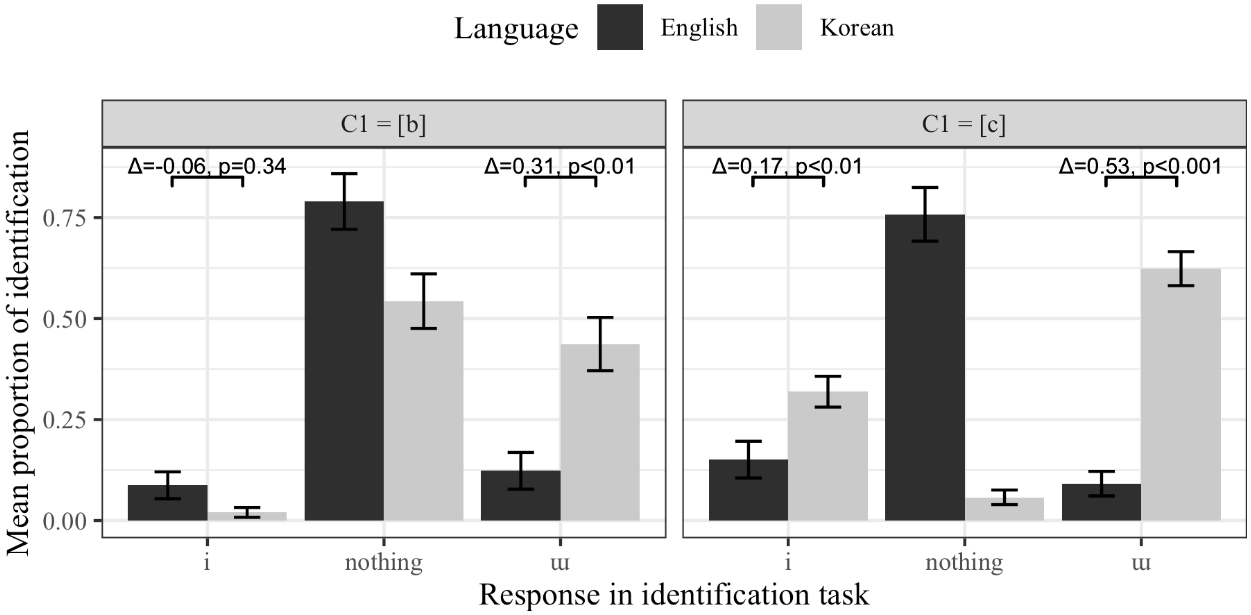

Figure 1: Illusions in vowel identification in [VCCV] stimuli for Korean, but not English, speakers. The error bars represent 95% confidence intervals.

Experiment 1 also allowed us to study the Korean wordlikeness judgments for different sequences. The results are presented in Figure 2. A linear mixed-effects model was fitted to the ratings with random intercepts of participants and stimulus. The relevant statistical comparisons from this model are incorporated into the figure as pairwise comparisons. For both the nonce words with palatal obstruents and those with labial obstruents, the Korean participants judged the illicit [VCma] nonce words worse than both the [VCima] nonce words and the [VCɯmV] nonce words (see Figure 2). Most important for our purposes is that amongst the [VCma] nonce words, the words with a palatal obstruent as the first consonant in the sequence were given higher scores than those with a labial obstruent as the first consonant ($\mathrm{MeanDiff}_{c-b} = 1.34, p < .0001$). Therefore, the acceptability results suggest a gradient: [Vbma] ≪ [Vcma] ≪ [VCVma]. [Vcma] were rated better than [Vbma] despite the fact that [c+m] sequences are phonotactically at least as bad as [b+m] sequences, and both are systematically absent from the language. In fact, labial nasals are possible in coda positions while palatal sounds are systematically absent from coda positions (Sohn 1999). This suggests that, at least based on positional featural licensing considerations, that palatal-consonant sequences are phonotactically *worse* than labial-consonant sequences more generally in the language. This makes the lower judgment for [Vbma] quite surprising at first. One gets further insight, however, by comparing the results of the acceptability task with those of the identification task. The comparison shows that the wordlikeness judgments of illicit sequences co-vary with the illusory vowel rates, i.e., parallel to the wordlikeness ratings, the illusory vowel perception rate with [Vcma] sequences is also generally higher than with [Vbma] sequences. These results suggest the important possibility that the observed gradience in wordlikeness judgments between [Vbma] and [Vcma] is better attributed to the gradient output of the perceptual system instead of the grammar.

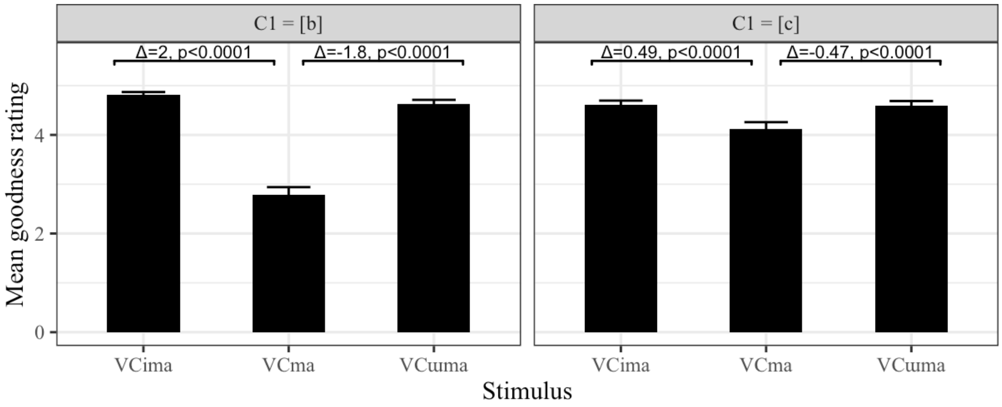

Figure 2: Gradient Korean wordlikeness judgments in Experiment 1: [Vbma] ≪ [Vcma] ≪ [VCVCV]. The error bars represent 95% confidence intervals, and the Δ indicates raw difference.

Experiment 2 shows the same pattern of wordlikeness results despite not having an identification task, as shown in Figure 3. This thereby confirms that the pattern of wordlikeness results observed for Korean speakers in Experiment 1 were not biased by the identification task.

Our results suggest that the wordlikeness judgments are based on perceived input, not the actual input, and that at least some of the gradience observed can be attributed to gradience in output of the perceptual system and not the grammar per se, i.e., if the perceptual output is a distribution of percepts, and the grammar is categorical over each percept, one should observe a gradient wordlikeness judgment. However, it is worth bearing in mind that the experiments involved a very small set of stimuli; therefore, the results are suggestive and in need of verification based on a more elaborate set of stimuli. Our proposed experiments do exactly this.

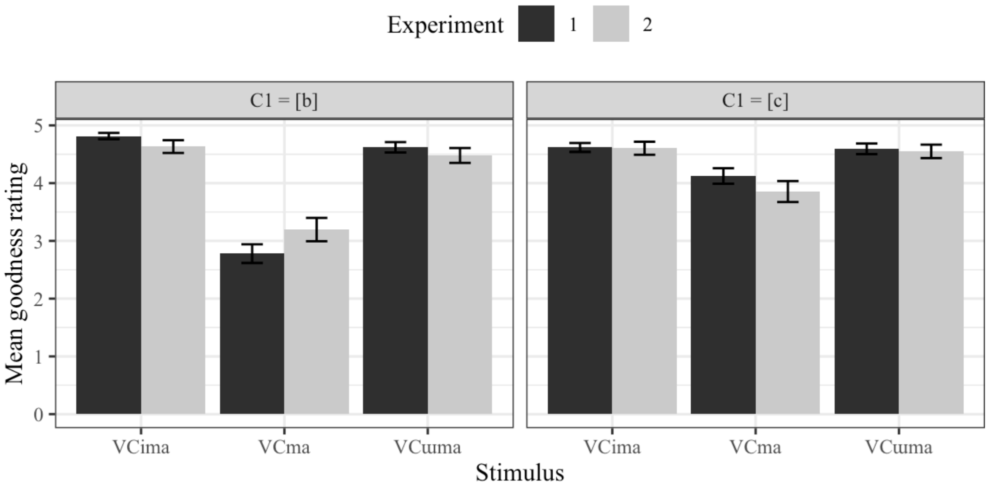

Figure 3: A comparison of Korean wordlikeness judgments in Experiments 1–2. The same acceptability cline is observed in both: [Vbma] ≪ [Vcma] ≪ [VCVma]. The error bars represent 95% confidence intervals, and Δ in figure represents raw difference.

3.3 Timeline

The project timeline is summarized in Table 4.

Fall 1	• Design Exp. 1a (English). • Record all audio stimuli. • Pilot perception experiments to test audio stimuli.
Spring 1	• Collect data for Exp. 1a. • Design Exps. 1b (English) and 3a (Korean). • Train graduate student to annotate typed responses.
Summer 1	• Annotate results for Exp. 1a. • Collect data for Exp. 1b in East Lansing, Michigan. • Collect data for Exp. 3a in Seoul, South Korea.
Fall 2	• Train graduate student to analyze experimental results. • Analyze Exps. 1a–b. • Design Exps. 2 (English) and 3b (Korean). • Outreach: Science Cafe in Oxford, Mississippi.
Spring 2	• Collect data for Exp. 2 in East Lansing, Michigan. • Design Exp. 4 (English). • Present results of Exp. 1a–b, and 3a at a conference. • Outreach: K-Core in Brooklyn, New York.
Summer 2	• Collect data for Exp. 4 in East Lansing, Michigan.

	• Collect data for Exp. 3b in Seoul, South Korea. • Implement computational models. • Outreach: Grandparents University in East Lansing, Michigan.
Fall 3	• Analyze Exps. 2, 4 and 3b. • Present results of Exps. 2, 4 and 3a–b at a conference. • Train and evaluate computational models.
Spring 3	• Submit paper on computational modeling to journal (expected: *Computational Linguistics*). • Submit paper on Korean experiments to journal (expected: *Frontiers in Psychology*).
Summer 3	• Submit paper on English experiments to journal (expected: *Glossa*). • Create open-access database.

Table 4: Project timeline.

4 Broader Impact

This proposal has the following broader impact goals: (a) mentorship of graduate students, (b) community outreach, and (c) release of data and software.

4.1 Mentorship of graduate students

An important aspect of the current proposal is the training and mentorship of two graduate students. The first graduate student will be working at Michigan State University for the first two years of the project and will receive mentorship on experimental design, administration and statistical analysis of behavioral results. This will be done primarily by Durvasula at the local institution, with help from Kahng and Gorman. The second graduate student will be working at the Graduate Center, City University of New York for the last two years of the project and will be mentored in computational modeling of the wordlikeness results. This will be done primarily by Gorman at the local institution, with help from Kahng and Durvasula. Both students will help in coordinating the research across the multiple institutions, and therefore they will also be trained to run a research lab in a streamlined and healthy fashion. Each graduate student will meet bi-weekly with the corresponding PI at their institution, where they will discuss background research, and relevant technical skills to implement the project. This training program will help us to develop better-prepared, well-adjusted researchers.

4.2 Community outreach

All three PIs will also engage in public outreach activities. Durvusula will hold a summer workshop at Grandparents University, an annual three-day summer camp for children ages 8–12 and their grandparents held at Michigan State University. This workshop will discuss this research project in a fashion that is accessible to a naive audience. Gorman will present the project to students enrolled in the Kingsborough Collaborative Research and Conference Bootcamp (K-CORE), a training program for science students at Kingsborough Community College in the process of transferring to senior colleges in the CUNY consortium. Kahng will present the project at Oxford Science Cafe, a monthly public meeting for community members 7–80 years old at the University of Mississippi. All three PIs have experience presenting research projects to non-specialists. Note that the proposed outreach activities are, by design, budget-neutral.

4.3 Release of data and software

The deidentified data, and software libraries for our computational modeling will be made available under a permissive open-source license via the GitLab source control system. Furthermore, we will also develop a website—akin to the English Lexicon Project (Balota et al. 2007) website—providing external scholars quick access to the trial-level data from all experiments. This website will use SQLite databases, the Flask web framework, and the Gunicorn web server. Gorman has the required experience as an software engineer to ensure the software and data will remain widely accessible many years after project completion. We anticipate this data and software will become a benchmark for future researchers looking to test hypotheses about wordlikeness judgments, speech perception, and phonotactic knowledge in general.

5 Intellectual Merit

While there has been a lot of attention paid to the possibility of gradient linguistic knowledge in recent years, research on phonotactic knowledge (Durvasula 2020; Gorman 2013) and syntactic knowledge (Sprouse et al. 2018) suggests that simple categorical grammatical models often predict human judgments accurately. Therefore, there is a need for a more thorough comparison of relevant computational models in order to better understand the nature of the underlying grammatical system. Our proposal targets this with explicit comparison of multiple computational models, but it will also allow the research community to develop more sophisticated computational models in the future. Existing computational models make at best *ad hoc* use of the rich representational structure posited by phonologists (Clements 1985; Clements and Keyser 1983; Hayes 1980; McCarthy 1979; Nespor and Vogel 1986, among others), and we anticipate these rich structures will become increasingly important as our understanding of wordlikeness becomes clearer.

Furthermore, as highlighted earlier, any behavioral observation is the result of many interacting factors. This is true of wordlikeness judgments too and consequently, gradient responses may occur regardless of the nature of the underlying grammar and task model used to make them. For this reason it is important to separate the different factors influencing wordlikeness judgments. In our proposal, we attempt to study known effects, namely, perception and lexical search, in order to better study the nature of the underlying phonotactic grammar. Finally, the proposal also attempts to study the effect of presentation and response modalities. An understanding along both dimensions is necessary for researchers to design more probative experiments in the future. We believe the above issues are important for the field to explore in a systematic fashion and obtain a better understanding of acceptability judgment tasks and their relationship to grammatical knowledge.

6 Results from Prior NSF support

None of the current grant applicants have had prior NSF support in the last 5 years.

References

Albright, A. (2007). *Natural classes are not enough: Biased generalization in novel onset clusters* [Ms., Massachusetts Institute of Technology].

Albright, A. (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology*, *26*(1), 9–41.

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

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, *13*(3), 263–308.

Ashby, L. F. E., Bartley, T. M., Clematide, S., Del Signore, L., Gibson, C., Gorman, K., Lee-Sikka, Y., Makarov, P., Malanoski, A., Miller, S., Ortiz, O., Raff, R., Sengupta, A., Seo, B., Spektor, Y., & Yan, W. (2021). Results of the second SIGMORPHON shared task on multilingual grapheme-to-phoneme conversion. *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 115–125.

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, *44*(4), 568–591.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459.

Bender, E. M. (2009). Linguistically naïve != language independent: Why NLP needs linguistic typology. *EACL Workshop on the Interaction Between Linguistics and Computational Linguistics*, 26–32.

Berent, I., Lennertz, T., Jun, J., Moreno, M., & Smolensky, P. (2008). Language universals in human brains. *Proceedings of the National Academy of Sciences*, *105*, 5321–5325.

Berent, I., Lennertz, T., Smolensky, P., & Vaknin-Nusbaum, V. (2009). Listeners' knowledge of phonological universals: Evidence from nasal clusters. *Phonology*, *26*(1), 75–108.

Berent, I., Steriade, D., Lennertz, T., & Vaknin, V. (2007). What we know about what we have never heard: Evidence from perceptual illusions. *Cognition*, *104*(3), 591–630.

Breiss, C. (2020). Constraint cumulativity in phonotactics: Evidence from artificial grammar learning studies. *Phonology*, *37*(4), 551–576.

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.

Chomsky, N., & Halle, M. (1965). Some controversial questions in phonological theory. *Journal of Linguistics*, *1*(2), 97–138.

Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English*. Harper; Row.

Clements, G. N. (1985). The geometry of phonological features. *Phonology Yearbook*, *2*, 225–252.

Clements, G. N., & Keyser, S. J. (1983). *CV Phonology: A Generative Theory of the Syllable*. MIT Press.

Coetzee, A. W. (2008). Grammaticality and ungrammaticality in phonology. *Language*, *84*, 218–257.

Coetzee, A. W., & Pater, J. (2008). Weighted constraints and gradient restrictions on place cooccurrence in Muna and Arabic. *Natural Language & Linguistic Theory*, *26*(2), 289–337.

Coleman, J., & Pierrehumbert, J. B. (1997). Stochastic phonological grammars and acceptability. *Proceedings of the 3rd Meeting of the ACL Special Interest Group in Computational Phonology*, 49–56.

Coltheart, M., Davelaar, E. J., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and Performance VI* (pp. 535–555). Lawrence Erlbaum.

Daland, R., Hayes, B., White, J., Garellek, M., Davis, A., & Norrmann, I. (2011). Explaining sonority projection effects. *Phonology*, *28*(2), 197–234.

Daland, R., Oh, M., & Davidson, L. (2019). On the relation between speech perception and loanword adaptation. *Natural Language & Linguistic Theory*, *37*(3), 825–868.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance*, *25*(6), 1568–1578.

Durvasula, K. (2020). *O gradience, whence do you come?* [Plenary talk at the 2020 Annual Meeting on Phonology, Santa Cruz, USA].

Durvasula, K., & Kahng, J. (2015). Illusory vowels in perceptual epenthesis: The role of phonological alternations. *Phonology*, *32*(3), 385–416.

Durvasula, K., & Kahng, J. (2016). The role of phrasal phonology in speech perception: What perceptual epenthesis shows us. *Journal of Phonetics*, *54*, 15–34.

Durvasula, K., & Liter, A. (2020). There is a simplicity bias when generalising from ambiguous data. *Phonology*, *37*(2), 177–213.

Durvasula, K., Huang, H.-H., Uehara, S., Luo, Q., & Lin, Y.-H. (2018). Phonology modulates the illusory vowels in perceptual illusions: Evidence from Mandarin and english. *Laboratory Phonology*, *9*(1), 7.

Frisch, S. A., Pierrehumbert, J. B., & Broe, M. B. (2004). Similarity avoidance and the OCP. *Natural Language & Linguistic Theory*, *22*(1), 179–228.

Gaskell, M. G., & Marslen-Wilson, W. D. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(1), 144–158.

Gorman, K. (2013). *Generative phonotactics* (PhD dissertation). University of Pennsylvania.

Gorman, K. (2016). Pynini: A Python library for weighted finite-state grammar compilation. *Proceedings of the ACL Workshop on Statistical NLP and Weighted Automata*, 75–80.

Gorman, K., Ashby, L. F. E., Goyzueta, A., McCarthy, A. D., Wu, S., & You, D. (2020). The SIGMORPHON 2020 shared task on multilingual grapheme-to-phoneme conversion. *17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 40–50.

Gow, D. (2003). Feature parsing: Feature cue mapping in spoken word recognition. *Perception & Psychophysics*, *65*(4), 575–590.

Greenberg, J. H., & Jenkins, J. J. (1964). Studies in the psychological correlates of the sound system of American English. *Word*, *20*(2), 157–177.

Halle, M. (1962). Phonology in generative grammar. *Word*, *18*(1), 54–72.

Halle, M., & Stevens, K. (1962). Speech recognition: A model and a program for research. *IEEE Transactions on Information Theory*, *8*(2), 155–159.

Hay, J., Pierrehumbert, J. B., & Beckman, M. E. (2004). Speech perception, well-formedness and the statistics of the lexicon. In J. Local, R. Ogden, & R. A. M. Temple (Eds.), *Phonetic*

Interpretation: Papers in Laboratory Phonology VI (pp. 58–74). Cambridge University Press.

Hayes, B. (1980). *A metrical theory of stress rules* (PhD dissertation). Massachusetts Institute of Technology.

Hayes, B. (2000). Gradient well-formedness in Optimality Theory. In J. Dekkers, F. van der Leeuw, & J. van de Weijer (Eds.), *Optimality Theory: Phonology, Syntax, and Acquisition* (pp. 88–120). Oxford University Press.

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

Hermes, A., Mücke, D., & Grice, M. (2013). Gestural coordination of Italian word-initial clusters: The case of 'impure s'. *Phonology*, *30*(1), 1–25.

Kabak, B., & Idsardi, W. J. (2007). Perceptual distortions in the adaptation of English consonant clusters: Syllable structure or consonantal contact constraints? *Language and Speech*, *50*(1), 23–52.

Kingsbury, P., Strassel, S., McLemore, C., & MacIntyre, R. (1997). *CALLHOME American English Lexicon (PRONLEX)* [LDC97L20].

Lakens, D. (2021). *Sample size justification* [PsyArXiv preprint]. doi: 10.31234/osf.io/9d3yf

Lee, J. L., Ashby, L. F., Garza, M. E., Lee-Sikka, Y., Miller, S., Wong, A., McCarthy, A. D., & Gorman, K. (2020). Massively multilingual pronunciation mining with WikiPron. *Proceedings of the 12th Language Resources and Evaluation Conference*, 4223–4228.

Marty, P., Chemla, E., & Sprouse, J. (2020). The effect of three basic task features on the sensitivity of acceptability judgment tasks. *Glossa*, *5*(1), 72.

Mayer, C., & Nelson, M. (2020). Phonotactic learning with neural language models. *Proceedings of the Society for Computation in Linguistics*, 291–301.

McCarthy, J. J. (1979). *Formal problems in Semitic phonology and morphology* (PhD dissertation). Massachusetts Institute of Technology.

McCarthy, J. J., & Prince, A. (1986). *Prosodic morphology* (tech. rep. RuCCS-TR-32). Rutgers University Center for Cognitive Science.

Mitterer, H., Kim, S., & Cho, T. (2013). Compensation for complete assimilation in speech perception: The case of Korean labial-to-velar assimilation. *Journal of Memory and Language*, *69*(1), 59–83.

Mücke, D., Hermes, A., & Tilsen, S. (2020). Incongruencies between phonological theory and phonetic measurement. *Phonology*, *37*(1), 133–170.

Nespor, M., & Vogel, I. (1986). *Prosodic Phonology*. Foris.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *30*(2), 1113–1126.

Oh, Y., Todd, S., Beckner, C., Hay, J., King, J., & Needle, J. (2020). Māori-speaking New Zealanders have a Māori proto-lexicon. *Scientific Reports*, *10*(1), 1–9.

Ohala, J. J., & Ohala, M. (1978). Testing hypotheses regarding the psychological manifestation of morpheme structure constraints. *Report of the Phonology Laboratory (Berkeley)*, *2*, 156–165.

Ohala, J. J., & Ohala, M. (1986). Testing hypotheses regarding the psychological manifestation of morpheme structure constraints. In J. J. Ohala & J. J. Jaeger (Eds.), *Experimental Phonology* (pp. 239–252). Academic Press.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., & Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 48–53.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*, 195–203.

Peperkamp, S. (2007). Do we have innate knowledge about phonological markedness? Comments on Berent, Steriade, Lennertz, and Vaknin. *Cognition*, *104*(3), 631–637.

Pierrehumbert, J. B. (1994). Syllable structure and word structure: A study of triconsonantal clusters in English. In P. A. Keating (Ed.), *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III* (pp. 168–188). Cambridge University Press.

Pierrehumbert, J. B. (2001). Why phonological constraints are so coarse-grained. *Language and Cognitive Processes*, *16*(5–6), 691–698.

Pierrehumbert, J. B. (2003). Probabilistic phonology: Discrimination and robustness. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic Linguistics* (pp. 177–228). MIT Press.

Pizzo, P. (2015). *Investigating properties of phonotactic knowledge through web-based experimentation* (Ph.D. Dissertation). University of Massachusetts, Amherst.

Roark, B., Sproat, R., Allauzen, C., Riley, M., Sorensen, J., & Tai, T. (2012). The OpenGrm open-source finite-state grammar software libraries. *Proceedings of the ACL 2012 System Demonstrations*, 61–66.

Scholes, R. J. (1966). *Phonotactic Grammaticality*. Mouton.

Schütze, C. T. (1996). *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press.

Schütze, C. T. (2005). Thinking about what we are asking speakers to do. In S. Kepser & M. Reis (Eds.), *Linguistic Evidence* (pp. 457–484). De Gruyter.

Schütze, C. T. (2011). Linguistic evidence and grammatical theory. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*, 206–221.

Shademan, S. (2006). Is phonotactic knowledge grammatical knowledge? *Proceedings of the 25th West Coast Conference on Formal Linguistics*, 371–379.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569.

Sohn, H.-M. (1999). *The Korean Language*. Cambridge University Press.

Sprouse, J. (2011). A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language*, *87*(2), 274–288.

Sprouse, J., Yankama, B., Indurkhya, S., Fong, S., & Berwick, R. C. (2018). Colorless green ideas do sleep furiously: Gradient acceptability and the nature of the grammar. *The Linguistic Review*, *35*(3), 575–599.

Stanley, R. (1967). Redundancy rules in phonology. *Language*, *43*(2), 393–436.

Suárez, L., Tan, S. H., Yap, M. J., & Goh, W. D. (2011). Observing neighborhood effects without neighbors. *Psychonomic Bulletin & Review*, *18*, 605–611.

Valian, V. (1982). Psycholinguistic experiment and linguistic intuition. In T. W. S. Simon & R. J. Scholes (Eds.), *Language, Mind, and Brain* (pp. 179–188). Erlbaum.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science*, *9*(4), 325–329.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood density in spoken word recognition. *Journal of Memory and Language*, *40*(3), 374–408.

Vitevitch, M. S., Luce, P. A., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and syllable stress: Implication for the processing of spoken nonsense words. *Language and Speech*, *40*(1), 47–62.

Weide, R. L. (1994). *CMU Pronouncing Dictionary* [URL: http://www.speech.cs.cmu.edu/cgi-bin/cmudict].

Wilson, C., & Gallagher, G. (2018). Accidental gaps and surface-based phonotactic learning: A case study of South Bolivian Quechua. *Linguistic Inquiry*, *49*(3), 610–623.

Yun, S. (2016). *The sonority sequencing principle in consonant cluster perception revisited* [Paper presented at the CUNY Phonology Forum].