Can you judge what you don't hear? Perception as a source of gradient wordlikeness judgements*

Word count ≈ 11,250 (including abstract, figures, tables and references)

Abstract

A key observation about wordlikeness judgements, going back to some of the earliest work on the topic is that they are gradient in the sense that nonce words tend to form a cline of acceptability. In recent years, such gradience has been modelled as stemming from a gradient phonotactic grammar or from a lexical similarity effect. In this article, we present two experiments that suggest that at least some of the observed gradience stems from gradience in perception. More generally, the results raise the possibility that the gradience observed in wordlikeness tasks may not come from a gradient phonotactic/phonological grammar.

Keywords: Acceptability Judgements, Perception, Gradience, Korean, English.

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) observe certain unattested 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" judgements, *i.e.*, acceptability judgements of nonce words, and such studies increasingly include sophisticated computational modeling

*This article was made possible by the help and support of many individuals. We would like to thank: first and foremost, the associate editor and the anonymous reviewers for valuable criticism that helped to improve the article greatly; second, the members of the phonology-phonetics group at Michigan State University for helpful discussions; and finally, the audiences of the Annual Meeting on Phonology 2019.

of the judgements 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 judgements, 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 judgements across students, he discerned a cline of acceptability. In past work, some researchers have proposed that gradience reflects similarity to existing words, perhaps because the judgement task makes use of lexical search (Bailey and Hahn 2001; Ohala and Ohala 1978, 1986; Vitevitch and Luce 1998, 1999; Vitevitch et al. 1997). Alternatively, more recently, many researchers have explored the possibility 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). For example, the fact that the input [blik] is typically judged better than [bnik] which in turn is typically judged better than [bzik] is modelled within the grammar through gradient generalisations.[1] One way to model it within a gradient grammar is in the form of increasing weights associated with the statements *bl, *bn, and *bz, where a higher weight is associated with higher ungrammaticality and lower actual frequency of occurrence of structural factors (such as, featural combinations) in real words, and the grammar would assign a higher weight to words where the structural factors within them have higher weights. A second way to model the same fact within a

[1]Note, in the interest of not distracting the reader from the main point of the paper, we only provide a very high-level description of such gradient models. We refer the reader to the cited work and to Durvasula and Liter (2020) for more in-depth discussion of the relevant models.

gradient grammar is to consider phonotactic statements about possible sequences within the grammar as being attached to probabilities (again, reflecting actual frequency of occurrence of structural factors in real words), and the grammar would assign lower probabilities to words where the structural factors within them have lower probabilities.[2]

While one can obviously see the merits of such a strategy, we would like to raise awareness here that the approach of *directly* modelling gradient acceptability judgements as stemming from gradient generalisations within a grammar presupposes that a substantial portion of (if not all) the observed gradience in acceptability judgements stems from the grammar. As we will argue in this paper, this presupposition is incorrect.

Despite the dominance of gradient phonotactic knowledge or lexical similarity as the basis of gradient wordlikeness judgements in current thinking, there are likely many other interacting factors that affect such judgements. Indeed, linguists and psychologists have long recognised that all linguistic performance is a complex, multifactorial phenomenon (Chomsky 1965; Schütze 1996, 2011; Valian 1982). Therefore, we should be exploring how these factors affect acceptability judgements in tandem with the proposed grammar. Note, our larger point is not that such gradient models are inadequate to account for the observed acceptabilities (some of them, such as the Hayes and Wilson (2008) UCLA Phonotactic Learner, are often very powerful weighted grammars and can sometimes fit observed data extremely well) — it is more that such complex models might be unnecessary to account for the observed gradience, *if we better factor out the sources that affect acceptability judgements*. Therefore, such gradient grammatical models, while sometimes effective in accounting for observed data, might not be an accurate reflection of the underlying computations made by human beings.

One such additional factor on wordlikeness judgements is the perceptual system itself. There is now a clear consensus that the perceptual system does not result in a single categorical percept, as was thought before (Liberman et al. 1957), but has a gradient response (Massaro and Cohen 1983; McMurray et al. 2008; Norris and McQueen 2008; Norris, McQueen, and Cutler 2003, amongst

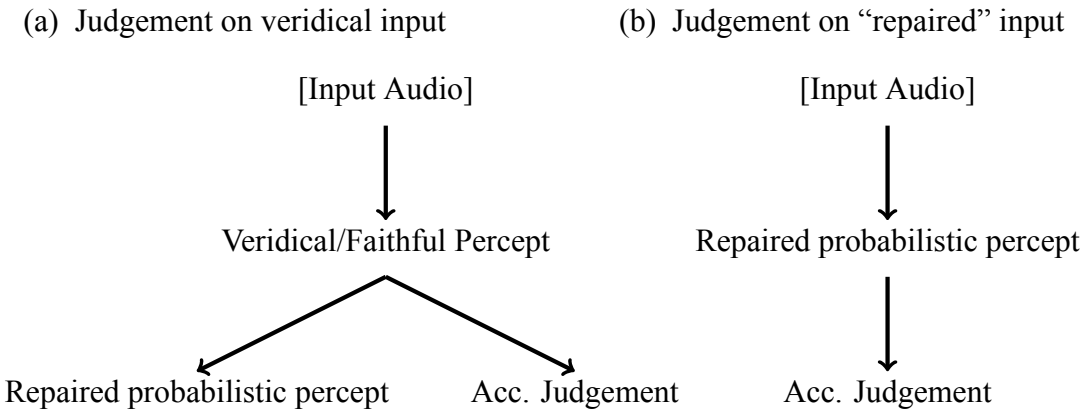
[2] As has been noted right from the earliest such attempts (Coleman and Pierrehumbert 1997), both of these strategies have the problematic consequence that longer words are always expected to be less acceptable than shorter words.

others). Note, however, the observation of a gradient response need not be analysed as the perceptual system recovering a gradient representation; in fact, there is robust research arguing it is possible to model the gradient response of the perceptual system as the product of recovering a probability distributions over possible categorical representations of the acoustic input (Caplan 2021; Feldman and Griffiths 2007; Kleinschmidt and Jaeger 2015; Norris and McQueen 2008; Norris, McQueen, and Cutler 2003; Sonderegger and Yu 2010, amongst others). More importantly for the purposes of the paper, given the possibility of gradience in the perceptual responses, some of the observed gradience in wordlikeness judgements may reflect this gradience.

Relatedly, stimuli which violate native language phonotactic generalizations have been observed to induce perceptual illusions (Dupoux et al. 1999; Hallé et al. 1998; Kabak and Idsardi 2007; Moreton 2002). For instance, Japanese speakers perceive /ebɯzo/ when [ebzo] is presented, likely because consonant clusters like [bz] are illicit in Japanese.[3] The preceding observations raise a question relevant to wordlikeness judgements: just how do speakers make judgements of unacceptability, if unacceptable stimuli are those which induce perceptual illusions, *i.e.*, those which are often unfaithfully perceived and perceptually "repaired"? Two possibilities suggest themselves: either (a) acceptability judgements are made before the perceptual repair, making them independent of the repair, or (b) the acceptability judgement is made after perceptual repair, making them contingent upon the post-repair percept(s). These two possibilities are schematised in (1). The second possibility would pose serious problems for the study of phonotactic knowledge via wordlikeness judgements, as such judgements would be based on a representation that has already been "repaired" by phonological knowledge to conform to the patterns in the language, *i.e.*, as per the second possibility, the grammar affects the actual percept itself, so it is not quite clear why a judgement would involve the use of the grammar *again* on the perceived input.

[3] In talking about illusions, we will discuss them as categorical and singular percepts so as to be consistent with the exposition of the issue in prior research in the field. However, in line with the main thrust of the article, we think that the illusion itself results in a gradient response from the perceptual system, and not a single categorical percept. Furthermore, as pointed out above, some important recent work suggests that the gradient response of the perceptual system is more likely to be a probability distribution over possible categorical representations of the acoustic input.

1. Two possibilities for how perception and acceptability judgements interact



Consistent with the view (in 1b), we will argue that wordlikeness judgements are in fact based on perceived (and "repaired") representations.

Furthermore, the issue of non-veridical (and gradient) perception is even present with the perception of *licit CV sequences* of nonce words. For instance, the classic Miller and Nicely (1955) study shows that there is a lot of variability in responses of licit monosyllabic Cɑ nonce words (even in the high signal-noise ratio condition), with some CV sequences showing far more variable identification than others. Therefore, there is a gradience in perception that is not equitably distributed across all sequences. However, typical word-acceptability modelling proceeds with the modelling assumption that the acceptability judgement is over a percept that is identical to the input stimulus, and that there is no bias or variance introduced by the perceptual system. Consequently, such models that implement gradient phonotactic knowledge attempt to account for the observed gradience purely through the grammar. In contrast to this assumption, we will argue that at least some of the gradience observed in such judgements stems from gradience in the perceptual system.

Of course, there are many other possible factors that influence both wordlikeness judgements and the gradience observed in such tasks—task effects (Armstrong, Gleitman, and Gleitman 1983; Schütze 2005), prosodic factors such as 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 judgements, but

we leave a systematic exploration of such factors to future work.

In this article, we focus on Korean to show that the gradience in acceptabilities, at least partly, stems from perceptual gradience. We specifically focus on Korean as there is a fair bit of work on how specific unobserved consonantal sequences trigger perceptual illusion (Berent et al. 2007; Durvasula and Kahng 2015, 2016; Kabak and Idsardi 2007; Yun 2016), and we utilise these illusory vowel patterns to explore the role that perception plays in wordlikeness judgements. In Experiment 1 (Section 2), we will show using a combined wordlikeness judgement and identification task that: (a) the acceptability of nonce words with such sequences is *contingent* on the presence/absence of an illusory vowel; which thereby suggests that such wordlikeness judgements are made based on the "repaired" perceived stimulus, not on a veridical percept of the acoustic input; (b) the acceptability of the nonce words depends on the rate at which illusory vowels are perceived in perceptual experiments across participants, which suggests that at least some of the gradience can be traced back to the perceptual system. In Experiment 2 (Section 3), we will show that the overall patterns of acceptability judgements observed in Experiment 1 were not due to a task effect related to the presence of an identification task as part of the experiment.

2 Experiment 1

As mentioned above in the *Introduction*, Korean is the focal language in this article. The language has no obstruent coda consonants word-internally that are not part of a geminate sequence (Sohn 1999). Therefore, word-medial sequences such as [*cm, *cn, *jm, *jm, *bn, *bm, *pm, *pn…] are not observed in the language. Furthermore, such word-medial sequences have been observed to trigger illusory vowels (Berent et al. 2007; Durvasula and Kahng 2015, 2016; Kabak and Idsardi 2007). More specifically, it's been observed that palatal obstruent stop+nasal consonant sequences such as [cm, cn] trigger both illusory [i] and illusory [ɨ], while labial stop+nasal consonant sequences such as [bm, bn] trigger illusory [ɨ] (Durvasula and Kahng 2015, 2016; Kabak and Idsardi 2007).

In order to understand the influence of perception on the gradience observed in acceptability

judgements, we collected acceptability ratings and perceptual identification information simultaneously from Korean participants. Furthermore, to establish that the Korean listeners are indeed hearing perceptual illusions based on their linguistic knowledge and not based on experimental/stimulus artefacts, we used American English listeners as a control group.[4] American English does allow word-medial obstruent+nasal consonant sequences, and consequently, such speakers have not been observed to hear perceptual illusions at the same rate as Korean listeners for such sequences (Durvasula and Kahng 2015, 2016; Kabak and Idsardi 2007). Note, the American English listeners are part of this experiment solely to replicate prior results that Korean listeners do hear illusory vowels in the target stimuli. Consequently, while they participate in the same experiment as the Korean participants (except for differences in instruction language and prompts), only their perceptual identification results will be presented here to establish the perceptual illusions of the Korean listeners.

2.1 Methods

2.1.1 Participants

Twenty-nine native Korean speakers (18 women and 11 men, age: 19-25 years) and 21 native American English speakers (14 women and 7 men, age: 19-30 years) participated in the experiment. All the Korean speakers were recruited in Daegu, South Korea. All the English speakers were recruited in Michigan, United States. The Korean participants were compensated for their participation and the English participants received extra credit for their participation.

2.1.2 Stimuli

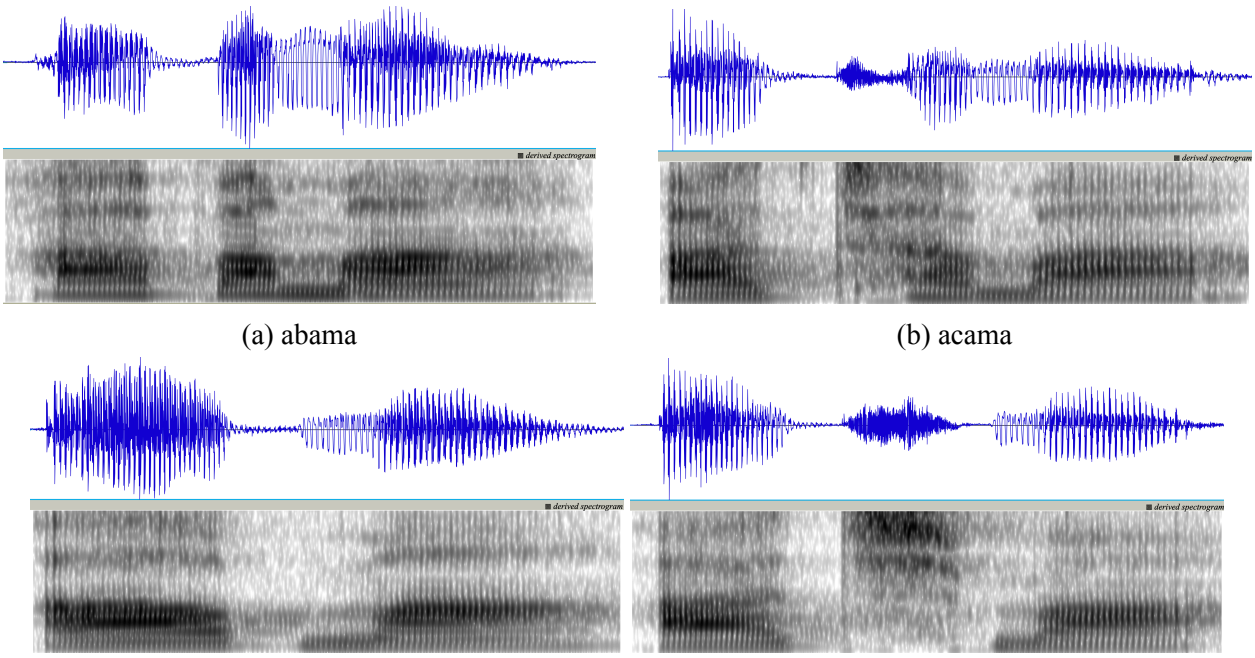
Forty-eight nonce words of the form [$V_1C_1V_2C_2$a] were used, in which V_1 was [a, i, u], C_1 was [c, b], V_2 was [i, ɨ, ∅ (null)], and C_2 was [m, n] (3*2*3*2 = 36). There were another 12 nonce word distractor items that had the template [$V_1C_1C_2$a], in which V_1 was [a, i, u]; C_1 was [m, n,

[4]The inclusion of a control group is now standard practice in the study of auditory illusions to guard against stimulus/experimental artefacts.

ŋ]; C_2 was [b, d, g, t]. Note, the distractor items were present to also ensure that there were clear examples of stimuli without a medial vowel (V_2) that did not violate any native phonotactics. Each of the 48 nonce words had two different recordings (total number of stimuli = 48 * 2 = 96). None of the stimuli were words in Korean or English. They had stress on the first vowel, and were naturalistic recordings by one of the authors, who is a female Korean-English bilingual. All items were recorded using Praat (Boersma and Weenink 2016), with a Blue Snowball USB microphone (frequency response 40 Hz–18 kHz) at a 44.1 kHz sampling rate (16-bit resolution; 1 channel). The stimuli were normalized in Praat to have a mean intensity of 65 dB SPL.

Given the rather large set of stimuli, we are unable to present the waveforms and spectrograms for all the stimuli. However, in Figure 1, we present waveforms and spectrograms of the stimuli [abama, acama, abma, acma]. In order to ensure that we did not present a biased sample, we chose to present the alphabetically first sample of each crucial case.[5] A reviewer wondered if the VCCa stimuli had strong burst releases for the first consonant [C]. In our stimulus selection and pronunciations, we took care to ensure there were no such clear bursts, and in fact, as can be seen in the figure, the VCma stimuli (1c-1d) did not have any observable burst releases that could trigger illusory vowels.

[5]The reader can find all the sound files, experiment files, and result files for all the experiments reported in this paper at the following Open Science Foundation repository: https://osf.io/e5hgz/.



(a) abama (b) acama

(c) abma (d) acma

Figure 1: Waveforms and spectrograms for 4 sample stimuli [abama, acama, abma, acma]

2.1.3 Procedure

The participants completed an identification task and an acceptability judgement task, which were scripted in Praat. First, the participants were asked to listen to a stimulus and determine the medial vowel by clicking on the corresponding box on the screen ('i', 'u'[6], or 'no vowel' for the English participants; `이', `으', or `없음' for the Korean participants). Immediately after identifying the medial vowel, the participants made an acceptability judgement on the stimulus they heard on a scale of 1-5 (1: Impossible / 불가능함, 5: Possible / 가능함). All the instructions were in English for the English speakers ('Choose the vowel between the two consonants, and then choose if the word is a possible word in English.') and were in Korean for the Korean speakers (`우선 두 자음 사이의 모음을 고르세요. 그 다음, 들은 단어가 한국어로 가능한 단어인지 고르세요.'). Before the actual experiment, each participant completed a practice session to ensure familiarity with the task. The practice session had 12 trials with a different set of nonce words. The intertrial interval

[6]As pointed out by Durvasula and Kahng (2015, 2016) and Kabak and Idsardi (2007), [ɨ] is very likely perceived as /ʊ/ by English listeners, and <u> is a good orthographic representation of the vowel.

was 1000 ms. All the trials were randomized for each participant. The experiment was conducted individually in a quiet room with a low-noise headset.

2.2 Results

In this article, we used the `tidyverse` suite of packages (Wickham 2017) for data analysis and plotting. The statistical analysis was done in base R (R Core Team 2021) using the packages `lme4` (Bates et al. 2015) and `lmerTest` (Kuznetsova, Brockhoff, and Christensen 2017). Finally, the statistical models were converted into LaTeX code using the R-package `stargazer` (Hlavac 2018).

2.2.1 General Acceptability Judgements

We first looked at the overall acceptability scores by the Korean listeners (Figure 2). As expected, for both the sets with the bilabial stop [b] and those with the palatal stop [c], the stimuli with no medial vowel were rated lower than those with a medial vowel.[7]

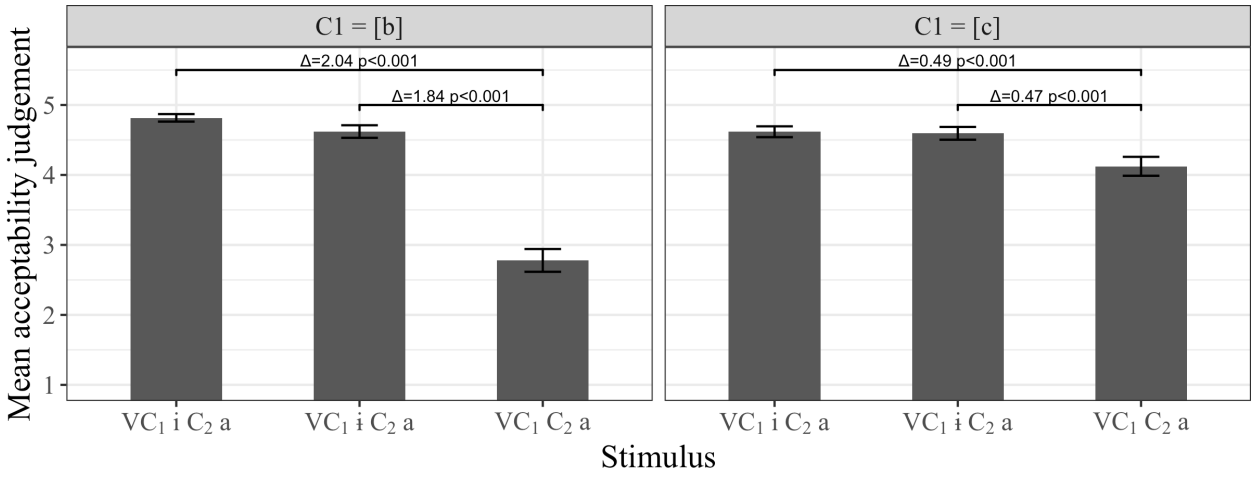


Figure 2: Korean acceptability judgements in Experiment 1 [Δ = difference in acceptability rating; the p-values correspond to results from pairwise linear mixed effects models discussed in the text; pair-wise comparison models had random intercepts of Subject, Stimulus and V_1, and a by-Subject random slope for V_2)].

To further probe the issue, we modelled the data using linear mixed-effects regression with

[7] While not relevant for the purposes of the paper, the interested reader might like to know that the English participants had no clear distinction in acceptabilities for the three types of stimuli for either place of articulation.

the acceptability ratings as the dependent variable. For all the mixed effects models presented in this article, we first identified the maximal random effects structure that converged for our data (Barr et al. 2013). Then we compared models with different fixed effects through a comparison of the Akaike Information Criterion (AIC), which measures model fit while controlling for over-parameterisation (Akaike 1974; Burnham, Anderson, and Huyvaert 2011). We use AIC instead of the more typical log-likelihood test since the latter suffers from the constraint that it can only be used to compare nested models. Note, the actual value of the AIC is unimportant for model comparison; instead, the difference in AIC value between models is more important. A lower AIC value suggests a better model, and Burnham and Anderson (2002) and Burnham, Anderson, and Huyvaert (2011) suggest that an AIC difference of more than 8 is an extremely large difference in the fits of any two models. Once the best model was identified, we looked at the fixed effects of that model.

For the current analysis, the fixed effects considered were V_2 (i, ɨ, ∅; baseline: ∅), C_1 (b, c; baseline: b), and an interaction between the two. The best model identified included all three fixed effects considered (Table 1); the maximal random effects structure is mentioned in the caption to the table (we follow this practice throughout to improve readability of the main text.). The model had an AIC differential of 87 from the next best model. The presence of the interaction suggests that there are differences in acceptabilities contingent on both the medial vowel and the preceding consonant.

Table 1: Linear mixed-effects model for the acceptability judgements in Experiment 1 [random effects structure: random intercepts of Subject, Stimulus and V_1, and by-Subject and by-V_1 random slopes for V_2]

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	2.78	0.15	22.98	18.68	< 0.0001
V_2-i	2.04	0.14	16.85	14.42	< 0.0001
V_2-ɨ	1.84	0.13	13.61	13.85	< 0.0001
C_1-c	1.34	0.06	28.20	23.91	< 0.0001
V_2-i : C_1-c	-1.54	0.08	28.20	-19.40	< 0.0001
V_2-ɨ : C_1-c	-1.37	0.08	28.20	-17.23	< 0.0001

In order to establish if the stimuli with no medial vowels were rated worse than the other two cases, we ran pairwise linear mixed effects models with a fixed effect of V_2 (i, ɨ, ∅; baseline: ∅). The random effects structure was the same as the above model except that it did not include a by-V_1 random slope of V_2. In the interest of concision, the results of these models were directly incorporated into Figure 2. As can be seen in the analysis presented there, the stimuli with no medial vowels were rated worse than either of the sets of stimuli with vowels. This suggests, parallel to the visual inspection, that participants rated stimuli with medial consonant sequences (*i.e.*, those with no medial vowels) as worse than those without.

We fitted another linear mixed effects model to the acceptabilities of V_1bC_2a and V_1cC_2a to see if one set was rated worse than the other (Table 2). The model revealed that indeed the V_1cC_2a stimuli were rated better than the V_1bC_2a stimuli.

Table 2: Linear mixed-effects model looking at the acceptability judgements of V_1bC_2a vs. V_1cC_2a in Experiment 1 [random effects structure: random intercepts of Subject, Stimulus, V_1 and C_2; by-Subject and by-V_1 random slopes of C_1)]

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	2.78	0.20	10.99	14.12	< 0.0001
C_1-c	1.34	0.17	6.97	7.87	< 0.0001

The results so far suggest that there is a gradience in the acceptabilities: V_1bC_2a « V_1cC_2a « $V_1C_1V_2C_2a$ (lowest-to-highest). Furthermore, for those who view the gradience in acceptability as stemming from a gradient phonotactic grammar, the results would typically be taken as evidence in favour of a gradient grammar that assigns a higher grammaticality to C_1VC_2 sequences than cC sequences, which in turn are assigned a higher grammaticality than bC sequences.

However, there are independent reasons to suspect that the above grammatical inference is not correct. The pattern of gradience is surprising based on Korean phonotactics. Note, both consonants (b and c) are unobserved in coda positions, and both consonant sequences (b+nasal and c+nasal) are unobserved. Furthermore, if one models phonotactic grammaticality at the level of featural co-occurences, one would expect V_1bC_2a to be *better* than V_1cC_2a, if anything. This is so given that a sequence of two voiced consonants is possible in Korean (*e.g.*, two nasals in a sequence). Consequently, there are more unobserved featural sequences in V_1cC_2a than V_1bC_2a. In fact, the sequential featural co-occurence violations in V_1cC_2a are a *superset* of those in V_1bC_2a. Therefore, the observed gradience is, contrary to the observed phonotactics of the language.

In line with the above reasoning, we argue that the observed gradience in this instance is not quite a product of the underlying grammar, but in fact is the product of gradience stemming from the perceptual system. To show this, we first establish that the stimuli do produce auditory illusions and then discuss how the acceptability ratings are correlated with the illusions.

2.2.2 Auditory Illusions

Next we turn to an analysis of the identification task. One of the aims of the identification task was to replicate prior research that showed that Korean listeners hear auditory illusions, particularly illusory vowels, when presented with illicit phonotactic sequences (Berent et al. 2007; Durvasula and Kahng 2015, 2016; Kabak and Idsardi 2007; Yun 2016). Therefore, as a first step, we focus on the stimuli with no medial vowels to see what the rates of illusory vowel perception are. These identification rates are presented visually in Figure 3. Note, as has been argued in prior work, it is necessary to compare the performance of a set of participants against speakers of a control

language in order to establish true phonological illusions, *i.e.*, illusions that are not due to stimulus or experimental confounds (Dupoux et al. 1999, 2011; Durvasula and Kahng 2015, 2016; Kabak and Idsardi 2007, amongst others). Therefore, we compare the medial vowel identification rates for the no vowel stimuli ($V_1C_1C_2a$) by the Korean listeners against those of the English listeners we tested. These are presented in Figure 3. Visual inspection reveals that the Korean listeners hear more illusory [ɨ] than English listeners for both V_1bC_2a and V_1cC_2a stimuli (note: orthographic <u> for both in the experiment), but they hear more illusory /i/ than the English listeners only for the V_1cC_2a stimuli (note: orthographic <i> for both in the experiment).

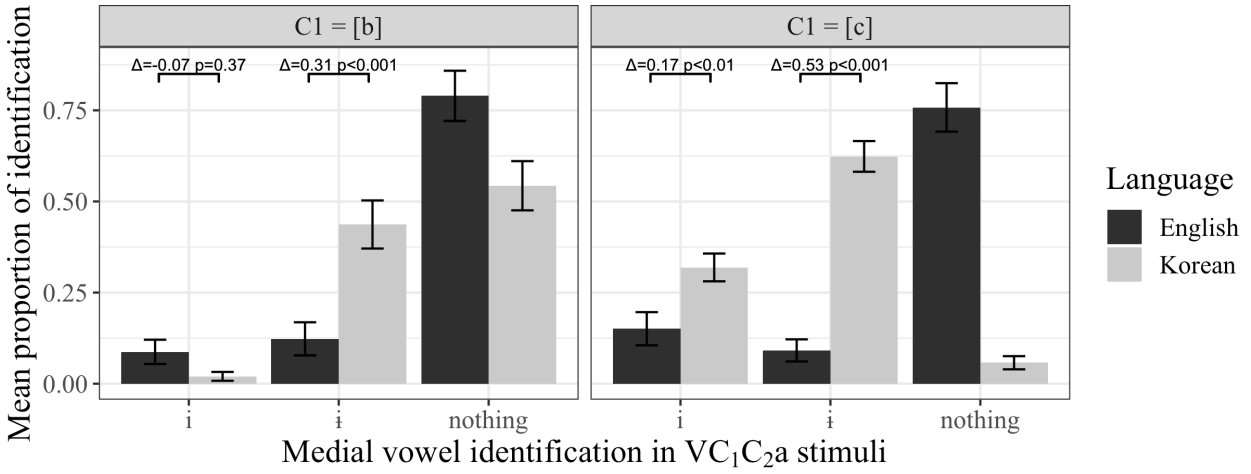


Figure 3: Comparing the identification of medial vowels in $V_1C_1C_2a$ stimuli by Korean and English listeners in Experiment 1 [Δ = difference in mean proportion of identification; the p-values correspond to results from pairwise logistic mixed effects models discussed in the text; comparison models had random intercepts of Subject, Stimulus and V_1]

For statistical analysis, we calculated the counts of i/ɨ/∅ responses for each subject for each stimulus and modelled the identification rates using mixed effects logistic regression. Furthermore, we subset the data to exclude the "no vowel responses"; the no vowel responses are perfectly predictable from the other two values, and therefore a statistical comparison including them does not add any information. The fixed effects that we included in the maximal model were Language (Korean vs. English; baseline = English), C_1 (b vs. c; baseline = b) and Chosen Vowel (i vs. ɨ; baseline = i). The best model was in fact the maximal fixed effects structure that included a 3-way interaction (Table 3). This model had an AIC differential of 68 from the next best model.

Given the three-way interaction, we followed up with pairwise comparisons for each chosen vowel for each C_1. Again, for the sake of concision, these results are directly incorporated into Figure 3. The statistical modelling replicated the visual observations and the patterns observed by Durvasula and Kahng (2015, 2016).

Table 3: Linear mixed-effects model looking at the responses to $V_1C_1C_2$a stimuli in Experiment 1 [random effects: random intercepts of Subject, Stimulus and V_1]

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.04	0.35	-8.62	< 0.0001
Language-Korean	-1.03	0.55	-1.89	0.06
ChosenVowel-ɨ	0.46	0.32	1.44	0.15
C_1-c	0.75	0.31	2.43	0.01
Language-Korean : ChosenVowel-ɨ	3.33	0.51	6.54	< 0.0001
Language-Korean : C_1-c	2.49	0.51	4.93	< 0.0001
ChosenVowel-ɨ : C_1-c	-1.15	0.44	-2.61	0.01
Language-Korean : ChosenVowel-ɨ : C_1-c	-1.26	0.62	-2.05	0.04

2.2.3 Gradience in acceptability stemming from gradience in perceptual inference

Now that we have confirmed prior illusory vowel findings in our own experiment, we are in a position to better understand the acceptability judgements of $V_1C_1C_2$a stimuli. In Figure 4, we present the acceptability ratings of $V_1C_1C_2$a stimuli partitioned by the perceived medial vowel. The visual inspection suggests that listeners rate both V_1bC_2a and V_1cC_2a stimuli as more acceptable when they hear an illusory vowel than when they do not hear an illusory vowel.

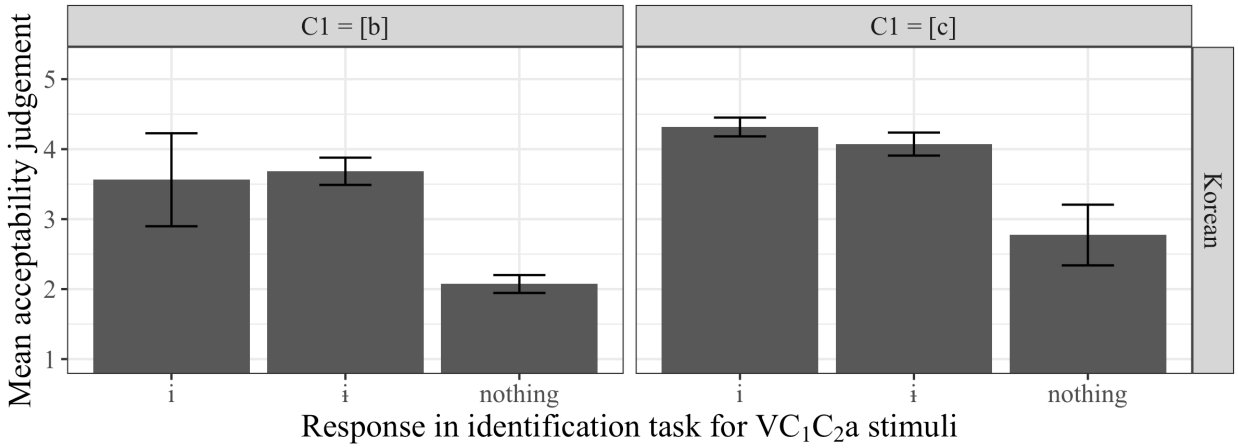


Figure 4: Acceptability judgements based on response in the identification task for the $V_1C_1C_2a$ stimulus

Statistical analysis of the acceptability rates suggests that the best model is one with two independent factors, Response (i, ɨ, ∅; baseline = ∅) and C_1 (b vs. c; baseline = b). This model is presented in Table 4. Crucially, the best model does not have an interaction term.

Note, the next best model had an interaction term in it and had an AIC 3.6 higher than the model presented. Based on the recommendations in Burnham and Anderson (2002) and Burnham, Anderson, and Huyvaert (2011), we therefore inspected this model too. However, the interaction terms were non-significant in this model too. This suggests that our interpretation in the previous paragraph that there is no evidence of an interaction is justified.

Table 4: Linear mixed-effects model looking at the acceptability judgements of V_1bC_2a and V_1cC_2a stimuli broken up by medial vowel identified [random effects structure: random intercepts of Subject, Stimulus and C_2)]

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	2.10	0.14	36.76	15.18	< 0.0001
Response-ɨ	1.48	0.10	687.45	14.99	< 0.0001
Response-i	1.73	0.13	514.97	13.16	< 0.0001
C_1-c	0.55	0.08	667.92	6.53	< 0.0001

We would like to clarify here that the fact that there is a main effect of consonant type [c vs.

b] in the above analysis recapitulates the observation that when the actual percept is not taken into account, there appears to be a clear difference in the acceptabilities of V_1bC_2a and V_1cC_2a, with the former rated worse. We further pointed out earlier that, for those who view the gradience in acceptability as stemming from a gradient phonotactic grammar, the results would typically be taken as evidence in favour of a gradient grammar that assigns a higher grammaticality to C_1VC_2 sequences than cC sequences, which in turn are assigned a higher grammaticality than bC sequences. Therefore, the crucial question is whether the Korean speakers thought that the stimuli without medial vowels adjacent to [b] were rated worse than those adjacent to [c], *because of not perceiving a vowel*. That is, the critical question is, is there an interaction in the results observed in Figure 4 and Table 4. And the answer seems to be that there is no evidence of an interaction effect, which in turn suggests that the data do not show evidence of a gradient grammar that assigns a higher grammaticality to cC sequences than bC sequences.

A final remaining issue is to understand why the overall acceptability rates of the V_1bC_2a are clearly lower than those of V_1cC_2a, when rates of illusions are not taken into account. The answer to this question is related to the fact that V_1cC_2a stimuli had far higher vowel illusion rates than V_1bC_2a stimuli. In fact, higher rates of vowel illusions have been observed in Korean for sounds with a noisy component (fricative, affricate, aspirated burst) in phonotactically illicit contexts (Daland, Oh, and Davidson 2019; de Jong and Park 2012; Durvasula and Kahng 2015; Kabak and Idsardi 2007). Note, however, the higher vowel illusion rates for V_1cC_2a stimuli cannot be purely due to the stimulus characteristics or for that matter the experimental paradigm itself, since the English listeners didn't show the same pattern, *i.e.*, it is a language-specific perceptual effect. An explanation for this language-specific perceptual effect, in our opinion, is likely to depend on the observation, discussed in the Introduction, that perception itself is known to be affected by phonological knowledge, wherein phonologically illicit sequences trigger higher rates of perceptual illusions (Daland, Oh, and Davidson 2019; Dupoux et al. 1999; Durvasula and Kahng 2015; Kabak and Idsardi 2007; Moreton 2002). Consequently, it is likely that the Korean and English listeners are behaving differently in the perceptual task due to how their phonological knowledge

interacts with the interpretation of the acoustic input during perception. One way to flesh out this possibility is to recognise that, as pointed out earlier, the sequential featural co-occurence violations in $C_{[palatal]}C$ sequences are a *superset* of those in $C_{[voiced\ stop]}C$ sequences. Furthermore, if the perceptual system is sensitive to sequential featural co-occurence violations (see Moreton (2002) for evidence), then it is straight-forward to see why the V_1cC_2a stimuli had higher rates of illusions, but only for Korean listeners.

A second way to flesh out the possibility of an interaction between phonological knowledge and the acoustic input during perception is by following up on the results of Durvasula and Kahng (2016): they noted that though both $C_{[voiced\ stop]}C_{[nasal]}$ and $C_{[voiceless\ stop]}C_{[nasal]}$ sequences are phonotactically illicit *within* words in Korean, the latter are perfectly acceptable across Intonational Phrases, while the former are illicit across all levels of the prosodic hierarchy. They observed that $C_{[voiced\ stop]}C_{[nasal]}$ sequences trigger more illusory vowels than the $C_{[voiceless\ stop]}C_{[nasal]}$ sequences. However, the cues for voicing in our stimuli are quite weak, particularly in coda positions (see Figure 1 for examples).[8] Therefore, it is possible that the listeners misperceived the voiced stop as a voiceless stop in some cases. If so, the corresponding inferred representation with a voiceless stop would have had a lower probability of triggering an illusory vowel. Similar to $C_{[voiced\ stop]}C_{[nasal]}$ sequences, sequences of $C_{[palatal]}C_{[nasal]}$ are also illicit across all levels of the prosodic hierarchy, but the misperception of voicing does not help in the case of input V_1cC_2a stimuli, therefore, they have higher rates of illusory vowels. In contrast to the Korean case, both $C_{[voiced\ stop]}C_{[nasal]}$ and $C_{[palatal]}C_{[nasal]}$ sequences are licit in American English, and even if the voicing is misperceived, $C_{[voiceless\ stop]}C_{[nasal]}$ sequences are also licit. Therefore, we would expect low rates of illusory vowels in both contexts. This potential perceptual parsing asymmetry, both within Korean and across Korean and American English, explains why V_1cC_2a stimuli trigger higher rates of illusions, but only for Korean listeners. At this point, we don't know which of the above explanations is at play (we suspect both), since our experiment wasn't designed (or intended) to tease them apart. Cru-

[8]This is not surprising given they were produced by a Korean-English bilingual, and both languages have been argued to have passive, and not true, voicing (Avery and Idsardi 2001; Beckman, Jessen, and Ringen 2013; Cho, Jun, and Ladefoged 2002; Iverson and Salmons 1995).

cially, while it might be possible to explain the language-specific perceptual results (*i.e.*, without circularity or simply restating the observation) in other ways, it is worth re-iterating that our interest here is not in the perceptual asymmetry *per se*, but in the effect this asymmetry has on the acceptability judgements.

Returning to the issue of acceptability judgements and the gradience observed in our experiment, one can describe a very simple mathematical model that derives the basic gradient fact. Imagine that the acceptability of an input stimulus was a simple weighted average of categorical word-likeness judgements over the probabilistic perceptual representations recovered from the input signal (note: this is true for all stimuli, both licit and illicit). Since listeners hear more illusory vowels in V_1cC_2a stimuli than in V_1bC_2a stimuli, they have a higher probability of perceiving a licit stimulus in the case of the former than the latter. Consequently, a listener is likely to give a higher acceptability score to the former than the latter. Therefore, the simple model derives the observed gradience in the acceptability ratings.

In order to make the same point in another way, we coded any medial vowel response (i or ɨ) to $V_1C_1C_2a$ stimuli as an illusion, and looked at the average illusion rates for different participants, and compared these illusion rates to the acceptability score given by the same participant (See Figure 5). What can be observed is that for both V_1bC_2a and V_1cC_2a stimuli, there is a positive correlation between the rate of illusion for a listener and their average acceptability rating for the same stimulus. Note, the x-axis range for V_1cC_2a is truncated because listeners were hearing illusions at very high rates for the stimuli.

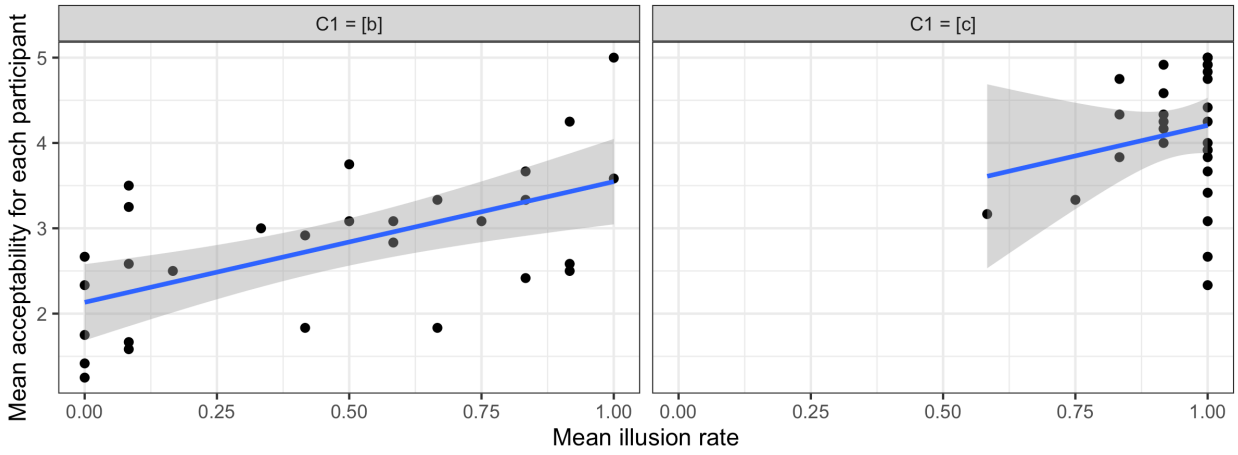


Figure 5: Correlation between acceptability judgements and illusion rate for $V_1C_1C_2a$ stimulus for the Korean listeners

We statistically modelled the mean acceptability values for each listener with linear regression. We considered three factors: Mean Illusion Rate, C_1 (c vs. b; baseline: b), and the interaction between the two factors. The model with the lowest AIC was one with the two simple factors, and no interaction term (Table 5). Given that the baseline for C_1 is V_1bC_2a, the table reveals that there is a clear positive relationship between the Mean Illusion Rate and mean acceptability for the V_1bC_2a stimuli. Furthermore, the model included a significant increase in the slope for V_1cC_2a stimuli. However, this latter effect should be taken with a grain of salt given the fact that two other models were within an AIC differential of 4: the first was a model with just a Mean Illusion Rate as a factor, which suggests that C_1 is unnecessary to model the data reasonably, and the second was a model with all three factors (including the interaction term) but in which C_1 is non-significant. Given that the AIC values are so close, for such different models, it is difficult to interpret the meaningfulness of the additional C_1 factor in the best model we presented, and therefore it is not clear that the slopes for V_1bC_2a and V_1cC_2a are meaningfully different. In contrast, the Mean Illusion Rate was clearly significant in all three models, suggesting that the factor is a useful one in modelling the data.

Table 5: Linear model comparing the acceptability judgements of $V_1C_1C_2a$ stimuli with rates of illusions for the same stimuli

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.13	0.21	10.06	> 0.00001
MeanIllusionRate	1.42	0.36	3.93	< 0.001
C_1-c	0.66	0.26	2.56	0.01

Finally, we'd like to return the reader's attention to the fact that the x-axis range for the V_1cC_2a stimuli is quite truncated (due to high rates of illusions). Therefore, it is likely that the estimate of the effect of the MeanIllusionRate for these particular stimuli is not as trustworthy. This is a second reason why the statistical significance observed for the C_1 factor should be taken with a grain of salt.

2.2.4 Discussion

Our results in Experiment 1 suggest that when we just look at the acceptability ratings of the stimuli, there appears to be a gradient, V_1bC_2a « V_1cC_2a « $V_1C_1V_2C_2a$. This result *could* be taken as evidence in favour of a gradient grammar that assigns a higher grammaticality to C_1VC_2 sequences than cC sequences, which in turn are assigned a higher grammaticality than bC sequences. However, further probing the perceived input of these stimuli reveals that the observed gradience in the acceptability judgements is an artifact of the differential rates of illusions (for licit vs. illicit sequences). Finally, we show that one can also see gradience in judgements as a function of the rate of illusion, which further corroborates that perception is a clear source of gradience in acceptability judgements.

One question that is reasonable to ask at this point is whether the crucial pattern of acceptability judgements observed in Experiment 1 was a task effect. Perhaps, the reason the acceptability judgement patterns appear the way they are is due to the fact that the participants were explicitly asked

for the medial vowel, which thereby might have focussed their attention on the presence/absence of that segment in making an acceptability judgement. We address this concern in Experiment 2.

3 Experiment 2

In Experiment 1, it is possible that the overall acceptability rating patterns and the dependence of the acceptability ratings on the perceptual repair were due to an experimental confound; we explicitly drew participants' attention towards the repairs through the identification task; consequently, the participants could have understood their objective during the acceptability task as rating the repaired stimulus. In order to address the confound, we simply removed the identification task for Experiment 2. Therefore, the participants in Experiment 2 only gave acceptability judgements for the same stimuli as in Experiment 1. In this experiment, our main interest was in seeing if the main patterns of gradience observed in Experiment 1 remain despite the absence of an identification task. While the task cannot probe the relationship between the acceptability task and the perceptual repairs, it does allow us to see if the overall acceptability patterns remain the same in the absence of an identification task. If we do see the same overall pattern of acceptability, and further see the same gradience in acceptability (namely, V_1bC_2a « V_1cC_2a « $V_1C_1V_2C_2a$), then we can be sure that the crucial results in Experiment 1 were not substantially affected by the presence of the identification task. We show evidence in favour of exactly this inference in what follows.

3.1 Methods

3.1.1 Participants

Twenty native Korean speakers (10 women and 10 men, age: 19-25 years) participated in the experiment, none of whom participated in Experiment 1. All the Korean speakers were recruited in Daegu, South Korea. All the participants were undergraduate or graduate students at a large university in Daegu and were compensated for their participation.

3.1.2 Stimuli

The stimuli for Experiment 2 were the same 96 nonceword tokens used in Experiment 1.

3.1.3 Procedure

In Experiment 2, the participants only completed an acceptability judgement task for each stimulus. The participants heard a stimulus and were asked to make an acceptability judgement on the stimulus they heard on a scale of 1-5 (1: Impossible `불가능함', 5: Possible `가능함'). The instructions were in Korean for the Korean participants (`들은 단어가 한국어로 가능한 단어인지 고르세요.' Choose if the word is a possible word in Korean.). Before the actual experiment, each participant completed a practice session, to ensure familiarity with the task. The practice session had 12 trials with a different set of nonce words. The intertrial interval was 1000 ms. All the trials were randomized for each participant. The experiment was conducted individually in a quiet room with a low-noise headset.

3.2 Results

For Experiment 2, since the crucial question is to observe the acceptability judgements in the absence of an identification task, we repeat the analytical strategy discussed for Experiment 1.

A visual inspection of the overall patterns of acceptability reveals that they are similar to those in Experiment 1 (Figure 6). The two primary observations are that the stimuli without medial vowels ($V_1C_1C_2a$) are rated lower than those with medial vowels ($V_1C_1V_2C_2a$), and that there seems to be a tendency for the V_1cC_2a stimuli to be rated better than the V_1bC_2a stimuli.

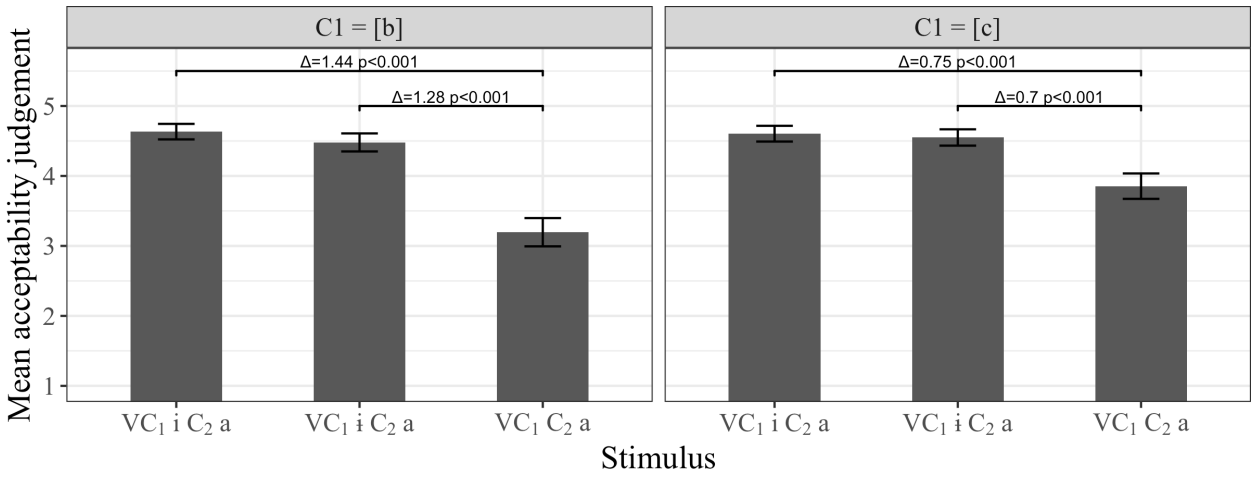


Figure 6: Korean acceptability judgements in Experiment 2 [Δ = difference in acceptability rating; the p-values correspond to results from pairwise linear mixed effects models discussed in the text; pair-wise comparison models had random intercepts of Subject, Stimulus and V_1, and a by-Subject random slope for V_2)]

Statistical analysis showed that the best model was the one with an interaction term (Table 6). The next best model had an AIC value that was higher by 27.[9]

Table 6: Linear mixed-effects model for the acceptability judgements in Experiment 2 [random effects structure: random intercepts of Subject, Stimulus and V_1, and a by-Subject random slope for V_2]

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	3.20	0.18	21.39	17.81	< 0.0001
V_2-i	1.44	0.14	26.12	10.24	< 0.0001
V_2-i̶	1.28	0.13	27.02	9.72	< 0.0001
C_1-c	0.66	0.07	28.00	8.88	< 0.0001
V_2-i : C_1-c	-0.69	0.10	28.00	-6.55	< 0.0001
V_2-i̶ : C_1-c	-0.59	0.10	28.00	-5.60	< 0.0001

[9]The same comparison in Experiment 1 also had a by-V_1 random slope for V_2. However, for Experiment 2, some of the models did not converge with that random effect added. Crucially, both the full fixed model (with three factors including an interaction term) and the one with two factors but no interaction term converged, which therefore allowed us to compare whether there was an interaction with the same random effects structure as in Experiment 1. As with the modelling in the main text, the model with the interaction term was better (by an AIC difference of 31), even when the same random effects structure as in Experiment 1 was used.

As with Experiment 1, given the interaction, we followed up with pairwise comparisons between the stimuli with and without vowel separately for the bilabial and palatal cases. The main effects are reported directly in Figure 6. The results confirm the visual observation that the $V_1C_1C_2a$ stimuli are indeed rated worse than either of the $V_1C_1V_2C_2a$ stimuli.

Finally, we also compared the acceptability of V_1bC_2a stimuli with V_1cC_2a stimuli directly (Table 7). As with Experiment 1, we found that the V_1cC_2a stimuli were rated better than the V_1bC_2a stimuli.[10]

Table 7: Linear mixed-effects model looking at the acceptability judgements of V_1bC_2a vs. V_1cC_2a in Experiment 2 [random effects structure: random intercepts of Subject and Segment; a by-Subject random slope of C_1)]

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	3.20	0.20	19.00	15.81	< 0.0001
C_1-c	0.66	0.16	19.00	4.02	0.001

The results replicate all the primary findings of Experiment 1 with respect to gradience in acceptability ratings, despite the absence of an identification task in Experiment 2. Therefore, we can infer that the primary findings in Experiment 1 were not due to the identification task.

The reader might ask why we did not include an overall model directly comparing the acceptabilities of Experiments 1 and 2 to see whether or not there was interaction effect with experiment (Experiment 1 vs. Experiment 2). The main issue with such an overall model is that it does not add to understanding of the question that Experiment 2 was designed to answer. It is true that if such a model had no significant main or interaction effect with experiment, that would be consistent with the claim that the the identification task did not affect the acceptabilities in Experiment 1. However, even if there was a main or interaction interaction effect with experiment, the primary result, namely, V_1bC_2a « V_1cC_2a « $V_1C_1V_2C_2a$, is still very clearly observable in Experiment 2. Consequently, no matter what the results of such an overall model, our interpretation that the per-

[10]The same comparison in Experiment 1 also included a by-segment random slope of C_1. However, that model did not converge for Experiment 2, and so the estimation is not trustworthy. However, it is worth noting that the fixed effect estimate and p-value for that model were nearly identical to the one presented in Table 7.

ceptual identification task in Experiment 1 is not the reason for the crucial pattern of acceptabilities would have been unaffected. For this reason, we do not include an overall model comparing the acceptabilities of the Korean participants in Experiments 1 and 2.

To re-iterate, our focus is not on showing that the acceptability results in Experiments 1 and 2 are identical. In fact, we doubt that is likely to be the case given that the tasks are different.[11] Instead, our focus is on showing that the main result related to the acceptabilities, namely, V_1bC_2a « V_1cC_2a « $V_1C_1V_2C_2a$, is not due to a task effect in Experiment 1. Our modelling in Experiment 2 show evidence for this.

4 Conclusion

In this article, we present two experiments and argue that there's a relationship between gradient phonotactic acceptability judgements and perception. In Experiment 1 (Section 2), we show that: (a) the acceptability of nonce words with such sequences is *contingent* on the presence/absence of an illusory vowel; (b) the acceptability of the nonce words depends on the rate at which illusory vowels are perceived in perceptual experiments across participants.[12] In Experiment 2 (Section 3), we show that the crucial pattern of acceptability judgements observed in Experiment 1 was not due to a task effect related to the presence of an identification task as part of the experiment.

Overall, our results suggest that wordlikeness judgements need a concomitant probing of the perceived stimulus in order to be interpreted as evidence for a particular grammatical claim. Furthermore, our results indicate that directly modelling acceptability judgements with a (gradient) grammar will necessarily, and incorrectly, incorporate the gradience from non-grammatical sources into the grammar. In fact, just observing gradience in acceptability judgements tells us nothing about the underlying grammar. For example, any categorical system that has a sufficiently noisy implementation will look gradient on the surface. Therefore, the shape of the response patterns by

[11] In fact, we fitted an overall model based on a reviewer's comment, and there is an interaction effect of experiment and C_1 on the acceptability judgements.

[12] Note, if this overall rate of confusability is ultimately derived from the rate of confusability within participants, then a stronger statement can be made, namely, that the acceptability of the nonce words depends on the *degree* of illusory vowels within the speaker.

themselves should therefore not be taken as any indication of the underlying state of the grammar; instead, studies like the current one are necessary to compare patterns of judgement with patterns of non-grammatical processes in order to understand how to factor out the different sources present in an acceptability judgement. More generally, our results go hand-in-hand with those of Schütze (2020), who argues that syntactic acceptability judgements cannot be taken at face-value and that there needs to be a better understanding of what the participant interprets the sentence to be before one can interpret the acceptability judgements. Both Schütze's (2020) and our results suggest that much more nuance is needed in interpreting acceptability judgements generally than is currently standard with results from formal experiments.

Having said the above, we would like to acknowledge here that our experiments involved very constrained stimuli and in fact a rather small set of phonotactic possibilities in any language. Furthermore, while we believe the use of such a constrained stimulus set was essential as a first step in order to carefully understand the underlying sources of the gradient acceptabilities based on established prior perceptual results, we do think there needs to be more such work done on much larger sets of stimuli in order to establish the main point beyond reproach. Therefore, the results presented herein should be taken as a proof of concept, and we encourage other researchers interested in modelling/testing acceptability judgements to include perceptual experiments as part of their experimental procedure.

One approach that might initially be considered reasonable to avoid the confound of probabilistic perceptual representations is the use of orthographic presentations. Orthographic presentations are sometimes seen as directly informing the participant of the representation to be judged (Daland et al. 2011; Pizzo 2015), thereby by-passing the issue of probabilistic perception raised above. However, it is possible that participants depend on orthotactic knowledge (Apel 2011)—*i.e.*, knowledge of spelling rules—rather than purely phonotactic knowledge to make such judgements. In fact, Daland et al. (2011) acknowledge the possibility of orthotactic knowledge influencing the judgement of nonce words, but argue that it is likely to only explain a "coarse difference in rating between attested, marginal and unattested onsets", and not any gradient judgements of unobserved clusters.

However, such a view presupposes that orthotactic knowledge is purely based on segment-like letters, and that speakers have no awareness that letters form clusters with similar patterns akin to natural classes in phonology. Using orthographic presentation further assumes that there isn't much gradience in inferring a phonemic representation from orthographic representations, contrary to fact (Gorman et al. 2020). We think orthographic presentation requires much more careful scrutiny before it can be used as a surrogate for auditory presentation, and one needs to incorporate more explicit theories of orthotactic knowledge into modelling judgements based on orthographic presentation. We would like to note that we believe this is a rich line of enquiry for those interested in using orthographic presentations to understand the nature of phonotactic knowledge.

Our results also raise the following questions: (a) What exactly do gradient phonotactic models of wordlikeness judgements capture? Note, perception itself uses phonological knowledge; therefore, perceived representations have already been in some sense "filtered" by the phonotactic grammar. Consequently, any judgements based on such repaired perceptual representations and any gradience observed in such judgements would not be an accurate reflection of the listeners phonotactic knowledge. This we believe is a genuine issue for researchers interested in using statistical modelling of such judgements to claim an underlying gradient phonotactic grammar, and raise it here to encourage future scrutiny. (b) To what extent are phonotactic perceptual repairs versus wordlikeness judgements perhaps comparable to ultimate parses versus (sometimes temporary) illusions of (un)grammaticality that have been discovered for sentence-level comprehension and judgements (Phillips, Wagers, and Lau 2011)? A careful study and comparison of related phenomena in syntax is likely to further our understanding not only of phonological knowledge, but the nature of linguistic knowledge more generally.

Relatedly, it is important to note that our argument extends beyond the context of illusory vowels. As we pointed out earlier, practically all acoustic input results in probabilistic percepts (Feldman and Griffiths 2007; Kleinschmidt and Jaeger 2015; Norris and McQueen 2008; Norris, McQueen, and Cutler 2003; Sonderegger and Yu 2010, amongst others). If wordlikeness judgements are based on probabilistic perceived representations, as discussed in the paper, and reflect some sort

of weighted average rating over such probabilistic percepts, it is easy to envision how a substantial amount of gradience observed in such judgements actually stems from perceptual gradience even in cases beyond illusory vowels. Crucially, variable or probabilistic percepts are true even in the case of segments (or segment sequences) observed in the language, as is known since the classic work of Miller and Nicely (1955). Their study shows that there is a lot of variability in responses even in the case of licit monosyllabic Cɑ nonce words (even in the high signal-noise ratio condition), with some CV sequences showing far more variable identification than others. For example, they observed that [θ] is often asymmetrically confused with [f] at extremely high rates (a phenomenon that is well-known today). So, a real word like <thick> [θɪk] might sometimes be perceived as the nonceword [fɪk]. Note, the Ganong effect (Ganong 1980; Rysling et al. 2015) may bias the listener towards the real word <thick>, but it doesn't completely exclude the possibility of interpreting the input as the nonceword [fɪk]. Consequently, the probabilistic perceived representation will include some nonceword possibilities, which will thereby decrease the overall acceptability judgement of the word.[13] In contrast, a segment such as [t^h] is far less confusable[14] with other segments; therefore, a word such as <tick> [t^hɪk] is likely to have a lower probability of illicit nonceword percepts — this in turn would suggest a higher rate of acceptability than the first case. We now have a situation where the words and the segmental sequences are both observed and perfectly grammatical, yet we have an expectation of gradience in acceptability judgements. In a similar vein, one could show that less frequent phonotatic sequences trigger higher rates of confusion, and therefore are more likely to be judged as less acceptable, than more frequent phonotactic sequences.

Finally, our results are consonant with the claims of Gorman (2013), who found that models of lexical similarity are often better models of human wordlikeness judgements than computational models implementing a gradient phonotactic grammar. In that study, gradient phonotactic grammars were often also outperformed by 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

[13] As a reminder, this follows if the judgement is viewed as a sort of weighted average rating over the probabilistic percepts.

[14] We intend this as a relative statement, and explicitly caution the reader against interpreting it as us claiming the perception of [t^h] is free of any confusion, which is nearly impossible according to us.

rimes.[15] Gorman's (2013) results along with ours suggest that it may be possible to maintain a *categorical* phonotactic grammar and account for the observed gradient wordlikeness judgements with various performance factors, consistent with traditional categorical grammatical architectures (Chomsky 1965; Schütze 1996, 2011; Sprouse et al. 2018; Valian 1982).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. doi: 10.1109/TAC.1974.1100705

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.

Apel, K. (2011). What is orthographic knowledge? *Language, Speech, and Hearing Services in Schools*, *42*(4), 592–603. doi: 10.1044/0161-1461(2011/10-0085)

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

Avery, P., & Idsardi, W. J. (2001). Laryngeal dimensions, completion and enhancement. *Distinctive feature theory*, 41–70.

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

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using `lme4`. *Journal of Statistical Software*, *67*(1), 1–48.

[15] Note, even with syntactic acceptability judgements, recent work has argued that despite the welter of gradience observed in any acceptability task, the extant results are actually compatible with categorical syntactic models (De Santo 2020a,b; Sprouse et al. 2018).

Beckman, J., Jessen, M., & Ringen, C. (2013). Empirical evidence for laryngeal features: Aspirating vs. true voice languages. *Journal of Linguistics*, *49*(2), 259–284.

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.

Boersma, P., & Weenink, D. (2016). *Praat: doing Phonetics by Computer [Computer program]*.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference - a practical information-theoretic approach*. Springer-Verlag.

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*(1), 23–35. doi: 10.1007/s00265-010-1029-6

Caplan, S. P. (2021). *The immediacy of linguistic computation* (PhD dissertation). University of Pennsylvania.

Cho, T., Jun, S.-A., & Ladefoged, P. (2002). Acoustic and aerodynamic correlates of korean stops and fricatives. *Journal of phonetics*, *30*(2), 193–228.

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.

Coetzee, A. W., & Pater, J. (2008). Weighted constraints and gradient restrictions on place co-occurrence 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.

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.

De Santo, A. (2020a). Mg parsing as a model of gradient acceptability in syntactic islands. *Proceedings of the Society for Computation in Linguistics 2020*, 59–69.

De Santo, A. (2020b). *Structure and memory: A computational model of storage, gradience, and priming* (PhD dissertation). State University of New York at Stony Brook.

de Jong, K., & Park, H. (2012). Vowel epenthesis and segment identity in korean learners of english. *Studies in Second Language Acquisition*, *34*(1), 127–155.

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.

Dupoux, E., Parlato, E., Frota, S., Hirose, Y., & Peperkamp, S. (2011). Where do illusory vowels come from? *Journal of Memory and Language*, *64*(3), 199–210.

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.

Feldman, N. H., & Griffiths, T. L. (2007). A Rational Account of the Perceptual Magnet Effect. *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, 257–262.

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

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110.

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

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.

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.

Hallé, P. A., Segui, J., Frauenfelder, U., & Meunier, C. (1998). Processing of illegal consonant clusters: A case of perceptual assimilation? *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 592–608.

Hay, J., Pierrehumbert, J., & 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. (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.

Hlavac, M. (2018). *Stargazer: Well-formatted regression and summary statistics tables* [R package version 5.2.2]. Central European Labour Studies Institute (CELSI). Bratislava, Slovakia.

Iverson, G. K., & Salmons, J. C. (1995). Aspiration and laryngeal representation in Germanic. *Phonology*, *12*, 369–396.

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.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. doi: 10.18637/jss.v082.i13

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, *54*(5), 358.

Massaro, D. W., & Cohen, M. M. (1983). Categorical or continuous speech perception: A new test. *Speech communication*, *2*(1), 15–35.

McCarthy, J. J., & Prince, A. (1986). Prosodic morphology, Report no. RuCCS-TR-32. Rutgers University Center for Cognitive Science, New Brunswick, NJ. [Retrieved from http://ruccs.rutgers.edu/tech_rpt/pm86all.pdf].

McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., & Subik, D. (2008). Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(6), 1609.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, *27*(2), 338–352. doi: 10.1121/1.1907526

Moreton, E. (2002). Structural constraints in the perception of english stop-sonorant clusters. *Cognition*, *84*, 55–71.

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

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.

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.

Phillips, C., Wagers, M. W., & Lau, E. F. (2011). 5: Grammatical illusions and selective fallibility in real-time language comprehension. *Syntax and semantics* (pp. 147–180). Brill. doi: https://doi.org/10.1163/9781780523750_006

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.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rysling, A., Kingston, J., Staub, A., Cohen, A., & Starns, J. (2015). Early Ganong effects. *Proceedings of the 19th International Congress of Phonetic Sciences*.

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.

Schütze, C. T. (2020). Acceptability ratings cannot be taken at face value. In S. Schindler, A. Drozdzowicz, & K. Brøcker (Eds.), *Linguistic intuitions*. Oxford University Press.

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

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

Sonderegger, M., & Yu, A. (2010). A rational account of perceptual compensation for coarticulation. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society (CogSci10)*, 375–380.

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.

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.

Wickham, H. (2017). *Tidyverse: Easily install and load the 'tidyverse'* [R package version 1.2.1].

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].