An experimental study of Catalan consonant alternations[1]

Abstract

We describe two "wug-test" experiments covering consonant alternations in Catalan. We chose our alternations and stimulus items with the goal of testing hypotheses in phonological theory. Our findings provide support for the following claims: (1) Speakers are capable of frequency-matching exceptionful phonological patterns in the lexicon. To a limited degree, speakers productively extend (2) opaque and (3) saltatory phonology. In post-hoc examination of the data, we sought explanations for why speakers deviate from frequency matching and located several possible mechanisms. First, speakers seem to tacitly reference orthography, matching more closely the alternations that are spelt. Second, speakers often respond to lexical patterns that are poorly attested by "avoidance": using unusual morphology that lets them avoid making a phonological choice. Lastly, participants differ substantially both in their degree of avoidance and in their detailed phonological preferences.

Keywords: Catalan phonology, wug test, frequency-matching, opacity, saltation

Word count: 15585

[1] Acknowledgments suppressed.

1. Introduction

Catalan is a Romance language spoken in Catalonia and neighbouring regions. Its sound pattern has been the subject of meticulous generative analysis, notably Mascaró (1975) and Wheeler (2005). We describe here an experimental study—a "wug test," in the sense of Berko (1958)—of consonant alternations found at the right edges of Catalan stems. These alternations include deletion of singleton /n/ and /r/,[2] final cluster simplification, and a manner-shifting version of Final Devoicing (/ʒ/ → [tʃ]). These are illustrated in (1) with masculine and feminine forms of the same stem.

(1) *Alternations chosen for study*

a. */n/-deletion*

/san/ → [ˈsa] 'healthy.M.SG'; cf. F.SG [ˈsan-ə]

b. */r/-deletion*

/klar/ → [ˈkla] 'clear.M.SG'; cf. F.SG [ˈklar-ə]

c. */nt/-cluster simplification*

/sant/ → [ˈsan] 'holy.M.SG'; cf. F.SG [ˈsant-ə]

d. */ʒ/ → [t͡ʃ] alternation*

/bɔʒ/ → [ˈbɔt͡ʃ] 'crazy.M.SG'; cf. F.SG [ˈbɔʒə]

Catalan is a well-studied system that has played an important role in the evolution of phonological theory; it is also widely employed in textbooks (Kenstowicz & Kisseberth 1979; Odden 2005) and computational study (Cotterell et al. 2015; Shilen & Wilson 2022; Rasin et al. 2021; Wang & Hayes 2025). Thus, it seems sensible to evaluate the productivity of its phonological patterns.

The particular phonological phenomena we chose for study stand out for the light they might shed on four open questions in phonological theory (2):

(2) *Research questions*

a. **Frequency-matching** (Zuraw 2000; Ernestus & Baayen 2003): When experimental participants are wug-tested with probes involving exceptional phonology, do their

[2] For convenience we transcribe the singleton rhotic [ɾ] as [r]. For description of the segmental phonemes of Catalan, see Wheeler (2005: ch. 2).

answers *frequency-match* the statistical patterns of the lexicon? When speakers *deviate* from frequency-matching (e.g. Becker et al. 2012; Hayes & White 2013), what is the cause?

b. **Productivity of opacity** (Kiparsky 1973; Sanders 2003): Can speakers productively extend an opaque phonological pattern?

c. **Stability of saltatory alternations** (Hayes & White 2015): Do speakers exposed to saltatory phonological alterations (A alternates with B, passing over stable intermediate C) tend to "repair" saltation by replacing them with non-saltatory patterns?

d. **UR inference** (Ernestus & Baayen 2003): Can speakers use the statistical patterns of their lexicon to guess the underlying representation of a neutralised surface form?

Post-hoc examination of our data reveals other areas of theoretical interest. We argue that our participants often adopt *"avoidant"* strategies, using aberrant morphological choices to avoid having to make a decision about the phonology. We find this is particularly the case with low-frequency patterns and suggest that this is a consequence of uncertainty about the appropriate outcome. We also note the existence of differences among individual participants, both in their tendency to be avoidant, and in their detailed phonological preferences.

1.1 Outline

In §2, we describe in more detail the consonant alternations of Catalan on which we focus. Our description relies on a lexical corpus of Catalan paradigms which permits us to assess the numerical representation of rival patterns. §3 covers the design of our experiments: stimuli, experimental procedure, and choice of participants. §4 gives the results, assessing how they bear on the research questions of (2), and offering our speculations concerning the less-expected outcomes. We also employed a MaxEnt analysis to uncover differences among individual participants. §5 offers a summary and directions for further research.

2. The phonology of stem-final consonants in Catalan

We rely heavily on the descriptions and analyses given in Mascaró (1975) and Wheeler (2005). Wheeler covers several Catalan dialects, but here we will focus entirely on Central Catalan, the variety spoken in Barcelona and surrounding regions.

In this section we employ, for terseness, a rule-based analysis (Chomsky & Halle 1968); a constraint-based analysis follows in §4.3.3.

2.1 Database

For our purposes, we need quantitative assessment of how often the various patterns are represented in the lexicon. To this end we gathered a collection of paradigms from Wiktionary (www.wiktionary.org; 185K words). The paradigm for ‘small’ is given in (3).

(3) *Sample paradigm for regular Catalan nouns and adjectives*

a.	Masculine singular	Ø	[pəˈtit]
b.	Masculine plural	[-s]	[pəˈtit-s]
c.	Feminine singular	[-ə]	[pəˈtit-ə]
d.	Feminine plural	[-ə-s]	[pəˈtit-ə-s]

This is the regular pattern; there are also irregular paradigms, which turn out to play an important role here. In particular, in some cases, the masculine takes an [-u] ending (e.g., [muˈrɛn-u] ‘brunet-M.SG’), or a [-ə] ending (e.g., [pəriuˈdist-ə] ‘journalist-M.SG’).

We used *Wikiextract* (Ylonen 2022) to obtain all complete paradigms from Wiktionary, then removed all words not known to the native-speaker second author (n = 619), since such forms are less likely to be known to our participants. With these deductions our list came to 5761 paradigms. Several of the alternations we study are not expressed in Catalan orthography, so we needed IPA transcriptions. These were taken from Wiktionary where available; else, they were generated with the *catalan2ipa* package (Groß 2019). All transcriptions were checked by the native-speaker author.

A few forms show optionality in the relevant phonology; for example [upsˈku.rə] ‘dark.F.SG’ has as the corresponding masculine either [upsˈku] or [upsˈkur] ‘dark.M.SG’. In calculating application rates for such cases, we counted each variant as 0.5.

2.2 /n/-deletion

The first process we examine, /n/-deletion, deletes /n/ in postvocalic word-final position.[3] In the nominal/adjectival paradigms studied here, this results in alternations, as deletion takes place in the unsuffixed masculine forms but not in feminines. An example is given in (4b).

(4) a. */n/-deletion*

n → Ø / V ___]$_{\text{word}}$ ‘Delete postvocalic [n] word-finally’

[3] The postvocalic restriction is illustrated by [əˈtɛrn] ‘eternal-m.’

b. *Derivations*

*‘flat.*M.SG*’*	*‘flat-*F.SG*’*	
/plan/	/plan-ə/[4]	underlying representations
∅	—	/n/-deletion
[ˈpla]	[ˈplanə]	surface representations

As Mascaró (1975) and Wheeler (2005) point out, /n/-deletion involves a fair number of exceptions. For instance, we might expect the UR /nɛn/ ‘child.M.SG,’ (F.SG [ˈnɛn-ə]) to surface in the masculine as *[ˈnɛ], but in fact it is pronounced [ˈnɛn]. Pairs like /plan/ vs. /nɛn/ demonstrate that /n/-deletion is not entirely predictable. However, as with exceptionality elsewhere (Zuraw 2000), exceptionality for /n/-deletion is *patterned*, in the sense that one can locate particular environments in which application is especially frequent or especially unusual among the words of the existing lexicon. Because of this patterning, the predictability of /n/-deletion is better than random. Importantly for this study, experimental work shows that, at least in other languages, participants are able to use such patterning to guide their responses in a wug-test—in other words, they show a tendency to *frequency-match* the lexicon (Hayes et al. 2009:826; Becker & Gouskova 2016; O'Hara 2020; Song & White 2022; Gouskova 2025).

There is also a substantial body of work that attempts to explain the cases where speakers deviate from precise frequency matching. For example, Becker et al. (2012) give that participants deviate from frequency-matching in order to respect a UG bias disfavouring alternations in initial syllables; for other cases see e.g., Zhang et al. (2011), Moreton and Pater (2012), and Hayes and White (2013). While establishing principles of learning bias is of great theoretical interest, our goal is more modest, namely, just to establish that there is a tendency toward frequency-matching in Catalan.

As a basis for studying frequency-matching, we consider four environments that affect /n/-deletion, shown in (5). Most of these were noticed earlier by Mascaró or Wheeler.

(5) *Environments for /n/-deletion*

a. **Particular suffixes**. In words formed with /-in/, a frequent adjectival suffix, /n/-deletion applies 100% of the time (105/105 cases in our corpus). An example is [məʎurˈk-in-ə] ~ [məʎurˈk-i] ‘Mallorc-an-F.SG/M.SG’.

b. **Penultimately-stressed stems.** Various stems of Catalan have stress on their second to last syllable, hence penultimate word stress in the masculine and antepenultimate in the feminine; e.g. [əwˈtɔktu**n**] ~ [əwˈtɔktun-ə] ‘autochthonous-F.SG/M.SG’ both have

[4] We adopt /-ə/ as the UR for the [-ə] suffix simply for convenience. Since /e/, /ɛ/ and /a/ all reduce to schwa in stressless position, any of them could serve as the UR. For discussion, see Wheeler (2005:§2.3).

penultimately-stressed stems. In such forms, /n/-deletion is unusual (1/27 in our corpus, or 3.7%). The example just given is, thus, typical.

c. **Monosyllabic stems**. These often permit deletion, but in only about half (8/15, 53.1%) of the cases. Examples of deleting and non-deleting monosyllables were noted above: [ˈplan-ə] ~ [ˈpla] ‘flat-F.SG/M.SG’, but [ˈnɛn-ə] ~ [ˈnɛn] ‘child-F.SG/M.SG’.

d. **Other**. The remaining category, a large one, is “elsewhere,” meaning: /n/ in final position of a polysyllabic stem, with a stressed stem-final syllable, and not in the suffix /-in/. Here, deletion is frequent (390/410, 95.1%), but not the 100% observed with /-in/. An example is [kətəˈlan-ə] ~ [kətəˈla] ‘Catalan-F.SG/M.SG’.

As an example of what guided our experimentation on frequency-matching, we offer a specific prediction: if our test includes a set of feminine wug words that have stem-final stress and are unsuffixed (such as [prəˈtɔn-ə] or [ˈbrɔn-ə]), we expect the group of experimental participants collectively to give masculine responses involving /n/-deletion more frequently when the wug stem is polysyllabic (like [prəˈtɔ]) than when it is monosyllabic (like [ˈbrɔ]), thus matching the difference in lexical frequencies. More generally, the sample of responses from the participants, taken as a whole, will be the result of their stochastically matching the frequency patterns given in (5).

2.3 /r/-deletion

The consonant /r/ is deleted in word-final position or before the plural ending [-s]; see (6).

(6) a. */r/-deletion*

r → ∅ / ___ (s) $]_{\text{word}}$ ‘Delete [r] word-finally or before a final /s/’

b. *Derivations*

‘clear.M.SG*’*	*‘clear*-F.SG*’*	*‘clear*.M-PL*’*	
/klar/	/klar-ə/	/klar-s/	underlying representations
∅	—	∅	/r/-deletion
[ˈkla]	[ˈklarə]	[ˈklas]	surface representations

/r/-deletion has many lexical exceptions, which follow a four-way statistical pattern that turns out to be very similar to the pattern observed for /n/-deletion, see (7).

(7) *Environments for /r/-deletion*

a. **Particular suffixes**. /r/-deletion applies with perfect regularity (205/205, 100%) in forms with the suffix /-dor/ ‘agentive’, as in [ədministrə-ˈdor-ə] ~ [ədministrə-ˈdo] ‘administrator-F.SG/M.SG’.

b. **Penultimate-stress stems.** /r/-deletion seldom applies (2/24, 8.3%) in these stems. A typical example of non-application is [ˈprɔspər-ə] ~ [ˈprɔspər] ‘prosperous-F.SG/M.SG’.

c. **Monosyllabic stems**. /r/-deletion applies only about half the time (3/7, 42.9%) in monosyllabic stems, as with [ˈklar-ə] ~ [ˈkla] ‘clear-f./m.’ vs. [ˈpur-ə] ~ [ˈpur] ‘pure-F.SG/M.SG’.

d. **Other**. In the remaining cases (polysyllabic, not ending in /-dor/, finally-stressed), application is very frequent (250/256, 97.7%) but not as frequent as with /-dor/: [priˈmer-ə] ~ [priˈme] ‘first-F.SG/M.SG’.

As with /n/-deletion, our interest in /r/-deletion resides in whether speakers in a wug-test will tend to frequency-match the deletion rates seen in these four subsets of the lexical data.

2.4 /nt/-cluster simplification

In broad terms, word-final homorganic clusters are simplified; the full pattern is given in detail by Mascaró and Wheeler. Here, we focus on a single cluster, /nt/. When no suffix follows, underlying /nt/ is rendered as [n]. An example is the stem for ‘holy’, /sant/, which surfaces intact in the feminine [ˈsant-ə] but with loss of /t/ in masculine [ˈsan]. Our database has 40 stems with final underlying /nt/, and every one of them surfaces with final [n].[5] More generally, irrespective of paradigm structure, Catalan does not tolerate word-final [nt]. /nt/-cluster reduction is stated and exemplified in (8), which also gives the crucial ordering relationship with /n/-deletion.

(8) a. */nt/-cluster reduction*

t → Ø / n ___]$_{\text{word}}$ ‘Delete [t] word-finally after /n/.’

b.

‘holy-F.SG’	*‘holy.M.SG’*	*‘good.M.SG’*	
/sant-ə/	/sant/	/bɔn/	underlying representations
—	—	Ø	/n/-deletion
—	Ø	—	/nt/-cluster reduction
[ˈsantə]	[ˈsan]	[ˈbɔ]	surface representations

[5] Certain highly unassimilated borrowings, such as (*Microsoft) Paint* and *PowerPoint*,can be pronounced with final [nt] (Pons-Moll 2015); these do not appear in our Wiktionary source.

As can be seen, in rule-based phonology, /nt/-cluster simplification *counterfeeds* (Kiparsky 1968) /n/-deletion, since [n] rendered word-final by /nt/-cluster reduction is not deleted (8).

Irrespective of which analytic approach we take, the key point is that this data pattern involves *opacity* (Kiparsky 1973), since /nt/-cluster reduction creates surface exceptions to /n/-deletion. Starting with Kiparsky's original work, it has been widely suggested (see also Sanders 2003; Bowers 2019; Mayer, in press) that opacity, at least in some of its forms, is difficult to learn and likely to lead to restructuring by language learners. The research question here, then, is whether the opaque Catalan pattern is productive, and to the extent that it is not, what sort of "errors" (forms not conforming to the existing language pattern) might be committed by speakers when tested on novel stems. Such cases might take the form of just not applying /nt/-cluster reduction (i.e. /sant/ → [ˈsant]), or perhaps shifting (as Kiparsky originally suggested) to the transparent outcome, which would be [ˈsa], derived in rule-based phonology with feeding order: /nt/ → [n] → ∅.

2.5 [ʒ] ~ [tʃ] alternation

Catalan has Final Devoicing: voiced obstruents at the end of stems are rendered as voiceless when the stem is followed by no suffix, thus placing the obstruent in final position. We give the rule with illustrative derivations in (9).

(9) a. *Final Devoicing*

$[-\text{sonorant}] \rightarrow [-\text{voice}] \; / \; \_\_\_ \,]_{\text{word}}$ 'Devoice obstruents finally'

b. *Derivations*

'grey.M.SG*'*	*'grey*-F.SG*'*	*'fat*.M.SG*'*	*'fat*-F.SG*'*	
/griz/	/griz-ə/	/gras/	/gras-e/	underlying representations
s	—	—	—	Final Devoicing
[ˈgris]	[ˈgriz-ə]	[ˈgras]	[ˈgrasə]	surface representations

The case of interest here concerns a set of masculine-feminine paradigms where [ʒ] alternates with [tʃ], as in [ˈbɔʒ-ə] ~ [ˈbɔtʃ] (*[ˈbɔʃ]) 'crazy-F.SG/M.SG' Following Bonet and Lloret (1998) and Wheeler (2005), we take /ʒ/ be the underlying form. Given Final Devoicing, we would expect /ʒ/ to surface as [ʃ], not [tʃ]. The appearance of [tʃ] is *not* the result of any ban on final [ʃ], for there also exist non-alternating /ʃ/ stems, such as [ˈfluʃ-ə] ~ [ˈfluʃ] 'loose-F.SG/M.SG.' In rule-based phonology, one option is to add a special version of the Final Devoicing rule, applicable only to /ʒ/, that also changes continuancy, as in (10a). As (10b) shows, it must be applied before ordinary Final Devoicing.[6]

(10) a. */ʒ/ to [tʃ] Rule*

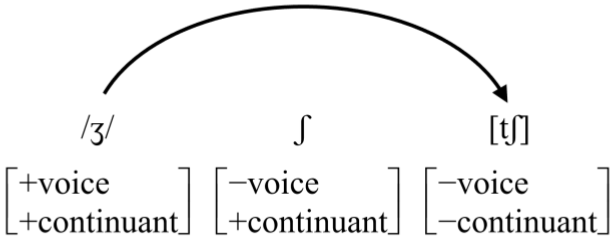
$$\text{ʒ} \rightarrow \begin{bmatrix}-\text{voice}\\ -\text{continuant}\end{bmatrix} \;\_\_\;]_{\text{word}} \qquad \text{‘Replace /ʒ/ by [tʃ] finally’}$$

b. *Derivations*

‘crazy-F.SG*’*	*‘crazy*.M.SG*’*	*‘grey*.M.SG*’*	*‘loose*.M.SG*’*	
/ˈbɔʒ-ə/	/ˈbɔʒ/	/ˈgriz/	/ˈfluʃ/	underlying representations
—	tʃ	—	—	/ʒ/ to [tʃ] Rule
—	—	s	—	Final Devoicing
[ˈbɔʒə]	[ˈbɔtʃ]	[ˈgris]	[ˈfluʃ]	surface representations

The shift of /ʒ/ to [tʃ] is an example of what Hayes and White (2015) call **saltation**, a term based on the Latin word for leaping. The underlying /ʒ/ of /bɔʒ/ surfaces as [tʃ], even though /ʃ/ (phonetically intermediate between [ʒ] and [tʒ]) is phonotactically legal in this position; thus, the /ʒ/ leaps across /ʃ/, as illustrated in (11).

(11) *Saltation in Catalan*



Since the work of Łubowicz (2002) and Ito and Mester (2003), saltation has been known to be underivable in “classical” Optimality Theory, defined as Prince and Smolensky (1993) with the Faithfulness constraints of McCarthy and Prince (1995). Saltation *is* derivable in OT with

[6] There are alternatives. Instead of /ʒ/ to [tʃ] we could have /ʒ/ to [dʒ], with the derivation completed by Final Devoicing. More fundamentally, Torres-Tamarit (2016) and Bonet and Lloret (2018) suggest that the underlying form for stems like [bɔʒ-ə] ~ [bɔtʃ] should be /bɔdʒ/, with [ʒ] derived by a rule of Intervocalic Spirantisation and [tʃ] by Final Devoicing. This analysis would require further apparatus, such as stem structure constraints, to explain why Catalan has no [ʒ] ~ [ʃ] alternations. In other words, to match the Catalan data pattern, the generalisation must be enforced that all stem-final phonetic [ʒ]s get derived from /dʒ/ rather than /ʒ/. If this alternative should prove correct, then a different theoretical question must be addressed, i.e. how language learners apprehend a principle like “all Xs are derived from Y”, perhaps itself a rather marked phenomenon; see McCarthy (2005) for discussion.

various non-standard Faithfulness constraints; for example, Hayes and White use the *Map constraints of Zuraw (2007, 2013).

Hayes and White also suggest that saltation is, in some sense, a marked phenomenon. It evidently only arises through quirky sequences of historical change and is never innovated per se as a phonological pattern.[7] Moreover, Saltation appears to be difficult to learn, as is shown both by cases of historical restructuring—"saltation repair"—given by Hayes and White, as well as White's experimental results in artificial grammar learning (2014, 2017). Lastly, a principled explanation of the marked status of saltation is available, based on principles of the "P-map", laid out by Steriade (2009) and developed formally by Zuraw (2013); deriving saltation requires constraint rankings or weightings that ban short phonetic paths like [ʒ] → [ʃ] more strictly than long paths like [ʒ] → [tʃ].

The research question here is whether the productivity of the /ʒ/ → [tʃ] process falls short of the lexical pattern, which is indeed exceptionless (about 15 stems, including some outside our database,[8] undergo /ʒ/ → [tʃ], and there are no stems at all with [ʒ] ~ [ʃ] alternation). Our expectation is that if the participants do "repair" saltation, they will do so in the natural way, devoicing final /ʒ/ to the phonetically close [ʃ] rather than to phonetically distant [tʃ]. Such forms, which mismatch the language pattern, may be interpreted as possible support for the view that saltation is hard to learn.

2.6 Probabilistic UR inference

In languages with phonological neutralisation, language learners who have not yet heard a complete paradigm are often in a position of needing to guess the underlying representation of a stem. To give an example, Dutch, like Catalan, has Final Devoicing. For this reason, a speaker who knows only the singular for a word like [ˈkaus] for 'stocking' cannot, in principle, know whether the UR is /kaus/ or /kauz/. If the former is chosen, then the plural form should be [ˈkaus-ən], which turns out to be correct; if the latter, a *[ˈkauz-ən].

One might imagine that the language learner who only knows neutralised forms would simply delay setting up the UR, waiting for other inflected forms to provide the crucial information. However, Ernestus and Baayen (2003) demonstrate that for Dutch this is not so; instead, speakers evidently construct some sort of statistical model that uses contextual cues to predict the voicing of the UR. For example, when the stem-final consonant is a velar fricative ([x]) in isolation, it is very likely that the UR will have /ɣ/, not /x/. This is because 97% of all Dutch stems ending in phonetic [x] have /ɣ/ as their UR. Similarly, when the preceding stem vowel is

[7] Indeed, this is so for the Catalan saltation, which arose from the diachronic events that are synchronically recapitulated in the analysis of Bonet and Lloret (2018:229-230), cf. Moll, (1952).

[8] There are 7 stems in our lexical database; our figure of 15 reflects additional forms listed in Wheeler (2005:12, 260).

long, it becomes more likely that the stem consonant will be voiced. With similar generalisations and suitable formalisation, a model that relies on these cues can predict the voicing of the URs with surprising if imperfect reliability. That Dutch speakers are tacitly aware of these generalisations is shown by Ernestus and Baayen's wug-test study, where the volunteered forms strongly followed the patterns seen in the lexicon. Indeed, UR-guessing seems to be a second area where speakers turn out to be good frequency-matchers.

Catalan, like Dutch, has extensive word-final neutralisation, raising the question of whether Catalan speakers can likewise use statistical cues to guess URs. In particular, when an isolation stem (here, a masculine form) ends in a stressed vowel, there are three plausible URs for it, because of /n/-deletion and /r/-deletion. Thus the hypothetical form [nəˈlo] might be underlyingly /nəˈlon/, patterning like [ˈsa] ~ [ˈsan-ə]; or it might be /nəˈlor/, like [ˈkla] ~ [ˈklar-ə]; or it might simply be /nəˈlo/, since there exist vowel-final stems like [ˈkru] ~ [ˈkru-ə] 'raw.M.SG/F.SG' Participants who tacitly adopt one of these three possible URs would render their choice detectable by providing feminine forms like [nəˈlon-ə], [nəˈlor-ə], and [nəˈlo-ə], respectively.

What cues might be available to assist Catalan speakers in making such guesses? We suggest that these could come from the quality of the preceding stem vowel. As Table (12) shows, there are strong asymmetries present based on this factor. For testing, we chose the vowels [ɛ], [o], and [u], which involve the strongest such asymmetries.

(12) *Guessing URs based on the rightmost stem vowel: lexical statistics*

Masculine-final vowel	*Feminine in* […n-ə]	*Feminine in* […r-ə]	*Feminine in* […-ə] *(hiatus)*
[ˈɛ]	120 100%	0 0	0 0
[ˈo]	38 12.4%	269[9] 87.6%	0 0
[ˈu]	4 30.8%	7 53.8%	2 15.4%

Thus, if Catalan speakers use vowel quality to project the URs of stem-final consonants, it is likely they would favour [n] for stems that end in [ɛ] and [r] for stems ending in [o]. There is no vowel that favours hiatus, but [u] is the vowel most compatible with this option.

In sum, the last of the four research questions we are addressing is whether Catalan speakers use probabilistic cues to guess URs. The particular cues we investigate involve the association of particular stem-final consonants with the preceding vowel quality.

[9] Of these, 205 had the suffix /-dor/; if these are omitted, the percentages become 37.3%, 62.7%, and 0%.

3. Experimental study

We first review (§3.2) our wug stimuli, which were designed to address the four research questions just given. Using these stimuli, we conducted two experiments. In Experiment 1, participants produced their favoured wug responses orally, which were then transcribed by the authors in IPA. In Experiment 2, we asked the participants to rate on a 1-7 scale the responses given most frequently by the participants in Experiment 1.

3.1. Participants

For Experiment 1, 43 participants were recruited. Six were excluded, for various reasons: their first exposure to Catalan occurred after age 3 (*n* = 2),[10] they self-identified as non-Central Catalan speakers (*n* = 2), or they failed a training task item (see §3.3; *n* = 2). This left 37 participants. For Experiment 2, 51 participants were recruited. Fourteen were excluded due to late age of acquisition (*n* = 4), self-identifying as non-Central Catalan speakers (*n* = 7), and/or failing at least one of the control items (§3.3, *n* = 3), leaving 37 participants. Eight participants completed both Experiments 1 and 2; this proved insufficient to draw conclusions about how responses to the two experiments are related.

In a post-experiment questionnaire, all included participants reported that they spoke Catalan at home during childhood, and the vast majority had completed their mandatory education (ages 5-16) in Catalan. We note that our participants were strongly multilingual—the median rate of Catalan usage was only 60%—and that this may have influenced our results (see §4.2.2). It would be logically possible to avoid such effects by recruiting a participant population consisting of monolinguals, but this seems unrealistic given that Catalonia is a highly multilingual society.[11]

Participants were recruited online with the assistance of government agencies and nonprofit organisations that promote the Catalan language; we also found some participants through word of mouth. Participants were compensated for their time with a USD$15 electronic gift card.

3.2. Materials and Design

3.2.1. Stimuli

The stimuli fell into two groups. For /n/-deletion, /r/-deletion, /nt/-final cluster reduction, and /ʒ/ final obstruent devoicing, the participants were provided with feminine forms (from which the

[10] For evidence that degree of language exposure in Catalan strongly influences the propensity to alternate, see Jovanovich-Trakál (2021).

[11] According to the Statistical Institute of Catalonia, in 2023, 43.4% of Catalans aged 15 years or older speak English at a conversational level (https://www.idescat.cat/indicadors/?id=basics&n=10367&tema=cultu). For Spanish, the percentage is 99.2%.

underlying representation is readily recoverable) and asked to produce the corresponding masculines. For the question of probabilistic UR inference, participants were provided with masculines and asked to produce feminines. Examples of our wug words are given in Table (13). It can be seen that the stimuli closely follow the scheme of §3.2.2: there are five conditions for particular phonological phenomena, of which /n/-deletion and /r/-deletion involve four subconditions. The rightmost column gives the choices that were offered to participants in Experiment 2. Most were drawn from from actual responses given in Experiment 1, but two additional candidate types, [bəˈzɛ-ə] and [nəˈlo-ə], were included in order to maintain full symmetry across the set of possible responses.

(13) *Experimental conditions and subconditions.*

Condition	*Subcondition*	*Presented to participants*	*Choices for Experiment 2*
/n/-deletion	frequent suffix /-inə/	[bəlunˈtrin-ə]	[bəlunˈtri], [bəlunˈtrin]
	monosyllabic	[ˈfrun-ə]	[ˈfru], [ˈfrun]
	penultimately stressed	[ˈdɔstun-ə]	[ˈdɔstu], [ˈdɔstun]
	other	[gəˈmɛn-ə]	[gəˈmɛ], [gəˈmɛn]
/r/-deletion	frequent suffix /-dorə/	[gruəˈdor-ə]	[gruəˈdo], [gruəˈdor]
	monosyllabic	[ˈlɛr-ə]	[ˈlɛ], [ˈlɛr]
	penultimately-stressed	[ˈsɔlir-ə]	[ˈsɔli], [ˈsɔlir]
	other	[kəˈnar-ə]	[kəˈna], [kəˈnar]
/nt/ final cluster reduction (opacity)	—	[mirˈbunt-ə]	[mirˈbun], [mirbunt], [mirˈbu] (feeding order)
/ʒ/ final obstruent devoicing (saltation)	—	[səˈlɔʒ-ə]	[səˈlɔt͡ʃ], [səˈlɔʃ]
Consonant restoration (UR inference)	[ɛ]	[bəˈzɛ]	[bəˈzɛn-ə], [bəˈzɛr-ə], [bəˈzɛ-ə]
	[o]	[nəˈlo]	[nəˈlon-ə], [nəˈlor-ə], [nəˈlo-ə]
	[u]	[pəˈmu]	[pəˈmun-ə], [pəˈmur-ə], [pəˈmu-ə]

We characterise the stimuli in more detail below.

/n/- and /r/-deletion: We chose our wug items to test the four environments by which deletion rates vary in the lexicon (see §2.2 and §2.3; research question (2a)).

a) **Frequent suffixes.** 20 polysyllabic stems, 10 with the suffix *-ina* and 10 with the suffix *-dora*. We expected high rates of deletion given the unanimous pattern found in the lexicon.
b) **Monosyllabic stems.** 20 monosyllabic stems, 10 ending in /n/ and 10 ending in /r/. We expected fewer cases of deletion, both because this matches the lexicon and because of the possible learning bias against alternation in initial syllables observed by Becker et al. (2012).
c) **Penultimately-stressed stems.** 20 polysyllabic stems with stem-penultimate stress, 10 ending in /n/ and 10 ending in /r/. These were chosen to test the hypothesis that stem-penultimate stress discourages the application of deletion.
d) **Other.** 20 polysyllabic stems with stem-final stress and not ending in *-ina* or *-dora*: 10 ending in /n/ and 10 ending in /r/. These were chosen as a sort of baseline condition, assessing preference for deletion in the absence of any of the first three factors.

/nt/-final cluster simplification: 10 polysyllabic stems, with stem-final stress, no identifiable suffix, and ending in /nt/. These were intended to test the productivity of the opaque interaction of /nt/-cluster simplification and /n/-deletion (research question (2b)).

/ʒ/ final obstruent devoicing: 10 polysyllabic stems, with stem-final stress, no identifiable suffix, and appearing with [ʒ] in the feminine. These were intended as a test of the productivity of saltation (research question (2c)).

Probabilistic UR Inference: 30 polysyllabic stems, with stem-final stress and no identifiable suffix. We chose 10 stems ending in each of the vowels [ɛ], [o], or [u] and presented as masculine forms. These were meant for use in "masculine-to-feminine" wug testing, assessing the ability of speakers to guess underlying forms in a way that frequency-matches the lexicon (research question (2d)). As discussed in §2.6, [ɛ] is the most likely of the seven vowels to take [n], [o] particularly favours [r], and [u] is the most likely simply to add the feminine ending, in hiatus.

In total, 130 different wug forms were created across 13 subconditions. The complete list can be found in the Supplementary Materials (available from OSF).

In order to obtain wug items that were maximally confound-free, we employed native-speaker judgement, assisted by an algorithm. The search was guided by the following principles:

i. **Phonotactic acceptability**: We sought wug items that would have reasonable phonotactic probability. For this purpose we created a software script that generated wug items intended to match the phonotactic properties of real words. The script worked by concatenating vowels, consonants and consonant clusters in a way that matches the frequency with which these sequences occur in our lexical database. The pool of potential wug words from which we chose our stimuli were the candidates that had relatively high probability.

ii. **Novelty**: We sought to avoid real words and wug words that closely resembled real words. We checked this not just for the wug forms presented to the participants, but also for all answers likely to be given in the test.[12] For example, the masculine wug form [pəˈmu] is not a real word, nor does it closely resemble one; and the same is true for the plausible feminine forms that participants might offer: [pəˈmur-ə], [pəˈmun-ə], and [pəˈmu-ə].

iii. **Variegation**: We sought a list that would contain a wide range of distinct consonants and vowels, to control for (by balancing out) any confounding factors that might affect the participants' judgements.

Candidate wug items were assessed intuitively using the native-speaker judgement of the second author, and sometimes modified to ensure naturalness, novelty, and variegation.

3.2.2. Frame paragraphs

We created 26 frame paragraphs, which were used in both experiments; an example is given in (14). The frames were presented auditorily and in written form on the screen, but the wug words were only presented auditorily; on the screen, they appeared as blanks. The frames were designed to give the participants multiple exposures to the wug word before they actually rendered the crucial wug-test response. The frame paragraphs are listed in the Supplementary Materials.

(14) *Sample frame paragraph ([bə.lun.ˈtri.nə] as adjective)*

> *[bə.lun.ˈtri.nə]*$_1$
>
> *Una obra __[bə.lun.ˈtri.nə]__*$_2$ *era una peça d'art on s'havien aplicat tècniques mixtes amb ornaments de metalls i pedres precioses. Al segle XV, un artista català va crear la primera escultura ____*$_3$*, feta de marbre, pedres precioses, i or. El primer quadre ___*$_4$ *no es va crear a Espanya fins al segle XVII.*
>
> '[bə.lun.ˈtri.nə]'$_1$
>
> 'A [bə.lun.ˈtri.nə]$_2$ work was a piece of art where they had applied mixed media with precious metals and stone ornaments. In the 15th century, a Catalan artist

[12] After piloting and completing data collection for Experiment 1, we detected two items that could be or were in fact existing low-frequency words: [mu.ɲi.ˈdo.rə] could have been interpreted as 'milker', derived from *munyir* 'to milk (V)'; and one of the candidates for [ˈklɔ.rə], [ˈklɔr], means 'chlorine'. Neither item elicited responses that diverged from the items of their respective subconditions: we obtained 100% deletion for [mu.ɲi.ˈdo.rə] (/-dor/ subcondition average = 100%) and 33.33% deletion for [ˈklɔ.rə] (monosyllabic r-deletion average = 26%).

created the first ___$_3$ sculpture, made of marble, precious stones and gold. The first ___$_4$ painting was not created in Spain until the 17th century.'

A frequent problem in wug test design is the possibility that the participants will interpret the stimuli as foreign words, a natural response given that novel words in a particular language are likely to be recent loans. This is particularly important in light of the fact that /n/-deletion, /r/-deletion and /nt/-cluster simplification are underapplied in Catalan loanwords (Pons-Moll 2015, 2021). For this reason, we crafted the content of the paragraphs to encourage participants to treat the wug forms as antiquated words in Catalan that had gone out of use, rather than loanwords.

The key task for participants was to internalise the wug form as much as possible, then provide a response inflecting it in the opposite gender from what had been provided (Experiment 1) or to rate a number of possible candidates for the opposite gender (Experiment 2). The required gender could be easily apprehended by participants, since there were multiple words in the frame that grammatically agreed with the wug word. For instance, in (14) the words *el* 'the', *primer* 'first', and *quadre* 'painting' are grammatically masculine and essentially force the choice of a masculine counterpart of *bə.lun.ˈtri.nə*. Half of the frames elicited a noun, the other half an adjective.

Participants first heard the wug in isolation (position 1), meant for high audibility and clarity, followed by the wug used in a sentence (position 2). Position 3 was left as a gap in Experiment 1; it required participants to repeat the wug word. This served as a check that they had internalised it correctly. Trials in which participants did not repeat the wug word correctly were discarded (*n* = 46/962). In Experiment 2, where no recording was collected, Position 3 was simply pronounced again as part of the frame. Position 4 implemented the actual wug test: participants were required to speak their response in Experiment 1 and choose from among options ((13) above) for Experiment 2.

There were a total of 130 wug words and 26 frame paragraphs, hence five different words per frame. We paired each wug word with a single frame in pseudo-random fashion, such that the wug words for testing /n/ and /r/-deletion, /nt/-reduction, and /ʒ/-devoicing were placed in one of 20 feminine-to-masculine frames and the vowel-final wug words for testing probabilistic UR inference were placed in one of six masculine-to-feminine frames. For the wug words ending in -*dora* and -*ina*, we chose frames compatible with nouns and adjectives, respectively. Other than those wug items, half of the remaining wug words were randomly placed in nominal contexts, the others in adjectival contexts.

We divided the resulting 130 frame + wug combinations into five scripts, such that within each script, participants would encounter two stems each from our 13 subconditions ((13)). Each frame was included exactly once in each script. In the experiment itself, participants were assigned a particular script in pseudo-random fashion, respecting the requirement that the five scripts should be used in nearly equal numbers.

The 130 frame + wug combinations were recorded by the second author in a sound booth using Praat (Boersma & Weenink 2024) with a mono audio channel, 44.1 kHz sampling rate, and 16 bits per sample. The recording was made in a way meant to ensure prosodic realism and maximum uniformity. Thus, for (14), the sentence containing the wug was recorded separately for each of the five wug forms affiliated with this frame, then spliced to a single recording of the remainder of the frame. Each gap was edited to be 1.5s long. All recordings were modified to have 72 dB of intensity. The audio files are available in the Supplementary Materials.

3.3. Procedure

The experiments were created and distributed using LabVanced (Finger et al. 2017). Participants received a link to the experiment and participated online using their own computer.

3.3.1. Experiment 1: Production task

Each participant completed 26 test trials (two items per subcondition). An example trial was given in (14). Before the start of the experiment, each participant saw two practice trials using similar paragraphs, one with a real masculine word ([bu.ˈta.nik] ‘botanist.M.SG’), and one with a feminine wug word that did not exhibit any alternation of interest, [ˈgrɛ.zə]. Participants were required to inflect the real word as expected in order to be included in the study. Participants were not allowed to skip trials. Their responses were timed (end of sentence frame to click). In total, participants produced four training items (two repetitions and two inflected forms), and 52 test items (26 repetitions and 26 inflected forms). The experiment took the participants on average 44 minutes.

After the experiment, participants were asked a number of demographic and linguistic questions, given in the Supplementary Materials. Some were intended to ensure that the included participants were exposed to Catalan during infancy, spoke Catalan at home, and self-identified as Central Catalan speakers. We additionally asked their age, sex, educational level, and how often they speak in other languages during a typical week.

Lastly, we asked participants to rate from 1 to 7 whether their responses throughout the experiment were “1 = purely intuitive” or “7 = based on conscious reasoning.” Our interest in asking this is based on the possibility, emphasised by Moreton and Pertsova (2023), that participants return different answers to wug test questions when they reason through their response consciously rather than relying on intuition.

3.3.2. Experiment 2: Acceptability judgement task

The procedure was identical to Experiment 1, except that the first two gaps in the frame paragraph contained the prerecorded wug word while the last gap prompted participants to evaluate two or three prerecorded candidates on a Likert scale from 1 to 7. For the choices made available to the participants, and the research questions thereby addressed, see §3.2.1.

Before the start of the experiment, each participant underwent three practice trials. For the real word *botànic* ‘botanist.M.SG’ they rated a correct inflected form ([bu.ˈta.ni.kə]) and an incorrect one ([bu.ˈta.ni.ə]). For the wug word [ˈgrɛ.zə] they were given the (very likely) choice [ˈgrɛs] and an unlikely choice [ˈgrɛ]. There was also a practice trial with the feminine wug form [əz.ˈma.βə], which, following Catalan phonology (Wheeler 2005:§10.3) has two acceptable masculine forms, namely [əz.ˈmap] or [əz.ˈmaw]. The purpose was to ensure participants understood that they could accept multiple candidates by rating them all high, or reject multiple candidates by rating them all low. The practice trials were accompanied by hints about how to rate these words (e.g., “for option (a), [ˈgrɛ], most Catalan speakers would assign a low score (1 or 2)”, for [əz.ˈma.βə]) “sometimes, there may be more than one plausible answer”).

There were four control items, embedded in similar paragraphs. These were interspersed with the test items and contained real words with only one possible correct answer. Two were masculine forms (*catedràtic* ‘professor.M.SG’*, filòsof* ‘philosopher.M.SG’) and two were feminine (*analítica* ‘analytical.F.SG’, *autònoma,* ‘autonomous.F.SG’). None of them exhibited the alternations of interest. Participants had to rate the correct candidate higher than the incorrect candidate in order to be included in the study. For example, for *analítica* the candidates were [ə.nə.ˈli.tik] (correct) vs. [ə.nə.ˈli.ti] (incorrect). In total, participants had to evaluate six training items (within three trials), 60 test items (within 26 trials), and eight control items (within four trials). The experiment took on average 21 minutes.

After completing Experiment 2, participants were asked to complete the same questionnaire used in Experiment 1 (§3.3.1).

4. Results and Discussion

For Experiment 1, the recordings were transcribed by two phonetically-trained listeners with disagreements adjudicated by a third native-speaker transcriber. Forty-six tokens were excluded due to the participants incorrectly repeating the wug form (blank 3 of frame (14)). This left 916 tokens to be included in the analysis. For Experiment 2, we report raw rating scores[13] In total, 2516 ratings were reported across the various wug items and candidates. Full results for both experiments, including individual participant responses and the R scripts for statistical analysis, are available in the Supplementary Materials.

[13] We also performed the analyses using Z-score normalised ratings; however, the statistical patterns were the same.

Given the similarity of the results, we report the two experiments together.

4.1. Non-phonological responses

While our study was focused on outcomes related to phonology, we also cover results that did not fit into the preconceived scheme of our experiments.

Notably, some participants employed Catalan morphology in unexpected ways. As noted in §2.1, there are two forms of irregular masculine inflection. The suffix [-u] is found in examples like [buˈra.t͡ʃ**-u**] ~ [buˈra.t͡ʃ-ə] 'drunk-M.SG/F.SG'. It occurs in 0.95% (55/5761) of the paradigms in our database. The irregular masculine schwa ending, seen in [əˈlegr**-ə**] ~ [əˈlegr-ə] 'happy-M.SG/F.SG', is found in 0.71% (41/5761) of paradigms in our database. In Experiment 1, we found that 1.6% (15/916) of participant responses employed [-u], a higher value than the percentage observed in the Catalan lexicon. For [-ə], such responses were even more frequent: 12.8% (117/916).

We offer a conjecture for why participants volunteered such morphologically unusual masculine forms: since the irregular endings provide a vowel after the stem, they allow a participant to circumvent the task of answering the phonological question at hand. Based on this possible explanation, we will refer to responses with [-u] or [-ə] as **avoidant**. Avoidant responses in wug-testing have been noticed before, particularly among children. Do (2018) found abundant cases in which Korean children chose a morphologically somewhat inappropriate response that allowed them to bypass the choice of what form of phonological alternation to deploy. See also Zamuner, Kerkhoff and Fikkert (2012) and Kerkhoff (2004, 2007) for Dutch-speaking children, Pérez-Pereira (1989) for Spanish-speaking children, and Kim (2025) for Spanish-speaking adults.

We find that avoidance is related to sparse lexical data; that is, a low number of exemplars for a given subcondition in the lexicon predicts avoidant responses ($F(1,12) = 21.8$, $p < 0.001$). We think this pattern is sensible—native speakers are tacitly aware that any guess they might give is based on few data points and so are reluctant to guess. We know of no phonological models that predict this correlation and see it as a challenge for future research.

There was also a non-significant trend ($F(1, 915) = 2.94$, $p = 0.087$) such that when participants gave avoidant responses in Experiment 1, their response times (total time of trial – total time of recorded frame) were somewhat longer: 23.3 s for avoidant responses vs. 20.9 s for non-avoidant. Combined with the lexical frequency effect, the greater response times suggest greater uncertainty: perhaps they reflect a process whereby participants tacitly weighed multiple phonological options, felt unable to choose between them, and ultimately opted for a non-phonological solution.

We are intrigued that the participants’ avoidant responses were predominantly with [-ə], not with [-u], even though [-u] is slightly more common in the lexicon.[14] This suggests that not all of the [-ə] necessarily reflect the [-ə] ending seen in the lexicon; rather, the bulk of them may be cases of the participant simply repeating back the form that was presented to them. The literature on avoidance includes several cases of avoidance by literal repetition (e.g., Pérez-Pereira 1989; Zamuner, Kerkhoff & Fikkert 2012; Kim 2025).

Aside from avoidant responses, we observed a number of responses (9%) that might be called **aberrant**, as they lacked any sort of reasonable interpretation. One instance occurred with [ʎuˈdaʒə], intended to test for saltation (§2.5): six participants responded with deletion, i.e. [ʎuˈda]. Conceivably, this response, which has no precedent in the Catalan lexicon, was also avoidant; though bizarre, it lets the participant avoid choosing between [ʎuˈdatʃ] and [ʎuˈdaʃ]. Just as with avoidant responses, aberrant responses involved greater average response time (31.2 s for aberrant responses vs. 20.4 s for non-aberrant (*F*(1, 915) = 20.9, *p* < 0.001). However, aberrance is not significantly predicted by low lexical attestation (*F*(1, 12) = 1.19, *p* = 0.30; see Supplementary Materials).

Below, in reporting response rates for individual phonological phenomena, we exclude non-phonological responses (both avoidant and aberrant) from the denominator.

4.2. Results by subcondition

4.2.1 /n/-deletion results

The first point that emerges from the /n/-deletion data is that this alternation is, to some degree, productive: indeed, the overall deletion rate was 80%. This is of interest because previous work (Wheeler 2005, Pons-Moll 2015) has suggested that /n/-deletion is unproductive, observing in particular that it usually fails to apply in loanwords. The high application rate we obtained is perhaps to be attributed to our efforts to characterize the wug forms as unknown native words, thus overcoming the natural tendency to treat loanwords faithfully.

The rates of /n/-deletion differed substantially across subconditions, as shown in Figure 1.

[14] Note that in most of the masculine schwa forms, the schwa can be treated as the result of epenthesis, repairing illegal final sequences (e.g., /əˈlegr/, → [əˈlegrə] ‘happy.M.SG’; Wheeler 2005:§8.3); there are very few cases like [puˈɛt-ə] ‘poet-M.SG/F.SG’ or [pəriuˈðist-ə] ‘journalist-M.SG/F.SG’, that cannot be derived by epenthesis (final [t] and [st] being phonotactically legal). None of the schwas in our wug test could be epenthetic. Therefore, the lexicon-wug test disparity is perhaps even larger than indicated above.

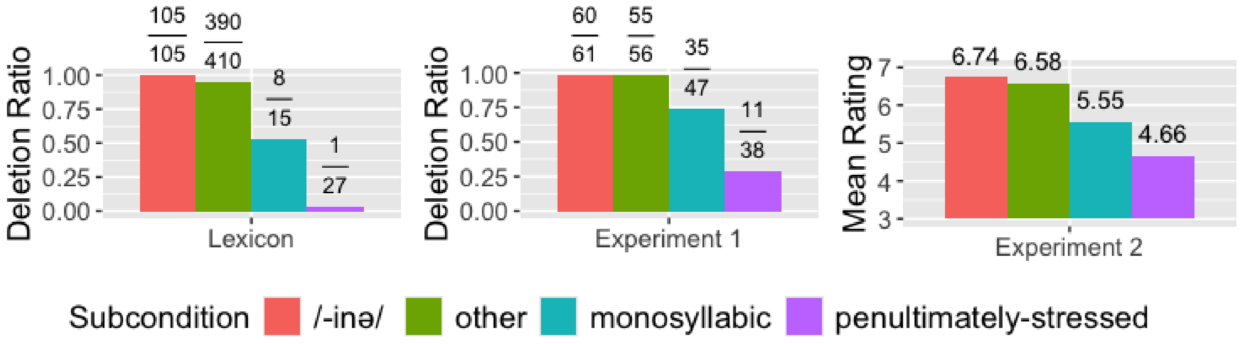


Figure 1. /n/-deletion in the lexicon, Experiment 1 (production), and Experiment 2 (ratings)

In rough terms, these data appear to reflect frequency matching. The environments where /n/-deletion applies most often in the lexicon match the environments where speakers most often applied /n/-deletion in Experiment 1 (production), following the order: *frequent suffix ≥ other > monosyllabic stems > penultimately-stressed stems.*[15] In Experiment 2, where the ratings again followed the pattern *frequent suffix ≥ other > monosyllabic stems ≥ penultimately-stressed*. The correlation of the plotted values for each subcondition from the Lexicon and Experiment 1 is r = 0.990 (p < 0.001); for Experiment 2, r = 0.996 (p < 0.001).

It is informative to look at cases that deviate from frequency matching in a systematic way. From Figure 1, it appears that in two conditions, participants overproduced and overrated deletion relative to the lexicon, namely monosyllabic stems and penultimately-stressed stems. Overapplication to monosyllables is particularly interesting in light of Becker et al.'s (2012) experimental evidence for a UG bias *against* alternation in monosyllables. We conjecture that here, two countering effects pair up to override this bias. **Simplicity bias** (Moreton & Pater 2012) would predict that because deletion is favoured overall in the lexicon (504/557 total cases), monosyllabic and penultimately-stressed stems might sometimes just follow the simplest available generalisation. Another possibility is **attestation bias** (Albright & Hayes 2003; Siah 2024): speakers will not take a generalisation as seriously when the data supporting them are sparse. In our database there are only 16 monosyllabic stems (9 with deletion) and 27 penultimately-stressed stems (1 with deletion). The numbers for the other conditions are much larger: 105/105 for the suffix [-ina] and 385/391 for 'other'.

[15] In statistical testing, for the Lexicon and Experiment 1 data, we used logistic regression models, incorporating random effects for Experiment 1. For Experiment 2, a cumulative-link mixed-effects model was employed. Post-hoc analyses were performed using Tukey's pairwise comparisons. In both the Lexicon and Experiment 1, there were significant differences between the following adjacent subconditions, as illustrated in Figure 1: *other > monosyllabic stems* (ps < 0.05) and *monosyllabic stems > penultimately-stressed stems* (ps ≤ 0.01). For Experiment 2, there was only a significant difference between the adjacent subcondition *other > monosyllabic stems* (p < 0.01). Across the Lexicon and the two Experiments, all non-adjacent subconditions were significantly different from each other (ps < 0.01). For statistical details of all comparisons discussed in this paper, see Supplementary Materials.

Our results on /n/-deletion replicate and extend those of the only previous wug-testing work on Catalan known to us, namely Jovanovich-Trakál (2021), who found modest productivity for /n/-deletion in plural alternations like [kləˈfon-s] ~ [kləˈfo(n)] ‘wug.M-PL/M.SG’. Our own participants showed higher productivity for /n/-deletion. We suggest that this may arise from (a) our use of spoken language rather than orthography; (b) our more stringent criterion for Catalan exposure in childhood; (c) the fact that we tested adults and Jovanovich-Trakál children.

4.2.2 /r/-deletion results

Again, we see evidence of productivity (overall application rate: 52%), despite the typical pattern of nonapplication in loanwords. The variation in application rate across contexts, seen in Figure 2, suggests frequency matching, just as for /n/-deletion. The correlation of the values for the subconditions from the Lexicon and Experiment 1 is ($r = 0.915$, $p = 0.08$); and for Experiment 2 is ($r = 0.881$, $p = 0.12$).[16]

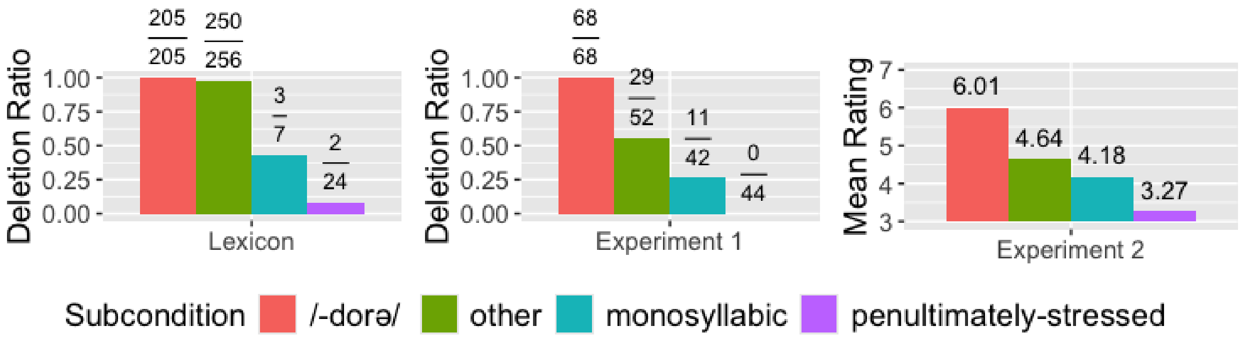


Figure 2: /r/-deletion in the lexicon, Experiment 1 (production), and Experiment 2 (ratings)

An interesting asymmetry surfaces when we compare deletion for /n/ and /r/: /n/-deletion closely matched the lexical frequencies but /r/-deletion matched only in *relative* terms. Speakers consistently disfavored [r]-deletion relative to [n]-deletion, as Figure 3 shows.[17]

[16] Using the same analyses as /n/-deletion, we found significant differences, in both the Lexicon and Experiment 1, between the following adjacent subconditions (Figure 2): *other* > *monosyllabic stems* (ps ≤ 0.001) and *monosyllabic stems* > *penultimately-stressed stems* (ps ≤ 0.05). In Experiment 1, there was also a significant difference of */-dora/* > *other* ($p < 0.001$). For Experiment 2, there was only one significant difference between the adjacent subcondition */-dora/* > *other* ($p < 0.001$). Across the Lexicon and the two Experiments, all non-adjacent subconditions were significantly different from each other (ps < 0.01).

[17] We modeled the combined /n/ and /r/ data using mixed effects logistic regression models and pairwise comparisons and found consonant type to be a significant predictor, such that participants were more likely to delete /n/ compared to /r/ in Experiment 1 ($p < 0.001$) and were more likely to rate /n/-deletion higher than /r/-deletion in Experiment 2 ($p < 0.001$).

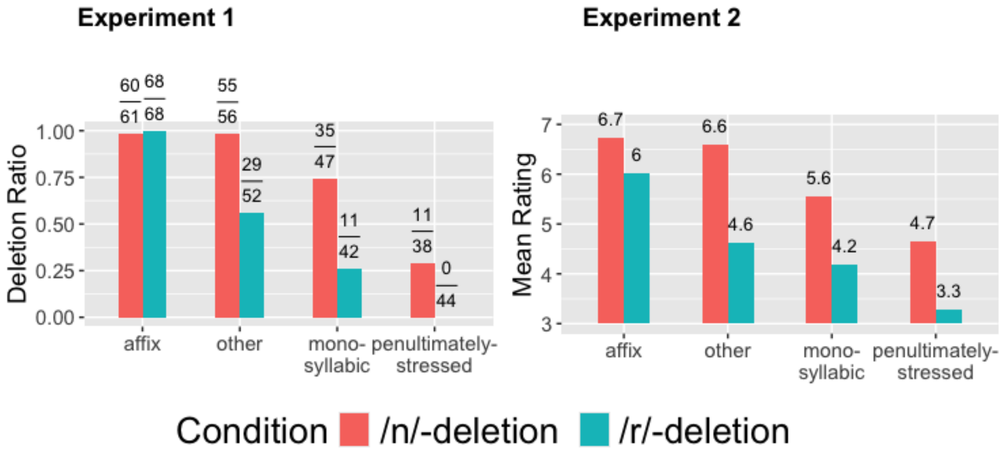


Figure 3: /n/- and /r/-deletion in Experiment 1 (production) and Experiment 2 (ratings)

One possible explanation for this difference is dialect variation, described in Wheeler (2005). Speakers of Central Catalan encounter speakers of another major dialect, Valencian, which lacks /r/-deletion; whereas /n/-deletion is pan-dialectal.

Another possibility is that participants were influenced by orthography, an effect observed in earlier research (Daland, Oh & Kim 2015; Kawahara 2018). In Catalan, /n/-deletion is spelt out: [ˈsan-ə] ~ [ˈsa] is spelt *sana* ~ *sa*, but /r/-deletion is not spelt out: [ˈklar-ə] ~ [ˈkla] is spelt *clara* ~ *clar*. Our participants may have been constructing appropriate orthographic representations for what they heard, preferring to pronounce these representations faithfully. Developing a model that expresses and incorporates this orthographic influence in speakers' verbal responses seems worth pursuing but is beyond the scope of this paper.[18]

4.2.3 /nt/-cluster simplification results

Our results show that /nt/-cluster simplification is productive: a majority of responses (52%) in Experiment 1 were of the form /nt/ → [n], Figure 4. However, we were surprised by the prevalence of [nt] answers: these formed 42% of the responses in Experiment 1, and in Experiment 2, [nt] was actually rated numerically higher than the expected [n]; see Figure 4.

[18] Yet another explanation for the /n/ vs. /r/ difference is that our stimuli inadvertently included items in which the outcome could be influenced by having two /n/s or two /r/s in the same word; a so-called “OCP” effect (McCarthy 1988 et seq.). Our statistical testing suggest a significant effect of OCP, in the single context [rVr], but the difference goes in the opposite direction, i.e., an OCP effect would predict more /r/-deletion than /n/-deletion, and the contrary is true in our data.

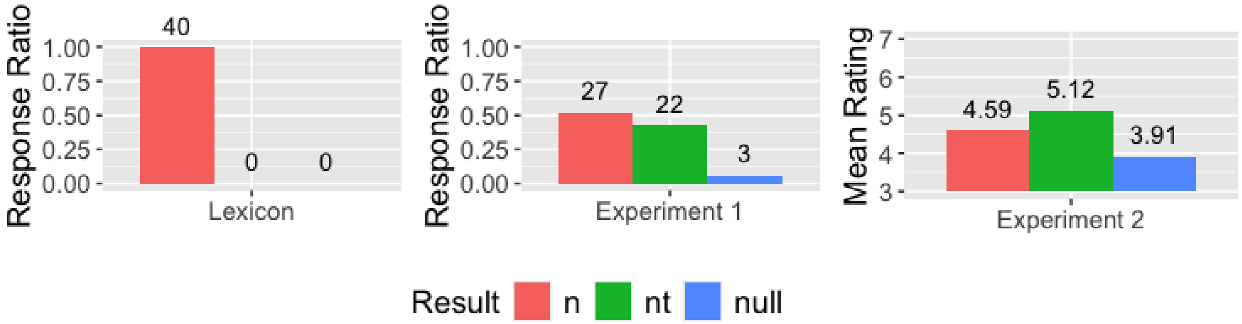


Figure 4: /nt/-cluster simplification in the lexicon, Expt. 1 (production), and Expt. 2 (ratings)

The frequency with which /nt/ is retained is surprising because the phenomenon is exceptionless in the lexicon, and because /nt/ simplification can plausibly be regarded as more natural than /n/- or /r/-deletion (see Pons-Moll 2015): it is typologically more common (Bonet et al. 2005), and imposes a lesser Faithfulness cost in terms of perceptual salience (Steriade 2009).

We conjecture three possibilities for the surprising degree to which participants preferred [nt]: (i) Exposure to other languages or Catalan dialects that do not have /nt/-cluster simplification (Wheeler 2005:221) weakens the native-language phonotactic constraint banning final [nt]; for a possible case of L2 learning leading to “weaker” phonology in L1 see Dmitrieva et al. (2010). (ii) Orthographic influence, as above: /nt/-cluster simplification is not spelt out. (iii) Opacity repair, discussed in the following paragraph.

The original reason we included /nt/-cluster simplification in the experiment was to test whether opaque phonology (here, counterfeeding of /n/-deletion) can be productive. To see how our results bear on this question, we examined the proportion of participants that fell into the patterns outlined in Table (15). In calculating the proportions in the table, we omitted participants who did not provide enough phonological responses (non-avoidant, non-aberrant) to make comparison possible; we also omitted ca;ses where the participant gave different responses for either the /nt/ pair or the /n/ pair.

(15) *Participant response patterns for /nt/-cluster simplification and /n/-other deletion.*

Pattern	*Fraction of participants in Expt. 1*	*Fraction of participants in Expt. 2*
a. *Opaque* nt → n, n → ∅	13 (46%)	7 (32%)

b. *Suppress /nt/ simplification* nt → nt, n → ∅	13 (46%)	12 (55%)
c. *Feeding* nt → ∅, n → ∅	1 (4%)	3 (14%)
d. *Delete /nt/* nt → ∅, n → n	1 (4%)	0 (0%)
e. *Suppress /n/-deletion* nt → n, n → n	0 (0%)	0 (0%)
f. *Fully faithful* nt → nt, n → n	0 (0%)	0 (0%)
g. *Inconsistent*	2	13
h. *Other*[19]	7	2

The Experiment 1 data suggest that counterfeeding opacity (15a) can be quite productive: it is found for 46% of the participants. Of the remaining cases, many were of type (15b): [mirˈbuntə] → [mirˈbunt], [gəˈmɛn-ə] → [gəˈmɛ]. As suggested above, this could be opacity-related—if you do not apply /nt/ deletion, the resulting output keeps /n/-deletion transparent. Only two participants volunteered the [mirˈbuntə] → [mirˈbu] pattern (15c-d). The [mirˈbu] outcome matches the prediction made in Kiparsky's (1968) pioneering work: by applying /nt/-cluster simplification and /n/-deletion in feeding order, it minimizes opacity. However, this outcome occurred so rarely that we are tempted to classify it as an aberrant response, similar to [ʎuˈdaʒə] → [ʎuˈda]. In the present case, the favoured response to opacity is not a shift to transparent rule order, but rather suppression of the feeding process.

The Experiment 2 data roughly match those of Experiment 1, though with slightly lower acceptance of the opaque pattern and a higher level of inconsistency, perhaps related to the participants' willingness to accept the stimuli as representing a different dialect.

4.2.4 Saltation results

We included the saltatory alternation [ʒ] ~ [tʃ] to test the hypothesis (2c) that saltation is unstable and vulnerable to restructuring. Our findings are summarized in Figure 5.

[19] This represents cases where the participant produced so many aberrant or avoidant responses that no comparison was possible.

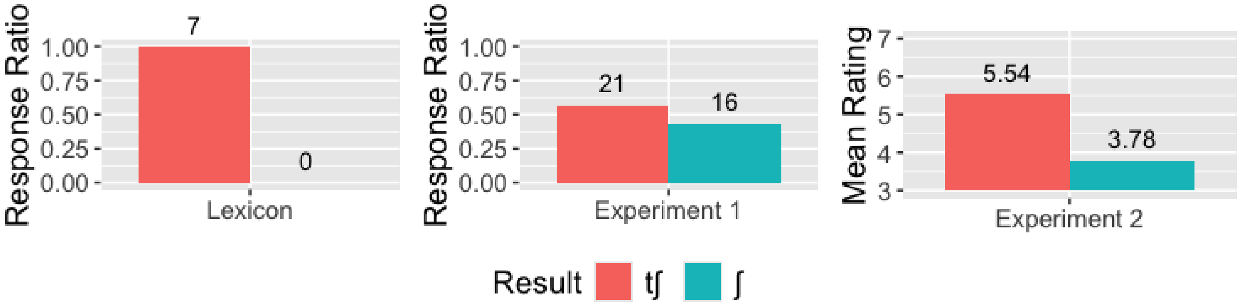


Figure 5: [ʒ] ~ [tʃ] alternation in the lexicon, Expt. 1 (production), and Expt. 2 (ratings)

To start, , saltation does seem to be productive in the sense that the majority of responses (57% or 21/37) took the saltating pattern [ʎuˈdaʒə] → [ʎuˈdatʃ]. However, many speakers produced forms that repaired saltation, like [ʎuˈdaʒə] → [ʎuˈdaʃ]; this was 43% or 16/37 of the phonological responses. The saltation repair occurred despite the fact that forms with [ʒ] ~ [ʃ] are not attested in the lexicon nor in any other dialect of Catalan. It is not compelling evidence for a learning bias against saltation, since there is an alternative explanation based on there being so few forms to learn from in the lexicon (only about 15; §2.5). The paucity of lexical forms may also explain the fact that the number of non-phonological responses (37/74 total) was higher than for in any other condition.

In Experiment 2, the [ʃ] outcomes were rated surprisingly low in light of the Experiment 1 results. This is in line with the predictions about production vs. acceptability tasks made by Smolek and Kapatsinski (2018).

The simplest saltation repair would have been non-alternating [ʎuˈdaʒ]. We suggest that this outcome never arose because final devoicing remains a powerful phonotactic principle of the language; for instance, unlike any of the other processes discussed here, final devoicing is a characteristic of Catalan-accented L2 speech (Pons-Moll 2015).

4.2.5. Probabilistic UR inference results

Our goal was to see if Catalan speakers would behave like the Dutch speakers studied in Ernestus and Baayen (2003) in using probabilistic cues from the phonological shape of stems to infer likely phonological underlying representations for novel stems. We hypothesized that varying the final vowel of a stem might lead them to set up URs that are most likely given that vowel, which would become apparent in feminine forms provided in response to a masculine. Recall from §2.6 that, in the lexicon, [ˈɛ]-final stems prefer [n] in the feminine, [ˈo]-final stems prefer [r], and [ˈu]-final stems generally prefer [r] but also tolerate hiatus. Thus, we expected responses like [bəˈzɛ] ~ [bəˈzɛnə], [nəˈlo] ~ [nəˈlorə], and, perhaps, [pəˈmu] ~ [pəˈmuə].

Nothing of the sort appeared; rather, responses with /n/ completely dominated the outcomes. This is shown in Figure 6, which compares the Lexicon with the responses and ratings given in Experiments 1 and 2.

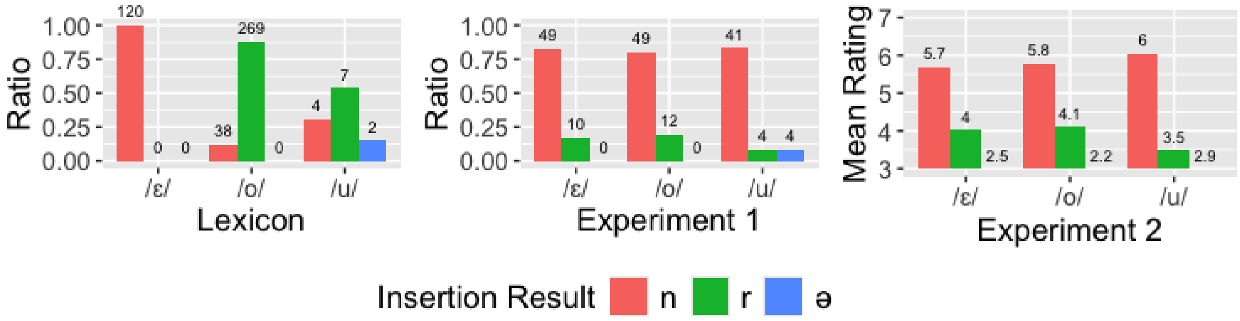


Figure 6: Probabilistic UR inference results in the lexicon, Expt. 1 (production), and Expt. 2 (ratings)

There are small differences between subconditions that fall in the expected direction, but these fall short of statistical significance; for details see the Supplementary Materials.

The outcome is reminiscent of the greater productivity of /n/-deletion relative to /r/-deletion (§4.2.2), and one of our explanations may be applicable. We suggest that when participants hear a vowel-final form like [bəˈzɛ], they posit an orthographic form like *bazè*. This is compatible with an orthographic feminine like *bazena*, since, as noted earlier, /n/-deletion is spelt out in the orthography. However, orthographic *bazè* could not correspond to feminine *bazera*, since /r/-deletion is not spelt, i.e., masculine forms are spelt with a final *r* that is not pronounced. The further possibility of *bazè* ~ *bazea* is not likely in any event, since few vowel-final masculines have simple schwa-adding feminines, even when the final stem vowel is [u], the maximally favourable environment.

Thus, we think that Ernestus and Baayen's claim that speakers can frequency-match the lexicon in inferring URs may be correct, but in Catalan the measurement of this effect is overwhelmed by the orthographic factor.

4.3. Studying the individual speaker

The assumption that speakers of a particular dialect share their linguistic systems in full detail has been questioned; see Jo (2024) for a comprehensive review. In light of this, we examined whether our experimental data offers any evidence for meaningful differences between participants.

4.3.1. Traits reported by participants

We begin with traits obtained from the demographic questionnaire (§3.3.1). With one exception, our statistical testing finds no effects from any of the following: sex, level of education, language of education, knowledge of foreign languages, and whether participants reported making their responses consciously or intuitively (Moreton & Pertsova 2023). The one factor that seems to matter is age: older participants are more willing to apply /r/-deletion ($\beta = 0.03$, $p = 0.02$).[20] See Supplementary Materials. We conjecture, in light of Catalan history, that such speakers received less of their education in Catalan and that the effect of orthographic suppression of /r/-deletion, discussed in §4.2.2, is weaker for them.

4.3.2. Consistency of responses and the idiolect mixture hypothesis

An alternative explanation for the observed variation in gradient phonology is *idiolect mixture*: the idea that every speaker possesses one single, nonstochastic grammar (it generates unique outputs), and that the appearance of variation in the outcome of a wug-test experiment is the consequence of mistakenly pooling results obtained from individuals with different grammars. Under this view, studying variation among individuals becomes very important, since it forms the sole basis for explaining variation in the aggregated data.[21]

This issue can be addressed with our dataset, since we arranged the test so that every participant responded to two different wug items for each subcondition. The prediction of the idiolect-mixture hypothesis is that participants should always give the same answer for every pair. For example, for /n/-deletion in penultimately-stressed stems, participant 955445 responded [ˈbatun] to [ˈbatunə] and [ˈsoðən] to [ˈsoðənə], consistent with a grammar that always avoids /n/-deletion in such stems.

However, the data show that participants frequently volunteer different answers under the same (sub)condition. For example, Participant 904843 responded [ˈbatu] to [ˈbatunə], but [ˈsoðən] to [ˈsoðənə], suggesting that 904843 internalized a stochastic grammar, and was simply drawing from the probability distribution that this grammar generates. Of course, 955445 too might also have internalized a stochastic grammar, and the forms elicited just happened to show the same pattern. Only 67% of responses involved identical response types, far short of the 100% predicted by the idiolect-mixture hypothesis. Indeed, the value is not much lower than the

[20] We used a mixed-effects logistic regression model with fixed effects of each demographic factor and a random intercept for Subject.

[21] We have not seen the idiolect mixture hypothesis put forth as a systematic principle. Halle and Vergnaud (1981) suggested it as an explanation for variation patterns in Finnish vowel harmony.

statistically expected value, 72%,[22] obtained by sampling responses at random from the observed probability distribution.

The Experiment 2 data likewise support the view that gradience is found within the individual speaker. For the possible candidates, we often find that the medial region of the rating scale is well-populated, reflecting gradience. Otherwise, either all ratings are high, or all are low. There are no candidates for which both extremes are well represented.

4.4 Inferences from MaxEnt modelling

We conclude with a phonological analysis of our data in the framework of MaxEnt Optimality Theory (Goldwater & Johnson 2003). We used this model as the core of larger model, also in MaxEnt, intended to address the following issues:

(16) *Four hypotheses addressable with a MaxEnt model*

a. Speakers differ in their **phonological preferences**, as reflected by providing parallel responses to wug words from the same subcondition. We have just established that it is impossible to attribute *all* variation to idiolect differences. However, it remains a possibility that more subtle idiolect differences are present.

b. Whether individual participants show an across-the-board preference for **Faithful responses**; i.e. those that avoid phonological alternation.

c. Whether there are **differences in avoidance behaviour** among participants, both in general and with regard to the particular suffix ([-u] or [-ə]) used for avoidance.

d. The possibility of **self-priming**: does giving a particular response to a recent wug from a given subcondition increase the chance of the same type of response for a wug of the same subcondition?

4.4.1 A MaxEnt phonological grammar

Our testing is based on a fairly ordinary MaxEnt OT grammar for this part of Catalan phonology. The grammar is assessed on its own, then supplemented by other model factors that address the four questions in (16).

Our grammar was set up to assign probabilities to candidates for inputs corresponding to the experimental subconditions. We fitted the weights of the grammar in two ways, first trying to match the frequencies of our lexical database, then the Experiment 1 response rates. In the interest of realism, we added some additional candidates: (1) about 4000 schematic forms

[22] For example, if one outcome has probability x, the other has probability (1 − x), the probability of getting two responses of the same response type is $x^2 + (1-x)^2$

(approximately the correct number) where none of the phonological processes we have studied are applicable. These provide information about the frequency with which Catalan employs the irregular masculine [-u] suffix (55 cases in our database) and [-ə] suffix (41 cases); (2) Schematic strings of the form /na/ (1,052), /ra/ (1,704), and /ta/ (1,807); these represent the many cases in the Catalan lexicon in which /n/, /r/, and /t/ occur in onset position and therefore surface faithfully (the exact counts are taken from our lexical database). We included these forms to offer a more realistic challenge to the model: it must find weights that allow for deletion of /n/, /r/, and /t/ in appropriate contexts, but retain them in onset position.

Turning to the specifics of the grammar, we found we could obtain better model fit with a system that maintains a separation between morphological and phonological components.[23] We adopted the scheme used in, e.g. Jarosz (2006) and Wang and Hayes (2025). This assumes that morphemes are affiliated with sets of URs, which bear probabilities summing to one. For example, for the Catalan masculine morpheme, we assumed the three rival URs /-∅/, /-u/, and /-ə/. In a model fitted to the lexical data, these will bear probabilities of about 0.98, 0.01, and 0.01, respectively. Suffixing each of these to the stem /sant/ 'holy', assumed to have just one UR, we obtain the word-level URs /sant/, /sant-u/, and /sant-ə/, also with probabilities 0.98, 0.01, and 0.01. Lastly, a probabilistic MaxEnt phonology assigns probabilities to rival candidates for each UR; e.g. for /sant-u/ the faithful candidate [santu] would receive probability 1. The final predicted probability for any surface form (SR), derived from lexical entry L (e.g. /sant/$_{\text{M.SG}}$) is found by multiplying the probability of the various URs by the probability that the SR will be derived from that UR, then summing over all candidate URs; i.e. $P(SR|L) = \Sigma_{UR}(P(SR|UR) \times P(UR|L))$.

This scheme is implemented in grammar (17), which includes the UR probabilities and the constraint weights obtained in fitting the system both to the lexicon and to the Experiment 1 results. The constraints are drawn from the following sources: Markedness and Faithfulness from McCarthy and Prince (1995), Beckman (1998), Mascaró (2007), and Wheeler (2005); *MAP from Zuraw (2007, 2013); and UR inference constraints (last four) from Kuo (2023:§2.2.1).

(17) *UR probabilities and constraints*

a. *UR probabilities for the masculine suffix*

	Lexicon	*Experiment 1*
Null	0.981	0.790
/-u/	0.011	0.024
/-ə/	0.008	0.186

[23] The rejected alternative was to include morphological constraints like "Use [-u] for masculines" in the same constraint set as the phonological constraints.

b. *Constraints with best-fit weights*

Constraint	*Meaning*	*Lexicon Weight*	*Expt. 1 Weight*
*CODA-[n]	Avoid [n] in coda position	17.01	10.20
MAX(n)	Retain /n/	14.04	6.96
MAX(n)-MONO	Retain /n/ in monosyllables	2.84	2.25
MAX(n)-POSTATONIC	Retain /n/ in penultimately-stressed stems	6.23	4.20
[-i] for M.SG	For /-in/$_{\text{M.SG}}$, select allomorph [-i][24]	9.17	0
*CODA-[r]	Avoid [r] in coda position	19.43	18.41
MAX(r)	Retain /r/	15.70	18.18
MAX(r)-MONO	Retain /r/ in monosyllables	4.02	1.27
MAX(r)-POSTATONIC	Retain /r/ in penultimately-stressed stems	6.13	14.83
[-do] for M.SG	For /-dor/$_{\text{M.SG}}$, select allomorph [-do]	18.63	0.28
*nt]$_{\text{word}}$	Avoid word-final [nt]	50	40.61
MAX(t)	Retain /t/	16.91	31.90
MAX(CC)	Penalise deletion of 2-consonant sequences	16.52	5.50
*MAP(ʒ-ʃ)	Avoid correspondence between [ʒ] and [ʃ]	11.98	0.40
*MAP(ʒ-tʃ)	Avoid correspondence between [ʒ] and [tʃ]	0	0.12
*HIATUS	Avoid two adjacent vowels	0.33	8.96
[∅]=/n/ / V__]	Surface [XV] should have the UR /XVn/	12.46	1.86
[∅]=/r/ / o__]	Surface [Xo] should have the UR /Xor/	14.41	0.45
[∅]=/n/ / u__]	Surface [Xu] should have the UR /Xur/	12.09	0.00
[∅]=/∅/ / u__]	Surface [Xu] should have the UR /Xu/	12.09	8.55

As can be seen, the fitted model parameters for the Lexicon and Experiment 1 are often very different, reflecting divergences from frequency matching already discussed. For instance, *MAP(ʒ-ʃ) is weighted highly for the lexicon but not for Experiment 1, reflecting the tendency of participants to carry out saltation repair for /ʒ/ stems. The higher UR probabilities for masculine [-u] and (especially) [-ə] reflect the adoption of these suffixes by participants as avoidance strategies.

The full model (for tableaux see Supplementary Materials) offers a very good match to the lexical data ($r = 0.992$ for candidate probabilities across 59 candidates) and a fairly good match to the Experiment 1 choices ($r = 0.965$). We suggest that the poorer performance against the Experiment 1 data is because, as argued above, participants tend to resort to avoidance when

[24] We suggest that the extreme preference for both [-i] and [-do] reflects not general phonology, but preference for listed allomorphs.

they are uncertain. Our model has no way to incorporate this effect and can only provide an *overall* best-fit value ((17a)) for the avoidant masculine URs /-u/ and /-ə/.

4.4.2 Using the grammar for hypothesis-testing

With this grammar, we turn to the testing of the four hypotheses in (16). To this end we incorporated the grammar into a larger model (Supplementary Materials), whose tableau includes a separate input for every instance in which a participant made a non-aberrant choice in Experiment 1. There are 798 such inputs, with the winner being the participant's choice. This larger model is not intended as a grammar but rather incorporates grammar (17) as part of a model of speaker behaviour.

To test whether particular individual participants tend to be especially Faithful (16b), we incorporated into our model 37 factors, each of the form BEFAITHFUL$_{\text{Participant } i}$, where *i* ranges over the set of 37 analysed participants. These factors record whether a participant favoured an unfaithful candidate, whether this be deleted {[n], [r], [t]} or [tʃ] from /ʒ/. We anticipated that in the best-fit weights, participants might vary in their weights for BEFAITHFUL$_{\text{Participant } i}$, reflecting their degree of personal preference for faithful candidates.

To check whether individual participants tended to be especially avoidant, and whether this avoidance involves a particular preference for either of the masculine suffixes [-u] or [-ə], we added 37 factors of the form EMPLOY[-u]$_{\text{Participant } i}$ and EMPLOY[-ə]$_{\text{Participant } i}$. These were similar to BEFAITHFUL$_{\text{Participant } i}$, but are assigned to the [-u] candidate or [-ə] candidate of the tableau, respectively, for all relevant inputs from participant *i*.

We also tested a factor CONSISTENT, which recorded when a candidate for a particular input matches the response in Experiment 1 that the same participant provided for the other stimulus from the same subcondition. This was intended to provide a further test of the hypothesis (16a) that individual participants tend to have particular preferences for individual phonological patterns.

Lastly, we included a factor for RECENCY, which records when a choice was made that was consistent with the choice made for the same subcondition when it occurred among the previous five test items (we used a sliding scale, with greater values for more recent exposures). This tests hypothesis (16d), that wug-test responses can *prime* subsequent responses (Jacobs, Cho & Watson 2019; James & Burke 2000; Stemberger 2004; White, Embick & Tamminga 2024).

In performing our statistical tests, it is not valid to test each of the features or feature sets separately, since they overlap in their effects. For instance, a participant who is very Faithful would also score high on Consistency. For this reason, we used the "step-up" method for model-building described in Walker (2010:§4.3). We started with just the phonological constraints and UR probabilities (17), then elaborated the model with further features in stepwise fashion, testing each added factor for statistical significance with the Likelihood Ratio Test. At each stage, we added to the model the feature or feature set with the lowest *p*-value, stopping when no further addition would pass a significance criterion of $p = 0.01$.

In our testing, we found that a model consisting solely of the UR probabilities and phonological constraints, fitted and assessed against the full dataset, obtained a log-likelihood value of −650.8. We then added in the 37 factors of the EMPLOY[-ə]$_{\text{Participant } i}$ family. Log-likelihood increased to −539.7, significant by the Likelihood Ratio Test (χ^2 (37) = 222.3, $p = 3 \times 10^{-28}$). The test confirmed what is already evident from inspection, namely that the participants differ greatly in whether they prefer the [-ə] avoidance strategy; indeed the range across participants was from 0% [-ə] masculines to 100%.

We then added the next best available choice, CONSISTENT. This yielded a smaller increase in log-likelihood than before, i.e. to −503.9; (χ^2 (1) = 71.5, $p = 2.8 \times 10^{-17}$). This suggests that there may be small but meaningful differences in the participants' phonological grammars, reflected in a tendency toward consistent outcomes for particular subconditions.

After this, the factor EMPLOY[-u]$_{\text{Participant } i}$ was entered into the model as the best choice. Log-likelihood rose modestly to −465.1, (χ^2 (37) = 77.5, $p = 0.00011$). The impact of this factor was due to just four particular participants who used [-u] frequently; unlike the use of [-ə], it was not a popular option in general. For discussion of the participants' strong preference for [-ə] over [-u], see §4.1 above.

With these factors, no further features produced significant improvement in the model. In particular, Faithfulness raised log-likelihood to −442.9, (χ^2 (37) = 44.4, $p = 0.19$), suggesting that our experiment provides no support for the (to us, intuitive) idea that there exists relatively faithful and unfaithful speakers. Recency raised log-likelihood only to −464.2, (χ^2 (1) = 1.87, $p = 0.17$), indicating no support for any self-priming effect.

In summary, the lexical survey and Experiment 1 results are fitted fairly closely by our MaxEnt model. Using this model as the core of a statistical study, we found that participants differed in how they responded to the wug test, most notably in their willingness to exaggerate the use of unusual morphology to avoid making phonological choices. In addition, we saw a modest tendency to give matched responses to items from the same subcondition, suggesting small inter-participant differences in the internalized phonological system. There was no basis to support any sort of Faithfulness preference, nor any sort of self-priming effect.

5. Summary and conclusions

5.1 Summary of findings

To conclude, we return first to the purely phonological research questions that motivated this study. Our first question was whether participants tested on novel words *frequency-match* the patterns of the lexicon (Zuraw 2000; Ernestus & Baayen 2003); the relevant phenomena were /n/-deletion and /r/-deletion. We found that these processes are productive and do indeed involve frequency-matching: the relative deletion rates observed in production (Experiment 1) and the ratings in the acceptability judgement task (Experiment 2) matched the lexically-observed order *frequent suffix* > *other* > *monosyllabic stems* > *penultimately-stressed stems*. However, at the level of detail there were discrepancies: notably, speakers were less likely to delete /r/ compared to /n/. We conjectured that this may be due either to exposure to dialects without /r/-deletion, or to the fact that only /n/-deletion is spelt.

Our next question was whether opaque processes can be productive (Kiparsky 1973; Sanders 2003); in this case /n/-deletion counterfed by /nt/-cluster simplification. The answer seems to be affirmative, in that the opaque pattern /n/ → ∅, /nt/ → [n] was repeatedly volunteered (Experiment 1) and found acceptable (Experiment 2). The main puzzle in the data was why /nt/-cluster simplification, which is exceptionless in the lexicon, should have so often been underapplied and underrated in testing. Possible explanations are non-deleting /nt/ dialects, orthography, or the very fact that /nt/-cluster simplification is part of an opaque interaction.

Our third research question concerned the stability of saltatory alternations, here [ʒ] ~ [tʃ]. Although participants did provide saltatory responses about half of the time, [ʃ]-final forms, with "saltation repair," were also frequent, and in Experiment 2 participants rated them as somewhat acceptable. This is despite the fact that forms with repaired saltation are not attested in the lexicon nor present in other dialects. This is consistent with the hypothesis that saltation is "unnatural phonology," liable to repair (White 2014; Hayes & White 2015), though again low attestation competes as a possible explanation.

Lastly, we addressed the question of UR inference (Ernestus & Baayen 2003) by giving participants neutralized, vowel-final wug-words and asking them to provide or rate a feminine form ending in [nə], [rə], or [ə] in hiatus. We found that /n/ is overwhelmingly preferred for the UR despite its only being marginally more frequent than /r/. This may be due to the same orthographic effects observed in /n/- vs. /r/-deletion.

Beyond the purely phonological questions we had in mind in designing the test, the results proved informative in other ways. Most notably, we found that our participants were frequently "avoidant," in the sense that they used unusual morphology to dodge the question at hand regarding phonology. Participants varied considerably in the degree to which they were avoidant, and in which suffix they used for the purpose. Avoidance appears to be related to uncertainty, as shown by its correlations with weak lexical support and with long response times.

Demographic factors, such as education or sex, were largely nonsignificant predictors. But we did find small differences in what may be actual phonological knowledge: participants showed a weak tendency to be consistent in how they responded to wug items from the same subcondition. There were no demonstrable effects of being Faithful overall, nor of self-priming.

5.2 Directions for future work

5.2.1. Explaining the puzzles through cross-language study

While our experimental data to some extent conformed to what we had expected in advance, the outcomes also included some puzzles. For several of these, we have speculatively offered multiple explanations, but nothing in our data permits us to distinguish between these explanations. For instance, it seems impossible at present to know whether the large number of saltation repairs we observed resulted from a bias against saltation itself, from their sparse lexical attestation, or indeed from some factor as yet unnoticed. Ideally, we would want to look at additional cases in which the combination of available explanations is different, in order to sort out the explanations that are truly effective. If so, it is clear that we need to wug-test other phonological systems. At present, it the set of systems that have been submitted to careful phonological analysis is surely larger than the set of systems that have also been submitted to wug-testing.

5.2.2. Challenges for phonological learnability theory

We offer our Catalan study (along with our full data in the Supplementary Materials) as a possible challenge to experts in computational learning theory. We have two particular issues in mind.

First, our data bear on the question of biased learning (Wilson 2006 et seq.): we seek a model that can input our lexical data, process it using appropriate UG biases and (perhaps) task-based factors, and output something like our wug-test results. Work that has been able to do this has employed data from artificial grammar learning experiments (Wilson 2006; White 2017) and paradigms from Hungarian (Authors in progress). We have put forth several possible biases that could qualitatively explain the deviations between wug-test and lexicon but have found no way to use the Gaussian-prior-based system employed in earlier work to implement these ideas in an effective formal learning model. It is possible that further theoretical development may be needed to make this possible.

Second, we are interested in modelling speaker uncertainty. We suggested (§4.1) that uncertainty is the most likely cause of avoidant responses, and that it is more common for phenomena that are poorly attested in the lexicon. That poor attestation should create uncertainty seems completely intuitive to us; however, it is not a prediction made by current constraint-based learning models (e.g. MaxEnt, or other frameworks like Stochastic OT, Boersma 1998). Such models frequency-match even when this matching is based on very few forms. One might imagine future models that incorporate uncertainty by allocating probability mass to a "Don't Know" candidate, but research is needed to find a principled way to do this.

References

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

Becker, Michael, and Maria Gouskova (2016). Source-oriented generalizations as grammar inference in Russian vowel deletion. *LI* 47. 391-425.

Becker, Michael, Andrew Nevins & Jonathan Levine (2012). Asymmetries in generalizing alternations to and from initial syllables. *Lg* **88(2)**. 231–268.

Beckman, J. N. (1998). *Positional faithfulness*. PhD dissertation, University of Massachusetts Amherst.

Berko, Jean (1958). The child's learning of English morphology. *Word* **14(2-3)**. 150–177.

Boersma, Paul (1998). *Functional phonology: Formalizing the interactions between articulatory and perceptual drives.* The Hague: Holland Academic Graphics.

Boersma, Paul & David Weenink (2019). Praat: Doing phonetics by computer. Version 6.0.49. http://www.praat.org/.

Bonet, Eulàlia & Maria-Rosa Lloret (1998). Fonologia Catalana. *Biblioteca filològica* **39(2)**. 75–120.

Bonet, Eulàlia, Maria-Rosa Lloret & Joan Mascaró (2005). How unnatural and exceptional can languages become?. Talk presented at the *3rd International Conference in Language Variation in Europe* (ICLaVE 3). Amsterdam.

Bonet, Eulàlia & Maria-Rosa Lloret (2018). Fricative–affricate alternations in Catalan. *Probus* **30(2)**, 215-249.

Bowers, Dustin (2019) The Nishnaabemwin restructuring controversy: new empirical evidence. *Phonology* 36:187–224.

Chomsky, Noam & Morris Halle (1968). *The sound pattern of English.* New York, NY: Harper & Row.

Cotterell, Ryan, Nanyun Peng & Jason Eisner (2015). Modeling word forms using latent underlying morphs and phonology. *Transactions of the association for computational linguistics* **3**. 433–447.

Daland, Robert, Mira Oh & Syejeong Kim (2015). When in doubt, read the instructions: Orthographic effects in loanword adaptation. *Lingua* **159**, 70–92.

Dmitrieva, Olga, Allard Jongman & Joan A. Sereno (2010). Phonological neutralization by native and non-native speakers: The case of Russian final devoicing. *JPh* **38(3)**. 483–492.

Do, Youngah (2018). Paradigm uniformity bias in the learning of Korean verbal inflections. *Phonology* **35(4)**. 547–575.

Ernestus, Miriam & Harald Baayen (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language* **79(1)**. 5–38.

Finger, Heinrich, Christo Goeke, Daniela Diekamp, Konstantin Standvoß & Peter König (2017). LabVanced: A unified JavaScript framework for online studies. Paper presented at the *International conference on computational social science*. Cologne.

Goldwater, Sharon & Mark Johnson (2003). Learning OT constraint rankings using a maximum entropy model. In Goldwater, Sharon, Mark Johnson, Jennifer Spenader, Anders Eriksson & Östen Dahl (eds.) *Proceedings of the Stockholm workshop on variation within Optimality Theory*. 111–120. Stockholm: Stockholm University.

Gouskova, Maria (2025). Phonological selection in small sublexicons. In Gerard Avelino, Merlin Balihaxi, Quartz Colvin, Vincent Czarnecki, Hyunjung Joo, Chenli Wang, Utku Zorbarlar, Adam Jardine & Adam McCollum (eds.) *Proceedings of the Annual Meeting on Phonology 2023-2024*. Amherst, MA: University of Massachusetts Amherst Libraries.

Groß, Johannes (2019). Catalan2IPA: Automatic transcription of Catalan into the International Phonetic Alphabet [Software].

Halle, Morris & Jean-Roger Vergnaud (1981). Harmony processes. In Wolfgang Klein, Willem Levelt (eds.) *Crossing the boundaries in linguistics: Studies presented to Manfred Bierwisch.* 1–22. Dordrecht: Springer Netherlands.

Hartigan, John A. & Patrick M. Hartigan. (1985) The Dip Test of Unimodality. *The Annals of Statistics* **13(1)**. 70-84.

Hayes, Bruce & Zsuzsa C. Londe (2006). Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology* **23(1)**. 59–104.

Hayes, Bruce, Péter Siptár, Kie Zuraw & Zsuzsa Londe (2009). Natural and unnatural constraints in Hungarian vowel harmony. *Lg* **82**. 822–863.

Hayes, Bruce & James White (2013). Phonological naturalness and phonotactic learning. *LI* **44**. 45-75.

Hayes, Bruce & James White (2015). Saltation and the P-map. *Phonology* **32(2)**. 267–302.

Ito, Junko & Armin Mester (2003). On the sources of opacity in OT: Coda processes in German. In Caroline Féry & Ruben van de Vijver (eds.), *The Syllable in Optimality Theory*. 271–303. Cambridge: Cambridge University Press.

Jacobs, Cassandra L., Sun-Joo Cho & Duane G. Watson (2019). Self-priming in production: Evidence for a hybrid model of syntactic priming. *Cognitive Science* **43**(7). E12749.

James, Lori E. & Deborah M. Burke (2000). Phonological priming effects on word retrieval and tip-of-the-tongue experiences in young and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **26(6)**. 1378.

Jarosz, Gaja (2006). *Rich lexicons and restrictive grammars: Maximum likelihood learning in Optimality Theory*. PhD dissertation, Rutgers University.

Jo, Jinyoung (2024). *Individual differences in phonology: the realization of stem-final coronal obstruents in Korean*. PhD dissertation, UCLA.

Jovanovich-Trakál, Marina (2021). *El procés d'elisió de la /n/ final en síl·laba tònica en català: Estat de la qüestió i experiment amb pseudoparaules*. Ms, Universitat de Barcelona.

Kawahara, Shigeto (2018). Phonology and orthography: The orthographic characterization of rendaku and Lyman's Law. *Glossa: A journal of general linguistics* **3**(1). 10.

Kenstowicz, Michael & Charles Kisseberth (1979). *Generative phonology: Description and theory*. New York, NY: Academic Press.

Kerkhoff, Annemarie Odilia (2004). Acquisition of Voicing alternations. In Sergio Baauw & Jacqueline van Kampen (eds.), *Proceedings of Generative Approaches to Language Acquisition (GALA) 2003*. Vol. 2. 269-280. Utrecht: LOT.

Kerkhoff, Annemarie Odilia (2007). *Acquisition of morpho-phonology: The Dutch voicing alternation.* PhD dissertation, Utrecht University.

Kim, Joo Kyeong (2025). *Phonetic/Phonological Analysis of Allomorphy Acquisition by Heritage Speakers.* PhD dissertation, UCLA.

Kiparsky, Paul (1968). Linguistic universals and linguistic change. In Emmon Bach & Robert Harms (eds.) *Universals in Linguistic Theory*. 171–202. New York, NY: Holt, Rinehart, and Winston.

Kiparsky, Paul (1973). Abstractness, opacity, and global rules. In Osamu Fujimura (ed.), *Three dimensions of linguistic theory*. 57–86. Tokyo: TEC.

Kuo, Jennifer (2023). Phonological markedness effects in reanalysis. PhD dissertation, UCLA.

Łubowicz, Anna (2002). Derived environment effects in Optimality Theory. *Lingua* **112**(4). 243–280.

Mascaró, Joan (1975). *Catalan phonology and the phonological cycle.* PhD dissertation, MIT.

Mascaró, Joan. (2007). External allomorphy and lexical representation. *LI* **38**(4), 715–735.

Mayer, Connor (in press). A large-scale corpus study of phonological opacity in Uyghur. To appear in *Phonology*.

McCarthy, John (1988). Feature geometry and dependency: A review. *Phonetica* **43**. 84–108.

McCarthy, John (2005). Taking a free ride in morphophonemic learning. *Catalan Journal of Linguistics* **4**:19–56.

McCarthy, John & Alan Prince (1995). Faithfulness and reduplicative identity. In Jill Beckman, Suzanne Urbanczyk & Laura W. Dickey (eds.) *University of Massachusetts occasional papers in linguistics 18: Papers in Optimality Theory.* 249–384.

Moll, Francesc de B. (1952). *Gramática histórica catalana*. Valencia: Gredos.

Moreton, Elliott & Joe Pater (2012). Structure and substance in artificial-phonology learning, part I: Structure. *Language and linguistics compass* **6**(11), 686–701.

Moreton, Elliott & Katya Pertsova (2023). Implicit and explicit processes in phonological concept learning. *Phonology* **40(1-2)**, 101–153.

Odden, David (2005) *Introducing phonology*. Cambridge: Cambridge University Press.

O'Hara, Charlie. (2020). Frequency matching behavior in on-line maxent learners. In Allyson Ettinger, Gaja Jarosz, Joe Pater (eds.) *Proceedings of the Society for Computation in Linguistics (SCiL)* 3(1). New York, NY: Association for Computational Linguistics.

Pérez-Pereira, Miguel (1989). The acquisition of morphemes: Some evidence from Spanish. *Journal of Psycholinguistic Research* **18**. 289–312.

Pons-Moll, Clàudia (2015). Comentaris a "Regularitat i excepcions en fonologia: Les reduccions vocàliques," de Joan Mascaró. In Maria-Rosa Lloret, Clàudia Pons-Moll & Eva Bosch-Roura (eds.) *Clàssics d'ahir i d'avui en la gramàtica del català* **15.** 71–104. Barcelona: Edicions Universitat.

Pons-Moll, Clàudia (2021). Estratègies en l'adaptació de manlleus en català i en altres llengües romàniques. In Maria-Rosa Lloret & Clàudia Pons-Moll (eds.) *L'Adaptació de manlleus en català i en altres llengües romàniques*. 111–144. Barcelona: Edicions Universitat.

Prince, Alan & Paul Smolensky (1993). *Optimality Theory: Constraint interaction in generative grammar*. Rutgers Center for Cognitive Science. New Brunswick: Rutgers University. Published 2004, Blackwell

Rasin, Ezer, Inbar Berger, Nathan Lan, Iris Shefi & Roni Katzir (2021). Approaching explanatory adequacy in phonology using Minimum Description Length. *Journal of Language Modelling* **9**(1). 17–66.

Sanders, Nathan (2003). *Opacity and sound change in the Polish lexicon*. PhD dissertation, University of California, Santa Cruz.

Shilen, Alexander & Colin Wilson (2022). Learning Input Strictly Local Functions: Comparing Approaches with Catalan Adjectives. In Allyson Ettinger, Tim Hunter, Brandon Prickett (eds.) *Proceedings of the Society for Computation in Linguistics (SCiL)* **5(27)**. 244–246. New York, NY: Association for Computational Linguistics.

Smolek, Amy & Vsevolod Kapatsinski (2018). What happens to large changes? Saltation produces well-liked outputs that are hard to generate. *Laboratory Phonology* **9(1)**. 10.

Song, Hanbyul & James White. (2022). Interaction of phonological biases and frequency in learning a probabilistic language pattern. *Cognition* **226**. 105170.

Stave, Matthew, Anna Smolek & Vsevolod Kapatsinski (2013). Inductive bias against stem changes as perseveration: Experimental evidence for an articulatory approach to output-output faithfulness. In *Proceedings of the Annual Meeting of the Cognitive Science Society* 35.

Stemberger, Joseph Paul (2004). Phonological priming and irregular past. *Journal of Memory and Language* **50**(1). 82–95.

Steriade, Donca (2009). The phonology of perceptibility effects: The P-map and its consequences for constraint organization. In Kristin Hanson & Sharon Inkelas (eds.) *The Nature of the Word: Studies in Honor of Paul Kiparsky*. 151–180. Cambridge: MIT Press.

Torres-Tamarit, Francesc (2016). *Algunes reflexions sobre la fonologia de les consonants postalveolars del català*. Paper presented at the Universitat Autònoma de Barcelona.

Walker, James A. (2010). *Variation in linguistic systems*. New York: Routledge.

Wang, Yang & Bruce Hayes (2025). Learning phonological underlying representations: the role of abstractness. *LI.* 1–44.

Wheeler, Max W. (2005). *The phonology of Catalan*. Oxford University Press.

White, James (2014). Evidence for a learning bias against saltatory phonological alternations. *Cognition* **130**(1). 96–115.

White, James (2017). Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a P-map bias. *Lg **93***(1). 1–36.

White, Yosiane, David Embick & Meredith Tamminga (2024). Affix priming with variable ING in English: Implications for unique vs. dual representation. *Journal of Memory and Language* **138**. 104535.

Wilson, Colin (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* **30**(5). 945–982.

Ylonen, Tatu (2022). Wiktextract: Wiktionary as Machine-Readable Structured data. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache et al. (eds.) *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC).* 1317-1325. Marseille.

Zamuner, Tania S., Annemarie Kerkhoff & Paula Fikkert (2012). Phonotactics and morphophonology in early child language: Evidence from Dutch. *Applied Psycholinguistics* 33(3). 481–499.

Zhang, Jie, Yuwen Lai, and Craig Sailor. (2011). Modeling Taiwanese speakers' knowledge of tone sandhi in reduplication. *Lingua* **121**. 181-206.

Zuraw, Kie (2000). *Patterned exceptions in phonology*. PhD dissertation, UCLA.

Zuraw, Kie (2007). The role of phonetic knowledge in phonological patterning: Corpus and survey evidence from Tagalog. *Lg* **83**. 277-316.

Zuraw, Kie (2010). A model of lexical variation and the grammar with application to Tagalog nasal substitution. *NLLT* **28**. 417–472.

Zuraw, Kie (2013). *MAP constraints. Ms, UCLA. https://linguistics.ucla.edu/people/zuraw/dnldpprs/star˙map.pdf.