Losing phonotactic distinctions in context

Abstract

Previous psycholinguistic research has demonstrated that sentence processing varies according to both syntactic and discourse context. However, systematic investigation of how such contexts influence how the processor manages low-level representations of linguistic structure has yet to be carried out. In this paper, we conduct a series of self-paced reading experiments which show how one well-established linguistic measurement – phonotactic distinctions between non-words – varies according to the phonological, syntactic, and discourse context that the non-words appear in. Our results demonstrate that the various types of context that we control for can influence both when and if phonotactic distinctions surface. More broadly, our findings suggest that well-established phonological and psycholinguistic effects may not generalize when tested in larger contexts.

1 Introduction

In this paper, we investigate how people incrementally regulate and prioritize different aspects of the linguistic signal. More specifically, we explore how context can influence on-line processing of low-level phonological detail during silent reading.[1]

For an example of how context can modulate sentence processing, consider the sentences in (1) and (2):

(1) Yesterday morning [Mikayla *told* the story about the dog].

(2) Srujan *heard* [Mikayla <u>told</u> the story about the dog].

Despite having similar semantic content, the two differ in their syntactic organization: while *Mikayla told the story about the dog* is shared between the two sentences, the bracketed clause is a matrix clause in (1) and an embedded clause in (2). Previous work on sentence processing has found that main clauses are processed more easily than embedded clauses in a variety of psycholinguistic paradigms (Jarvella and Herman 1972; Ko 1998; Lord 2002). Additionally, differences in processing difficulty between types of embedded clauses – such as subject-extracted relative clauses and object-extracted relative clauses – have also been observed cross-linguistically (Bader and Meng 1999; Hsiao and Gibson 2003; Gibson, Desmet, et al. 2005; Ishizuka 2005; Gibson and Wu 2013). We will refer to contexts like those in (1) and (2) as *syntactic contexts*.

Like the effects of syntactic structure that we observed in (1) and (2), previous work has also demonstrated that broader discourse can also manipulate sentence processing. For example, consider (3):

(3) When the boys strike the dog kills.

Prior psycholinguistic research has found that garden-path sentences like (3), when read in isolation, display significant processing slowdowns at *kills*, the disambiguating region (Frazier 1979; Ferreira and Henderson 1991).[2] However, the processing difficulty of garden-path sentences can be reduced if the garden-path sentence follows some relevant discourse (Crain and Steedman 1985; Warner and Glass 1987), as in (4):

(4) CONTEXT: The dogs become dangerous whenever boys attack.
When the boys strike the dog kills.

The discourse present in (4) biases the processor towards an intransitive reading of *When the boys strike*, meaning that people are less likely to follow the erroneous parse for the transitive reading (*When the boys strike the dog*), thus lessening the processing difficulty of the garden-path sentence overall. We will refer to contexts like that of (4) as *discourse contexts*. In total, the examples in (1)-(4) suggest that both local syntactic context and global discourse context can modulate how the processor handles linguistic input.

[1]This paper is concerned with how the processor handles written input; we acknowledge that our findings may vary from studies where the processor receives other kinds of input, such as sign or sound.

[2]For an example as to how the processor handles written input differently from spoken input, note that example (3) is unambiguous when read with the correct prosody.

In addition to processing the syntactic and semantic aspects of the linguistic signal in (1)-(4), the processor is also sensitive to phonological aspects of the linguistic signal during reading, with prior work showing on-line processing effects of stress (McCurdy, Kentner, and Vasishth 2013), metrical structure (Magne, Gordon, and Midha 2010; Breen and Clifton Jr 2011), rhyme (Acheson and MacDonald 2011), and binomial ordering preferences (Siyanova-Chanturia, Conklin, and Van Heuven 2011; Morgan and Levy 2016). Moreover, the phonological structures that are generated during reading are susceptible to the same phonological judgments for the same phenomena outside of reading contexts: stress must be placed in the correct position (McCurdy, Kentner, and Vasishth 2013), metrical structure should remain unviolated (Magne, Gordon, and Midha 2010; Breen and Clifton Jr 2011), rhymes take longer to process (Acheson and MacDonald 2011), and binomial preferences appear to follow previously-established phonological constraints (Siyanova-Chanturia, Conklin, and Van Heuven 2011; Morgan and Levy 2016). Given these results, it is likely that some phonological structure is projected or accessed during reading, though most of the aforementioned studies examine phonological phenomena of the *phrasal* or *prosodic* kind. In this paper, we study another well-established phonological phenomenon that is at the *segmental* level, a level which minimizes references to higher levels of linguistic organization: phonotactics.

Phonotactic structure corresponds to how sounds can pattern in a language. Phonotactic distinctions are differences in how viable a particular sound pattern is within a language: most English speakers distinguish *blick* as more acceptable than *bnick* (Chomsky and Halle 1968). Importantly, all languages display some language-specific constraints on what patterns of phonemes are viable. For example, word-initial /mb/ is phonotactically unviable in English, but is phonotactically viable in Mbay (Keegan 1997).

Language-specific differences between viable and unviable phonotactic structures have been investigated both theoretically and experimentally over the past few decades (Chomsky and Halle 1968; Vitevitch and Luce 1998; Vitevitch, Luce, Pisoni, et al. 1999; Frisch et al. 2001; Kirby and Yu 2007; Hayes and Wilson 2008; Albright 2009; Mollin 2012; Hayes and White 2013, *inter alia*). Most previous work on phonotactics often asks people to rate how "viable" or "well-formed" a non-word target may be to a native speaker, either in isolation or in comparison to another non-word target, usually in some form of spoken word recognition or comprehension task (Vitevitch, Luce, Charles-Luce, et al. 1997; Shademan 2006; Breiss 2020, *inter alia*); these paradigms will be discussed further in Section 2. To reduce the possibility of noise from other aspects of processing beyond those that are phonological, the majority of these studies isolate the non-word target from any context. However, as has been established, language processing requires the individual to manage multiple aspects of linguistic information simultaneously. As such, it is unclear whether previously-found distinctions in phonotactic acceptabilty persist when such targets are processed within different contexts, and if such distinctions also arise during reading. We contribute to the literature by investigating how these differences arise when phonological, syntactic, and discourse contexts are controlled and manipulated.

Broadly, we investigate how layers of context across multiple linguistic subfields can influence the processor's behavior. More specifically, we conduct four self-paced reading experiments that show how the processor's computation of phonotactic acceptability can vary according to three dimensions of context – syntax, discourse, and phonology

– during sentence processing, and how these results both conflict with how phonotactic distinctions[3] have surfaced in prior research and inform our understanding of sentence processing across different aspects of the linguistic signal.

The structure of this paper is as follows. In the following section, we detail more information surrounding phonotactics and the role of context during sentence processing. In Sections 3-6, we describe four self-paced reading experiments that place non-words of varying phonotactic acceptability in different contexts. In Section 3, we present Experiment 1, which investigates how phonotactic differences for TARGETS with varying onset phonotactics surface in different syntactic contexts. In Section 4, we present Experiment 2, which explores how the presence of a discourse context further modulates how phonotactic differences surface during reading. In Section 5, we present Experiment 3, which places the discourse context of Experiment 2 after the critical sentence; Experiment 3 replicates Experiment 1 in the two-sentence self-paced reading paradigm of Experiment 2, supporting the finding that the presence of discourse context in Experiment 2 greatly modulates how phonotactic distinctions appear. All non-word targets in Sections 3-5 have modifications to the word-onset position. In Section 6, we present Experiment 4, a follow-up to Experiment 1, which examines the influence of phonological context on phonotactic judgments by applying phonotactic modifications to the coda position instead of the onset position. In Section 7, we discuss our findings and their broader implications. In Section 8, we conclude.

2 Background

2.1 Phonotactic distinctions

As mentioned previously, much prior psycholinguistic research has investigated phonotactic distinctions between non-word targets using experimental paradigms where the non-words are removed from any linguistic context. Examples of such paradigms include rating tasks (Vitevitch, Luce, Charles-Luce, et al. 1997; Dankoviþová et al. 1998; Shademan 2006; Weber and Cutler 2006; Scholes 2016), speeded auditory-recognition tasks (Vitevitch, Luce, Charles-Luce, et al. 1997; Vitevitch and Luce 1998), or artificial grammar learning tasks (Adriaans and Kager 2017; Linzen and Gallagher 2017; Breiss 2020), among others.[4] These studies have found that phonotactic acceptability is gradient: $blick > b\overset{?}{w}ick > b\overset{*}{n}ick$ (Bailey and Hahn 2001; Shademan 2007; Albright 2009), though phonotactic acceptability can also display more categorical distinctions, as in the strong unacceptability of word-initial /mb/ in English.[5]

We argue that the investigation of phonotactic differences *in context* is warranted, for a number of reasons. First, even though differences in phonotactic acceptability are well-accepted, little (or no) prior work has examined whether differences in phonotactic acceptability generalize to more naturalistic linguistic contexts, where non-word

[3] When describing phonotactic structure, we use *distinctions* and *differences* interchangeably throughout this paper.

[4] Prior work suggests that other factors may interact with phonotactic distinctions, such as orthotactic effects and neighborhood density (e.g., Vitevitch and Luce 1998). While these factors are not the primary concern of this paper (and therefore not included in our main statistical analyses), post-hoc analyses that incorporate these factors can be found in Appendix B.

[5] We do not explore gradient phonotactic acceptability differences in this paper; we focus only on clearly viable (*blick*) and unviable (*bnick*) non-words. We limit our analyses to the edges of the acceptability spectrum in order to more clearly define the bounds of phonotactic distinction during reading.

targets surface in sentences. Second, while it is evident that some abstract phonological computation, most notably prosody, is used during sentence processing (Fodor 2002; Snedeker and Trueswell 2003; McCurdy, Kentner, and Vasishth 2013, *inter alia*), whether these computations occur at the sub-lexical level during sentence processing is unknown. Third, given the frequent link between prosodic structure and syntactic structure, studying phonotactic distinctions in context provides opportunity to reveal how syntactic and discourse information modulates the processing of non-syntactic, non-discourse phenomena. Some prior work has shown that lexical, syntactic, and discourse information interact during processing (e.g. Marslen-Wilson and Tyler 1980; Britt et al. 1992), but most studies in this domain focus on spoken word recognition and ambiguity resolution; our analyses extend this literature to phonological structure during reading. Given these motivations, we will now discuss how we studied phonotactic structure in context.

2.2 Integrating phonotactic distinctions into context

Consider again the examples in (1)-(4), here repeated as (5)-(6):

(5) Yesterday morning Mikayla *told* the story about the dog.
Srujan *heard* Mikayla told the story about the dog.

(6) The dog becomes loud whenever the boys protest.
When the boys strike the dog barks.

In (5), the syntactic context of a phrase modulates how difficult the sentence is to parse; in (6), we observe that embedding a sentence in a discourse context modulates the difficulty of that sentence. In tandem, these examples demonstrate that syntactic and discourse context can manipulate the processor's behavior.

We extend these generalizations regarding syntactic and discourse context to phonotactic acceptability by placing non-word targets in different layers of syntactic and discourse embedding. For explanatory purposes, consider the sentences in (7) and (8), where sentences differ by the number of layers of syntactic embedding and by the phonotactic acceptability of the non-word target in subject position:

(7) –Non-embedded Structure–
a. Phonotactically-viable non-word
After lunch the trar stopped for gas.
b. Phonotactically-unviable non-word
After lunch the tnar stopped for gas.

(8) –Embedded Structure–
a. Phonotactically-viable non-word
They doubted the trar stopped for gas.
b. Phonotactically-unviable non-word
They doubted the tnar stopped for gas.

Above, we place {viable, unviable} non-word targets in {matrix, embedded} clauses. If phonotactic distinctions are robust to syntactic variation, then we would expect each (a) sentence to be read faster than each (b) sentence in (7) and (8).

Furthermore, we can add another layer of embedding by placing a discourse context before the sentences above:

(9) –DISCOURSE CONTEXT–
There was a delay in the trip.

Unlike prior work that explores the influence of discourse context on sentence processing, we do not investigate how a discourse context can bias the processor towards a certain interpretation, instead choosing to study how the *presence* of context alone influences the processor. We argue that this methodological distinction from prior work is valid, given that it would be quite challenging to use a discourse context to bias a participant in favor of one phonotactic target or another. However, our research still contributes to the broader psycholinguistic literature that explores how the processor manages incremental linguistic input in the *presence* of any discourse context.

In summary, sentences like those in (7)-(9) help reveal how layers of syntactic and discourse embedding can be used to study phonotactic distinctions in context: if the findings of prior work on phonotactic distinctions are robust, then we would expect differences between non-word targets of varying phonotactic acceptability to arise *consistently*. In the following four sections, we present a series of experiments that study sentences similar to (7) and (8), sometimes placing them after discourse contexts like (9), in order to more systematically probe if and how people construct phonotactic distinctions in context during reading.

3 Experiment 1

In this experiment, we investigate how phonotactic distinctions bear out for non-word targets that are placed within one-sentence contexts of differing syntactic structures, leaving discourse aside for now.

3.1 Methods

3.1.1 Design & Experimental Stimuli

Participants read one of three TARGETS in one of three STRUCTURES. Phonological TARGETS differed in onset phonotactic acceptability: VIABLE non-words satisfied the phonotactic restrictions of English syllable structure, UNVIABLE non-words violated such restrictions, and REAL words were used as a control condition. Each STRUCTURE increased the complexity which the TARGET appeared in, as reported in previous research (Kluender and Kutas 1993; Gibson, Desmet, et al. 2005): MATRIX structures placed the TARGET as the subject of the main clause, EMBEDDED structures placed the TARGET as the subject of a full-CP embedded clause, and CENTER-EMBEDDED structures placed the TARGET as the subject of a reduced relative clause. STRUCTURES and TARGETS were fully crossed in a 3x3 within-participant design, creating nine conditions for each experimental item; see Figure 1 for three examples, with each row demonstrating one of each possible STRUCTURES and TARGETS in combination. To ensure that participants saw each of our conditions the same number of times, we constructed 27 experimental items, meaning each participant saw an item in each condition three times.

	1	2	3	4	5	6
MATRIX	Later	on	the	**trip**	was	worth...
EMBEDDED	She	decided	the	**glip**	was	worth...
C-EMBEDDED	The	price	the	**lgip**	was	worth...

Figure 1: Sample experimental item for Experiment 1. Color indicates phonological TARGETS: green indicates REAL WORD targets (control). Blue indicates phonologically VIABLE targets. Red indicates phonologically UNVIABLE targets. Each phonological TARGET could surface in position 4 for each STRUCTURE; all nine conditions are not shown.

As mentioned previously, all phonotactic differences between VIABLE and UNVIABLE targets occurred in onset position; targets were constructed such that rimes and codas were identical across all TARGETS. UNVIABLE targets were constructed with either metathesis of the first two consonants of the VIABLE equivalent (*blick* > *lbick*), or by a substitution of the second consonant of the VIABLE equivalent that led to a phonotactic violation (*blick* > *bnick*). Positional information was controlled across experimental items to allow for comparisons both within and across syntactic STRUCTURES: all TARGETS appeared in position 4 of the sentence. Words in positions 1 & 2 were identical within each syntactic STRUCTURE for each experimental item; words 3-6 were identical across all syntactic STRUCTURES.

In addition to the 27 experimental items, we also constructed 27 filler items of varying syntactic structures that did not match with any of the STRUCTURE conditions. Two sample filler items for this experiment were: *No one was able to forget the legacy of Mr. Smith.* and *We are worried that our bosses will fire us.* As seen in these examples, filler items did not have any phonological modifications.

3.1.2 Procedure

All participants ($N = 62$) were native speakers of English that were recruited on Prolific. Participants were paid approximately $15/hr for their participation (mean completion time: 10 minutes). All experiments were conducted via the online research platform PC Ibex (Zehr and Schwarz 2018).

The experimental procedure follows previous work using the moving-window self-paced reading paradigm (Just, Carpenter, and Woolley 1982; Ferreira and Henderson 1990). At the beginning of each trial, the participant saw a series of dashes, where each dash corresponded to a word in the upcoming sentence. When the participant pressed the SPACE bar, the first word appeared. Participants then pressed the SPACE bar to advance to the next word; the previous word disappeared after each press. Participants repeated this procedure until they had read the full sentence.

A yes-no comprehension question regarding the sentence would appear following one-third of all trials (both experimental and filler); comprehension questions following experimental trials would never refer to the TARGET, but to material that followed the target. To reduce the likelihood that participants would rapidly press their SPACE bar to get through the sentence and then randomly choose an answer to a comprehension question (therefore causing many trials to be excluded), participants were told that payment was contingent on high accuracy (>85%) across all comprehension

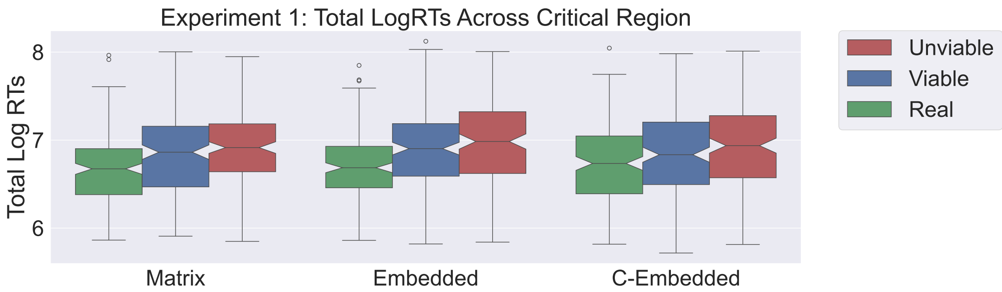

Figure 2: Total Log RTs of the critical region (log(position 4 + position 5)) across STRUCTURES for Experiment 1. Notches indicate 95% confidence intervals.

questions;[6] participants who did not achieve 85% accuracy across all comprehension questions were excluded.

3.2 Results

We collect reading times (RTs) at each position for all experimental items. We assume that phonotactic distinctions bear out during reading as follows: sentences that contain non-words with viable phonotactics will be read faster than the same sentences that contain non-words with unviable phonotactics, where greater differences in magnitude suggest a larger gap in phonotactic acceptability. Such differences in RTs may surface on the non-word target itself, the word following, or both, as prior research has shown that sentence processing during reading is not always immediate (Rayner, Garrod, and Perfetti 1992; McElree and Griffith 1995; Plummer and Rayner 2012); no significant differences in RTs should occur prior to the non-word. If we find no differences in RTs between viable and unviable targets at a critical position in a sentence, then we assume that participants were not sensitive to phonotactic distinctions when processing that word during the self-paced reading task.

All RTs less than 100ms and greater than 2000ms were excluded. Additionally, data from two participants who scored less than 85% on comprehension questions were excluded. All RTs were log-transformed. In all instances where RT differences surfaced between VIABLE and UNVIABLE targets, the VIABLE target was read more quickly than the UNVIABLE target.

We focus on RTs for position 4 (where the TARGET surfaces) and position 5 (the word following the TARGET); as mentioned previously, we study position 5 because prior research has found that sentence processing may spill over into the following word during reading (Rayner, Garrod, and Perfetti 1992; McElree and Griffith 1995; Plummer and Rayner 2012). We will refer to these two positions in tandem as the *critical region*. We conduct statistical analyses across the entire critical region and within each position of the critical region.

The results for the full critical region are summarized in Figure 2. To test for differences in RTs for the entire critical region, we fit a linear mixed-effects regression model to the log-transformed summed RT across the critical region (position 4 + position 5), with fixed effects for STRUCTURE, TARGET, and their interaction, along with by-item

[6]This additional condition on compensation was included after a previous iteration of this experiment required nearly 50% of trials to be excluded.

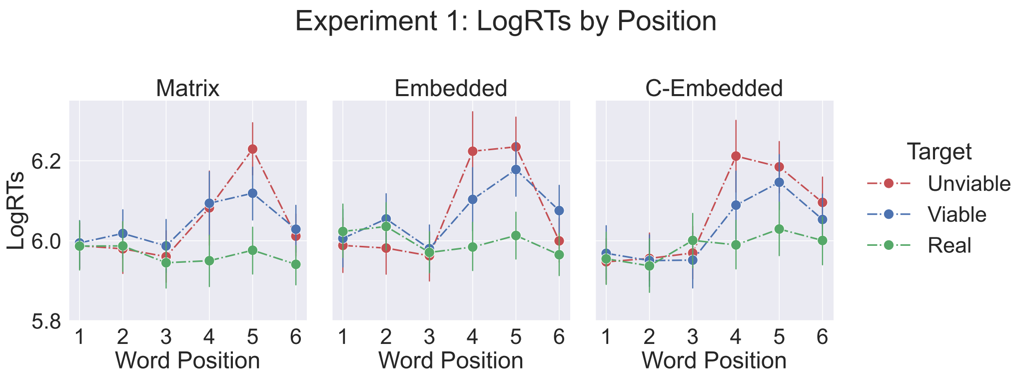

Figure 3: LogRTs for each condition by word position. Each subplot indicates STRUCTURE condition; each color represents TARGET condition.

and by-subject intercepts.[7]. In this analysis and in all following analyses, p-values were estimated using the `lmerTest` package (Kuznetsova, Brockhoff, and Christensen 2017). Across the entire critical region, we find that the REAL targets were read faster than the VIABLE targets (β = -0.163, SE = 0.036, t=-4.466, p <0.001), though we do not find any significant differences between VIABLE and UNVIABLE targets.[8] This finding indicates that neither phonological acceptability nor syntactic complexity affects the total processing time across the whole critical region.

However, phonotactic distinctions surface between VIABLE and UNVIABLE targets when looking at each position within the critical region. Positional distinctions are visualized in Figure 3. To test if there is an interaction between STRUCTURE type, phonological TARGETS, and position, we fit a linear mixed-effects regression model to the log RTs, with fixed effects for STRUCTURE, TARGET, position (4, 5), and their interactions (all possible permutations of two-way interactions, as well as the full 3x3x2 interaction), and random intercepts for both participants and items.[9]

Model outputs are reported in Table 1. We find a significant main effect for REAL targets (β = -0.138, SE = 0.043, t=-3.171, p <0.01). We report a significant two-way interaction between UNVIABLE targets and both the EMBEDDED condition (β = 0.143, SE = 0.062, t=2.291, p <0.05) and the C-EMBEDDED condition (β = 0.130, SE = 0.062, t=2.097, p <0.05). Finally, we find two significant three-way interactions, one between UNVIABLE targets, EMBEDDED clauses, and position 5 (β = -0.179, SE = 0.081, t=-2.200, p <0.05), and another between UNVIABLE targets, C-EMBEDDED clauses, and position 5 (β = -0.194, SE = 0.081, t=-2.392, p <0.05).

In sum, we find that all interactions involving non-word TARGETS and STRUCTURES are significant: phonotactic distinctions between non-words arise in different positions depending on the syntactic STRUCTURE that the TARGET appears in, with embedded clauses (EMBEDDED, CENTER-EMBEDDED) showing distinctions on the target position,

[7]The complete formula was: LogSummedRTs ~ TARGET*STRUCTURE + (1| subject) + (1| item). This model is the maximal model that reaches convergence. The TARGET baseline was the VIABLE condition. The STRUCTURE baseline was the non-embedded/MATRIX condition.

[8]For readability, full output for all of our statistical analyses on the entire critical region – for this experiment and the remaining three experiments – can be found in Appendix A.

[9]The complete formula was: LogRTs ~ TARGET*STRUCTURE*Position + (1| subject) + (1| item). This model is the maximal model that reaches convergence. The TARGET baseline was the VIABLE condition. The STRUCTURE baseline was the MATRIX condition. The position baseline was position 4.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.102	0.053	115.585	<2e-16
Unviable	-0.003	0.044	-0.077	0.939
Real	-0.138	0.043	-3.171	**0.002**
Embedded	0.005	0.044	0.113	0.91
C-Embedded	-0.004	0.044	-0.103	0.918
Position 5	0.02	0.04	0.498	0.619
Unviable:Embedded	0.143	0.062	2.291	**0.022**
Real:Embedded	0.014	0.062	0.234	0.815
Unviable:C-Embedded	0.130	0.062	2.097	**0.036**
Real:C-Embedded	0.033	0.062	0.542	0.588
Unviable:Position 5	0.104	0.057	1.815	0.070
Real:Position 5	-0.004	0.057	-0.070	0.944
Embedded:Position 5	0.052	0.057	0.906	0.365
C-Embedded:Position 5	0.025	0.057	0.438	0.661
Unviable:Embedded: Position 5	-0.179	0.081	-2.200	**0.028**
Real:Embedded: Position 5	-0.041	0.080	-0.505	0.614
Unviable:C-Embedded: Position 5	-0.194	0.081	-2.392	**0.017**
Real:C-Embedded: Position 5	-0.006	0.080	-0.079	0.937

Table 1: Model outputs for positional analysis for Experiment 1. Significant effects and interactions are bolded.

and non-embedded clauses (MATRIX) showing distinctions on the following word.

3.3 Discussion

The results of Experiment 1 demonstrate that phonotactic distinctions surface immediately for embedded conditions (EMBEDDED, CENTER-EMBEDDED), while they surface later for the non-embedded condition (MATRIX). We find that differences in timing for phonological distinctions are driven by the type of STRUCTURE the TARGET appears in: all interactions involving STRUCTURE and non-word TARGETS are significant.

Additionally, the summed syntactic and phonological processing costs incurred by the non-word TARGETS and STRUCTURES across the critical region appear to be uniform. We find no evidence for a cumulative interaction of difficulty: more difficult STRUCTURES with TARGETS of low phonological acceptability are not more challenging to process than more difficult STRUCTURES with TARGETS of high phonological acceptability, nor do we find faster processing in circumstances where phonological acceptability is high and syntactic complexity low. As such, it appears that there is a ceiling for processing costs, where the processor has computed enough of the possible signals to move on to the next word. These findings support a model of sentence processing where syntactic contexts modulate *when* phonotactic distinctions will appear: in syntactic contexts with no embedding (MATRIX), syntactic processing occurs immediately when the TARGET appears, with phonotactic differences surfacing on the following word. In syntactic contexts where there is one level of embedding, syntactic processing is delayed, thus allowing phonotactic distinctions to arise immediately on the target.

In total, we observe that previously-reported phonotactic distinctions for non-words continue to surface during on-line processing, though the timing of such differences varies according to the syntactic context that the non-word appears in.

4 Experiment 2

In Experiment 1, we explored how phonotactic distinctions of word-onsets arise in *one-sentence contexts*, finding that introducing one level of embedding shifts the timing of phonotactic differences earlier. In this experiment, we explore the influences of discourse context on phonotactic distinctions by embedding the one-sentence experimental stimuli from Experiment 1 after a one-sentence discourse context. As such, this experiment examines both how discourse context alone manipulates people's sensitivity to phonotactic information, while also investigating how layered embeddings – syntactic embedding in addition to discourse embedding – affect such low-level differences. If phonotactic distinctions are robust, we expect to see such distinctions arise, regardless of discourse context.

4.1 Methods

4.1.1 Design & Experimental Stimuli

A sample experimental item is presented in Figure 4. This experiment uses a 2 (CONTEXT) x 2 (TARGET) x 2 (STRUCTURE) design: a one-sentence CONTEXT {meaningful, random} preceded an experimental sentence from Experiment 1 that had an orthographically-transparent TARGET {viable, unviable} as the subject of a syntactic STRUCTURE {matrix subject, embedded subject}. We excluded REAL targets, as they behaved consistently in Experiment 1, and the focus of this study is on phonotactic differences between non-words; we excluded the CENTER-EMBEDDED structure, as the phonotactic distinctions appeared to follow the same pattern as those found in the EMBEDDED structure in Experiment 1.[10]

We used 24 of 27 experimental sentences (Figure 4b) from Experiment 1; 3 experimental sentences were randomly excluded to maintain a balanced number of participant exposures to conditions (2x2x2; 8 total conditions) per participant. TARGETS and STRUCTURES were identical to those from Experiment 1 besides the exclusion of REAL targets and CENTER-EMBEDDED structures. Additionally, each experimental sentence was preceded by a context sentence, thus embedding the experimental sentence in a discourse context. Given that this is an exploratory study, we were unsure how the type of context would modulate the results. As such, we constructed two kinds of contexts to survey for different possible discourse effects: MEANINGFUL or RANDOM. MEANINGFUL contexts anticipated the events (induced by the verb) that would be presented in the second sentence; RANDOM contexts were completely unrelated to the second sentence. For example, in Figure 4, the MEANINGFUL context in (A) is related to the event of the sentences in (B), while the RANDOM context in (A) is unrelated to the event of the sentences in (B). Differences between contexts are considered in our

[10]Our decision to exclude these structures in this experiment will also later be supported by how REAL targets and CENTER-EMBEDDED structures pattern identically in Experiment 4 as they do in Experiment 1.

[A] Discourse context:

	Context Type
Meaningful	There was a disagreement in the courtroom.
Random	There was a shimmer on the water.

[B] Experimental stimuli:

	Matrix Subject	Embedded Subject
Viable	By evening the *spudge* made a…	You realized the *spudge* made a…
Unviable	By evening the *psudge* made a…	You realized the *psudge* made a…
	1 2 3 4 5 6	1 2 3 4 5 6

Figure 4: Sample experimental item for Experiment 2 (A→B) and Experiment 3 (B→A). Colors for TARGETS are identical to those of Experiment 1, with the exclusion of REAL WORD targets: blue indicates phonologically VIABLE targets, red indicates phonologically UNVIABLE targets. Participants saw one of two possible context sentences (MEANINGFUL, RANDOM, then one of four experimental sentences from Experiment 1. The critical region is highlighted in grey.

statistical analyses. All contexts were seven words in length and followed the structure "There was a/an NP ADJUNCT", where NP is a noun phrase and ADJUNCT is an adjunct clause.[11]

To match our 24 experimental items, we randomly selected 24 of the 27 filler items from Experiment 1 and added a preceding context sentence as well. To minimize the risk of participants recognizing that non-words only appear in the second sentence of the experimental items, 12 of the filler items included nonsense character clusters in the initial sentence; importantly, unlike our non-word targets, these clusters were not pronounceable and did not have any relation to any real words. Two sample filler items were *It had been a fantastic season for the soccer team. The coach led the team onto the field.* and *Flashing fspcyc lights illuminated the dark night as the investigators arrived. The detective analyzed the brutal crime scene.*

4.1.2 Procedure

All participants ($N = 65$) were native speakers of English that were recruited on Prolific. Participants were paid at a rate equivalent to $15/hr for their participation (mean completion time: 11 minutes); compensation conditions were identical to those of the prior experiments presented in this paper, with participants needing to achieve an accuracy above 85% across all comprehension questions. All experiments were conducted via the online research platform PC Ibex (Zehr and Schwarz 2018).

As in the previous experiment, participants followed the standard moving-window self-paced reading procedure (Just, Carpenter, and Woolley 1982; Ferreira and Henderson 1990). However, participants instead read two sentences instead of one; the two sentences within each trial were line-separated.

4.2 Results

Results for RTs across the full critical region are visualized in Figure 5. To test for differences in total reading time between conditions, we fit a linear mixed-effects model that predicts the log-transformed sum of RTs across the critical region, with

[11] Note that the experimental stimuli from Experiment 1 were designed to not reveal significant semantic information until the verb in position 5. As such, the MEANINGFUL context condition, if significant, should modulate RTs at the end of the critical region; RTs of the non-word target should not be affected, given that the role of the target in the sentence is unclear until the verb.

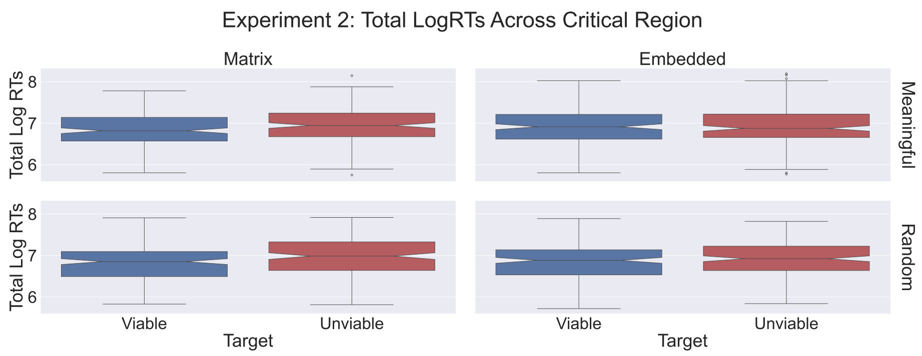

Figure 5: Total Log RTs of the critical region (log(position 4 + position 5)) across STRUCTURES and CONTEXTS for Experiment 2. Notches indicate 95% confidence intervals.

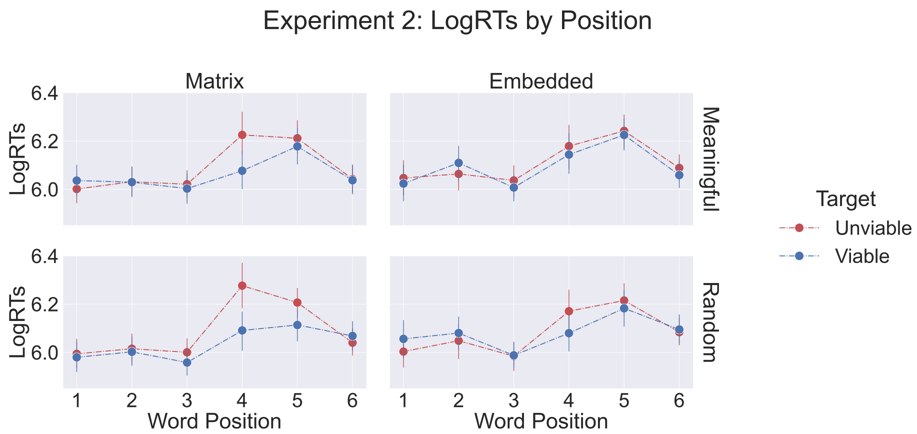

Figure 6: LogRTs for each condition by word position for Experiment 2. Each column indicates STRUCTURE condition; each row indicates CONTEXT condition; each color represents TARGET condition. We find phonotactic distinctions surface immediately in MATRIX conditions; no phonotactic distinctions surface in EMBEDDED conditions.

fixed effects of all predictors and their interactions.[12] We find a main effect of TARGET, such that VIABLE targets are read significantly faster than UNVIABLE ones (β = 0.144, SE = 0.361, t=3.172, p <0.01). No other significant effects or interactions are found.

Positional results for this experiment are visualized in Figure 6. To test how the presence of discourse context modulates phonological judgments, we fit a linear mixed-effects model that predicts the log-transformed RTs of each word in the critical region, with fixed effects of TARGET, STRUCTURE, CONTEXT, position, and their full interactions, along with random intercepts for participant and item.[13] The full model output is shown in Table 2. We find significant main effects of UNVIABLE targets (β=0.158,

[12] The complete formula was: LogSummedRTs ~ TARGET*CONTEXT*STRUCTURE + (1| subject) + (1| item). This model is the maximal model that reaches convergence. The TARGET baseline was the VIABLE condition. The CONTEXT baseline was the MEANINGFUL condition. The STRUCTURE baseline was the MATRIX condition.

[13] The complete formula was: LogRTs ~ TARGET*CONTEXT*STRUCTURE*Position + (1| subject) + (1| item). This model is the maximal model that reaches convergence. The TARGET baseline was the VIABLE condition. The CONTEXT baseline was the MEANINGFUL condition. The STRUCTURE baseline was the MATRIX condition. The position baseline was position 4.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.081	0.054	111.632	<2e-16
Embedded	0.057	0.043	1.321	0.187
Unviable	0.158	0.044	3.613	**0.002**
Random	0.031	0.041	0.760	0.448
Position 5	0.099	0.041	2.413	**0.016**
Embedded:Unviable	-0.103	0.062	-1.671	0.095
Embedded:Random	-0.087	0.058	-1.492	0.136
Unviable:Random	0.016	0.059	0.276	0.783
Embedded:Position 5	-0.009	0.058	-0.155	0.877
Unviable:Position 5	-0.134	0.059	-2.283	**0.023**
Random:Position 5	-0.085	0.058	-1.458	0.145
Embedded:Unviable: Random	0.027	0.083	0.325	0.745
Embedded:Unviable: Position 5	0.089	0.083	1.071	0.284
Embedded:Random: Position 5	0.093	0.082	1.13	0.259
Unviable:Random: Position 5	0.031	0.083	0.373	0.709
Embedded:Unviable: Random:Position 5	-0.048	0.117	-0.413	0.679

Table 2: Model outputs for interaction analysis for Experiment 2. Significant effects and interactions are bolded.

SD=0.044, t=-3.613 p <0.01) and position 5 (β=0.099, SD=0.041, t=2.413, p <0.05), as well as a significant interaction between UNVIABLE targets and position 5 (β=-0.134, SD=0.058, t=-2.283, p <0.05). We report no other significant predictors.

4.2.1 Post-hoc By-STRUCTURE Positional Analyses

In our primary analyses, we find significant main effects of TARGET and position, as well as a significant two-way interaction between TARGET and position; we do not find any significant effects involving STRUCTURE. However, the positional results of Experiment 2 – as visualized in Figure 6 – indicate that an interaction between STRUCTURE and TARGET is approaching significance (p=0.095), suggesting that the significant main effect of TARGET may be driven by differences in the MATRIX condition. To test if each structure displays phonotactic distinctions differently, we ran two post-hoc by-STRUCTURE positional analyses. For each STRUCTURE, we fit a linear mixed-effects model that predicts log-transformed RTs, with fixed effects of TARGET, CONTEXT, and position, and random by-participant and by-item intercepts.[14] Given that we are conducting multiple analyses of the same data, we applied the Bonferroni correction to our significance level (α=0.025).

The significant effects for these two models are visualized in Table 3; we report raw p-values. In the MATRIX model, we find significant main effects of TARGET (β=0.153,

[14]For each STRUCTURE, the complete formula was: LogRTs ~ TARGET*CONTEXT*Position + (1| subject) + (1| item). This model is the maximal model that reaches convergence. The TARGET baseline was the VIABLE condition. The CONTEXT baseline was the MEANINGFUL condition. The position baseline was position 4.

MATRIX	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.081	0.054	112.462	<2e-16
Unviable	0.153	0.046	3.332	**0.001**
Position 5	0.099	0.042	2.384	**0.017**
Unviable: Position 5	-1.34	0.060	-2.254	**0.024**
EMBEDDED	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.141	0.055	111.798	<2e-16
Position 5	0.091	0.040	2.251	**0.024**

Table 3: Model outputs for post-hoc by-STRUCTURE analyses for Experiment 2. Significant effects and interactions are bolded. MATRIX model output is above the EMBEDDED model output.

SD=0.046, t=3.332, p <0.01), position (β=0.099, SD=0.042, t=2.384, p <0.025), and their interaction (β=-1.34, SD=0.060, t=-2.254, p <0.025). In the EMBEDDED model, we only find a significant main effect of position (β=0.09, SD=0.040, t=2.251, p <0.025).

4.3 Discussion

In this experiment, we took the experimental sentences from Experiment 1 and embedded them in a one-sentence discourse context. Despite using the same experimental stimuli, the pattern of results between the two experiments differed significantly. In Experiment 1, phonotactic distinctions between non-word targets surfaced across all syntactic contexts, with distinctions surfacing immediately for embedded structures and later for non-embedded structures; in this experiment, post-hoc analyses for each structure revealed that phonotactic distinctions arose only in MATRIX structures, with no phonotactic distinctions surfacing in EMBEDDED structures.[15] We attribute the difference in results between Experiments 1 and 2 to the layering of a discourse context in addition to syntactic context. When only one layer of embedding was present – EMBEDDED structures in Experiment 1, MATRIX structures following a discourse context in Experiment 2 – phonotactic distinctions surfaced immediately.[16] However, when multiple layers of embedding were present – EMBEDDED structures in Experiment 2 – phonotactic differences between the targets did not influence reading times: readers did not recognize low-level distinctions involving segmental detail. Note that the type of context (MEANINGFUL, RANDOM) did not significantly affect how phonotactic distinctions arose: the *presence* of a context sentence alone was enough to modulate the results between Experiment 1 and Experiment 2.

Additionally, we found that people took longer to read the entire critical region when the UNVIABLE target was present, which differs from the findings of Experiment 1. We attribute this difference to the difference in reading times found in the MATRIX

[15]One possible reason why the interaction between STRUCTURE and TARGET only approaches significance instead of reaching it (in our primary analyses) may be due to the complexity of the model compared to the number of participants that we ran for study. In Experiment 1 (where we observe interactions between STRUCTURE and TARGET), we ran 62 participants. In this experiment, we introduced an additional condition (CONTEXT), but kept the number of participants approximately the same as in Experiment 1 (65 participants for Experiment 2).

[16]Furthermore, effect sizes in conditions with one layer of embedding are similar between the two experiments.

structures. In Experiment 1, we found that the structures "balanced" each other out – MATRIX structures showed differences in the target position, while EMBEDDED structures showed differences in the post-target position. In this experiment, only MATRIX structures display these distinctions.

In tandem, the summation and positional results of this experiment demonstrate that well-established phonotactic distinctions between non-word targets do not always influence the processor, depending on the syntactic and discourse context that the targets appear in. Moreover, one layer of discourse embedding produces patterns of phonotactic judgments that are quite similar to those found within one layer of syntactic embedding, suggesting that the processor may compute syntactic and discourse embeddings similarly; this finding is amenable to previous theoretical work on the interaction between syntax and discourse, such as Discourse Representation Theory (DRT; Kamp 1991).

5 Experiment 3

In the previous experiment, we found that phonotactic distinctions do not surface during on-line processing when syntactic embedding is layered into a discourse. However, we want to ensure that the findings of Experiment 1 and Experiment 2 differ because of the presence of a discourse context before the experimental item, rather than resulting from having to read two sentences. To confirm that the results of Experiment 2 are not an artifact of reading two sentences (Experiment 2) instead of one (Experiment 1), we ran a replication study of Experiment 1 using the two-sentence self-paced reading paradigm by placing the experimental sentences *prior* to the context sentence. Since there is no discourse context present while reading the experimental sentences, we expect that the results of this study should pattern identically to those of Experiment 1.

5.1 Methods

5.1.1 Design & Experimental Stimuli

We used the 24 experimental sentences and contexts from Experiment 2. As mentioned previously, we switched the order of the sentences in this experiment: the experimental sentences in Figure (4b) were read prior to the context sentences in Figure (4a).

5.1.2 Procedure

All participants ($N = 40$) were native speakers of English that were recruited on Prolific. Participants were paid at a rate equivalent to $15/hr for their participation (mean completion time: 11 minutes); compensation conditions were identical to those of the prior experiments presented in this paper, with participants needing to achieve an accuracy above 85% across all comprehension questions. All experiments were conducted via the online research platform PC Ibex (Zehr and Schwarz 2018).

The procedure was otherwise identical to that of Experiment 2.

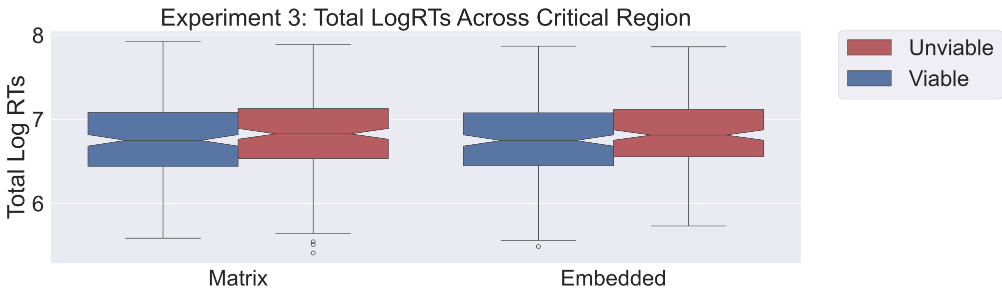

Figure 7: Total Log RTs of the critical region (log(position 4 + position 5)) across STRUCTURES and CONTEXTS for Experiment 3. Notches indicate 95% confidence intervals.

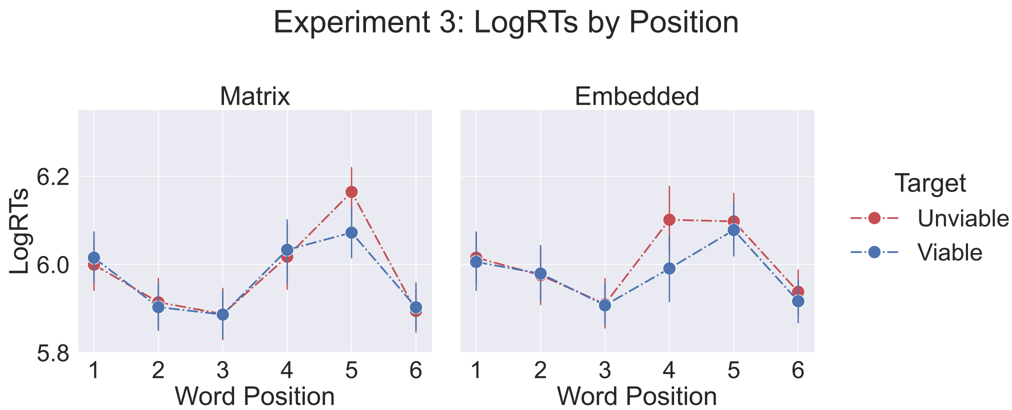

Figure 8: LogRTs for each condition by word position for Experiment 3. Each column indicates STRUCTURE condition; each color represents TARGET condition. We find phonotactic distinctions surface in MATRIX conditions immediately, and surface on the following word in EMBEDDED conditions.

5.2 Results

Summed reading times were analyzed using a linear mixed-effects model that predicts the log-transformed summed RTs across the critical region; this model was identical to the one used in Experiment 1.[17] As visualized in Figure 7, we found no significant differences between the total reading time for the critical region; this finding aligns with the finding of our prior experiments.

Positional results are visualized in Figure 8. As in the previous experiments, we fit a linear mixed-effects model that predicts the log-transformed RTs of each word in the critical region, with fixed effects of TARGET, STRUCTURE, position, and their full interactions, along with random intercepts for participant and item; this model is identical to the model found in Experiment 1.[18]

[17]The complete formula was: LogSummedRTs ∼ TARGET*STRUCTURE + (1| subject) + (1| item). This model is the maximal model that reaches convergence. The TARGET baseline was the VIABLE condition. The STRUCTURE baseline was the MATRIX condition.

[18]The complete formula was: LogRTs ∼ TARGET*STRUCTURE*Position + (1| subject) + (1| item). This model is the maximal model that reaches convergence. The TARGET baseline was the VIABLE condition. The STRUCTURE baseline was the MATRIX condition. The position baseline was position 4.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.029	0.066	90.77	<2e-16
Embedded	-0.025	0.050	-0.49	0.625
Unviable	0.029	0.050	0.578	0.565
Position 5	0.031	0.037	0.826	0.409
Embedded:Unviable	0.101	0.071	1.416	0.161
Embedded:Position 5	0.053	0.053	1.003	0.316
Unviable:Position 5	0.106	0.053	2.013	**0.044**
Embedded:Unviable: Position 5	-0.214	0.075	-2.838	**0.005**

Table 4: Model outputs for interaction analysis for Experiment 3. Significant effects and interactions are bolded.

Model output is presented in Table 4. We found a significant two-way interaction between UNVIABLE targets and position 5 (β = 0.106, SE = 0.053, t=2.013, p <0.05), as well as a significant three-way interaction between UNVIABLE targets, EMBEDDED structures, and position 5 (β = -0.214, SE = 0.075, t=-2.838, p <0.01).

5.3 Discussion

In this experiment, we observed that the general pattern of results that were found in Experiment 1 appear to hold. Said differently, the type of syntactic STRUCTURE that the TARGET appeared in modulated when phonotactic distinctions arose: distinctions surfaced immediately for embedded syntactic structures and later for non-embedded ones. In alignment with the previous two experiments, the summation analyses indicated that there were no processing differences for the critical region across conditions.

More broadly, it is unlikely that our previous results were the result of reading two sentences instead of one. Instead, the findings of this experiment support the conclusion that the difference in results between Experiment 1 and Experiment 2 were due to the presence of a discourse context prior to the experimental sentence.

6 Experiment 4

In the previous three experiments, we found that syntactic and discourse context can influence both when and if phonotactic distinctions surface: phonotactic differences between non-word targets arise immediately when the target is not embedded, arise on the following word when the target is embedded in a single layer of context, and do not arise when the target is embedded in more than one layer of context.

However, phonological modifications to the TARGETS of Experiments 1-3 only occurred in onset position. In this experiment, we modify the phonotactics of the targets' coda positions to explore whether the *phonological* context of a phonotactic structure can also be influenced by higher level structure during reading. Said differently, this experiment explores whether previous results are generalizable to additional phonological positions. Previous work has shown that word-initial segments both inform processing and pattern differently from word-final segments: word-initial segments are more computationally informative cross-linguistically (King and Wedel 2020; Pimentel, Roark, and Cotterell 2020; Pimentel, Cotterell, and Roark 2021), word-initial segments are read more closely and perceived more saliently (Nooteboom 1981; Pisoni

	1	2	3	4	5	6
MATRIX	This	week	the	**desk**	broke	in…
EMBEDDED	He	said	the	**dest**	broke	in…
C-EMBEDDED	The	pencils	the	**desg**	broke	in…

Figure 9: Sample experimental item for Experiment 4. Color for TARGETS are identical to those of Experiment 1: green indicates REAL WORD targets (control). Blue indicates phonologically VIABLE targets. Red indicates phonologically UNVIABLE targets. As in Experiment 1, each phonological TARGET could surface in position 4 for each STRUCTURE; all nine conditions are not shown.

et al. 1985; Hall et al. 2018), and word-initial segments are more resistant to phonological modification (Van Son, Pols, et al. 2003; Smith 2004; McCarthy 2007; McCarthy 2008).

6.1 Methods

6.1.1 Design & Experimental Stimuli

We revised the 27 experimental items from Experiment 1 to use TARGETS with word-final phonotactic violations instead of word-initial ones; phonotactic modifications were either a metathesis of word-final phonemes or a substitution of one of the phonemes in coda position. All contexts outside of the TARGET were unchanged from Experiment 1. A sample item is presented in Figure 9. Additionally, all 27 filler items were identical to those used in Experiment 1.

6.1.2 Procedure

All participants (N = 48) were native speakers of English that were recruited on Prolific. Participants were paid approximately $15/hr for their participation (mean completion time: 10 minutes); compensation conditions were identical to those of Experiment 1, with participants needing to achieve an accuracy above 85% across all comprehension questions. All experiments were conducted via the online research platform PC Ibex (Zehr and Schwarz 2018).

Experiment 4 followed the same self-paced reading procedure as used in Experiment 1: participants read sentences word-by-word by pressing their SPACE bar. Similarly, comprehension questions appeared randomly after one-third of all trials.

6.2 Results

We begin by reporting our analyses across the entire critical region. Results for the critical region are summarized in Figure 10. Given that the design of this experiment is identical to Experiment 1, we fit an identically-structured linear mixed-effects model to predict the log-transformed total reading time across the critical region.[19] Across the entire critical region, we find that VIABLE targets are read slower than REAL targets (β = -0.155, SE = 0.035, t=-4.462, p <0.001) but faster than UNVIABLE targets (β =

[19]The complete formula was: LogSummedRTs ~ TARGET*STRUCTURE + (1| subject) + (1| item). This model is the maximal model that reaches convergence. The TARGET baseline was the VIABLE condition. The STRUCTURE baseline was the MATRIX condition.

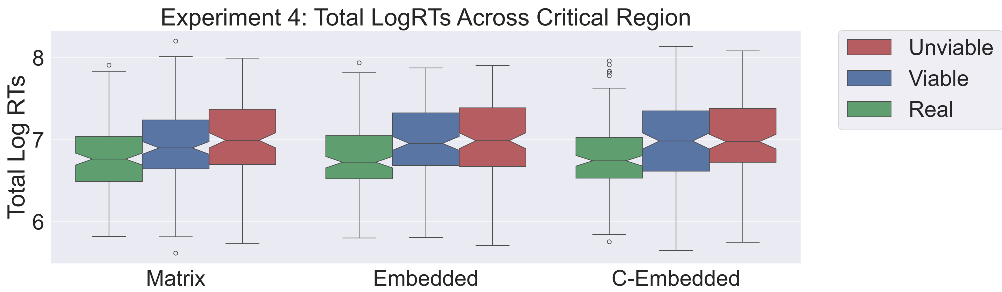

Figure 10: Total Log RTs of the critical region (log(position 4 + position 5)) across STRUCTURES for Experiment 4. Notches indicate 95% confidence intervals. No significant differences between non-word TARGETS were noted.

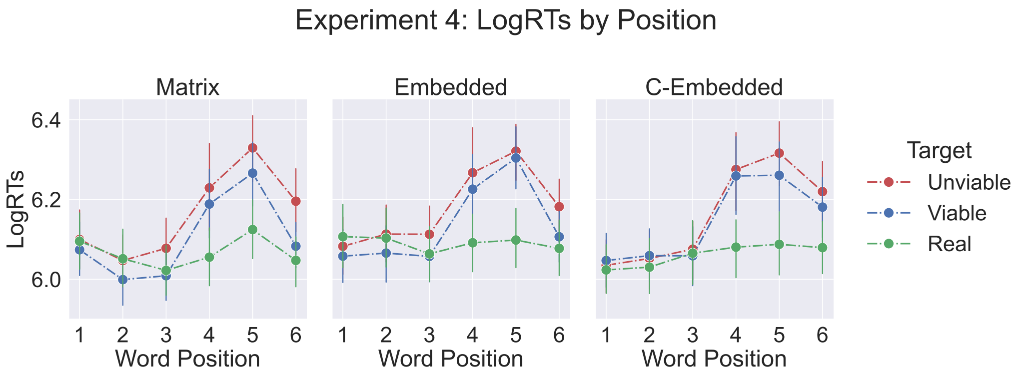

Figure 11: LogRTs for each condition by word position for Experiment 4. Each subplot indicates STRUCTURE condition; each color represents TARGET condition. No significant differences between non-word TARGETS are reported in any position.

0.079, SE = 0.036, t=2.206, p <0.05) . This result differs from the results of Experiment 1, where neither phonological acceptability nor syntactic complexity increased the total reading time of the critical region.

Positional results are visualized in Figure 11. Using an identically-structured linear mixed-effects model as to the one that was used in Experiment 1,[20] we report no significant differences between VIABLE and UNVIABLE targets at either position 4 or position 5 for all STRUCTURES; model outputs are presented in Table 5. The only significant differences that maintained from Experiment 1 to Experiment 4 were between the viable TARGETS and the REAL words in both positions, where we find viable TARGETS take significantly longer to read than the REAL words across all STRUCTURES. No significant differences between VIABLE and UNVIABLE targets were observed.

6.3 Discussion

The results of Experiment 4 only partially align with those of Experiment 1.

As in Experiment 1, Experiment 4 demonstrates that the total reading time for the critical region is different between non-word targets and real words: across all STRUC-

[20]The complete formula was: LogRTs ∼ TARGET*STRUCTURE*Position + (1| subject) + (1| item). This model is the maximal model that reaches convergence. The TARGET baseline was the VIABLE condition. The STRUCTURE baseline was the MATRIX condition. The position baseline was position 4.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.192	0.063	98.323	0
Embedded	0.041	0.044	0.932	0.352
C-Embedded	0.068	0.044	1.523	0.128
Unviable	0.065	0.045	1.423	0.155
Real	-0.138	0.044	-3.118	**0.002**
Position 5	0.065	0.041	1.585	0.113
Embedded:Unviable	-0.001	0.064	-0.009	0.993
C-Embedded:Unviable	-0.029	0.064	-0.445	0.656
Embedded:Real	-0.002	0.063	-0.027	0.978
C-Embedded:Real	-0.040	0.063	-0.64	0.522
Embedded:Position 5	0.009	0.059	0.146	0.884
C-Embedded:Position 5	-0.054	0.059	-0.913	0.362
Unviable:Position 5	-0.006	0.060	-0.101	0.920
Real:Position 5	0.004	0.059	0.069	0.945
Embedded:Unviable: Position 5	-0.033	0.085	-0.386	0.699
C-Embedded:Unviable: Position 5	0.013	0.085	0.146	0.884
Embedded:Real: Position 5	-0.070	0.084	-0.837	0.403
C-Embedded:Real: Position 5	-0.008	0.084	-0.100	0.92

Table 5: Model outputs for interaction analysis for Experiment 4. Significant effects and interactions are bolded.

TURES, we find that non-word targets take significantly longer to read than real words. However, we also find that critical regions of sentences with the VIABLE targets are read faster than UNVIABLE targets. Additionally, the positional phonotactic acceptability differences between VIABLE and UNVIABLE targets that we observed in Experiment 1 do not maintain in Experiment 4: neither embedded clauses nor non-embedded clauses display phonotactic distinctions between the two non-word TARGETS, in any position.

These results appear to display distinctions according to both lexical status and phonological acceptability. We interpret this finding to be a consequence of errors in retrieval during lexical access, as previously reported in the literature (Lukatela and Turvey 1994; Christophe et al. 2004): participants likely begin to retrieve a TARGET as they read, given the identical onset and rime of each TARGET. However, as they continue to read the non-word TARGETS, they encounter an unexpected sequence of phonemes (that are either phonotactically viable or unviable). As such, their partially-retrieved lexeme fails, so participants must dedicate some additional processing resources to accommodate this unexpected sequence; we do not observe these effects when modifying the onset position of the word, as lexical retrieval fails much earlier.

In sum, the findings of Experiment 4 partially align with prior psycholinguistic and phonological research: word beginnings matter more than word endings during reading when looking at each word in the sentence, but perhaps such distinctions accumulate across many words during lexical access. We contribute to this literature by demonstrating that processing slowdowns caused by phonotactic distinctions arise on a single word region in sentences with onset-modified non-words, but that such slowdowns arise across multiple word regions in sentences with coda-modified non-words.

7 General Discussion

In our first two experiments, we explored how non-words of varying phonotactic acceptability were processed in different syntactic and discursive contexts. In Experiment 1, we found that phonotactic distinctions between VIABLE and UNVIABLE targets surface immediately in embedded structures, whereas non-embedded structures display delayed phonotactic distinctions. In Experiment 2, we noted that the presence of a discourse context significantly affects how phonotactic differences arise during reading: if a non-word target is embedded both syntactically and discursively, phonotactic differences between non-word targets disappear, while differences between targets that are only discursively-embedded pattern identically to those that are only syntactically-embedded.[21] These findings both challenge previous studies on phonotactic acceptability – as we report that phonotactic distinctions do not always surface – and suggest that syntactic embedding and discourse embedding may influence the processor similarly.

In Experiment 3, we confirmed that the findings of Experiment 2 are due to the presence of a preceding discourse context: the patterns of phonotactic distinctions from Experiment 1 return – both in timing and presence – when the experimental sentence is placed before the context sentence. These results align with the patterns of Experiment 1, which wholly lacked discourse context.

In Experiment 4, we observed that the results of Experiment 1 do not maintain across different phonological contexts. Phonotactic distinctions surface in targets with modified coda positions, but when comparing the total reading times across the critical region. These findings align with previous research that has found differences in the behavior, patterning, and processing of word-initial and word-final phonological structure (Nooteboom 1981; Pisoni et al. 1985; Smith 2004; McCarthy 2007; McCarthy 2008; Hall et al. 2018; King and Wedel 2020; Pimentel, Roark, and Cotterell 2020, inter alia).

Across our experiments, we find that certain manipulations cause cumulative differences in total reading time for the critical region. Also, we observe that, following the critical region, the processor is not influenced by our experimental manipulations: reading times for words following the critical region do not appear to display significant spillover effects, and high accuracy on comprehension questions that asked about material following the non-words suggests that people still successfully processed material after the non-words.

[21] One anonymous reviewer notes that, because the non-word targets overlap with their real-word counterparts – rime overlap (Experiments 1-4), coda overlap (Experiments 1-3), and onset overlap (Experiment 4) – pre-activation of the real-word target in context may influence our results. While such pre-activation likely occurs to some extent, we argue that our experimental design attempts to minimize these effects. To avoid differences in pre-activation caused by syntactic context, we structured our stimuli to not reveal any significant semantic information about the non-word target until after the target appeared: in {Yesterday morning, Serena hoped} the {prant, psant} was at the store, the three words before the target word (in either syntactic condition) provide minimal expectations about the upcoming word. Controlling the structures in this way also helps reduce differences in pre-activation caused by discourse context, as participants do not know how (or whether) the first and second sentences are related until after the target has been read.

Given the varying differences in reading times *across* the critical region and *within* the critical region, the results of our four experiments support a model of sentence processing that is sensitive to the *incremental* interaction between different layers of linguistic information, where phonological, syntactic and discourse factors govern the presence and timing of linguistic distinctions (in our case, phonotactic distinctions between non-word targets of varying acceptability). Crucially, the processor does not necessarily linger on these distinctions beyond the critical region.

The mechanism by which the sentence processor manages these different layers of linguistic information is currently unclear. We propose one possible mechanism: good-enough processing (Ferreira and Patson 2007; Traxler 2014; Christianson 2016). As evident in the name of the mechanism, good-enough processing models suggest that people only construct representations of linguistic structure that make sense for the task; fully-specified details for each word or structure may not be used. We argue that this mechanism generally accounts for the findings of this paper: people are able to determine differences between fine-grained representations of sound when input complexity is minimal (isolated acceptability judgments, from prior work; conditions with no layers of embedding, in our experiments), but must move away from such low-level comparisons as input complexity increases (via the introduction of syntactic and/or discourse embedding, in our experiments).[22] Then, once people have developed good-enough representations of the word (or sentence), they continue onto the next word, thus accounting for why we do not notice significant differences across the full critical region in many of our experiments. To our knowledge, good-enough processing has not yet been extended to sub-lexical phonological structure.

In addition to furthering sentence processing research, this paper also contributes to theoretical linguistic research: computational models of phonotactic distinctions are well-established in the literature (Hayes and Wilson 2008; Albright 2009; Hayes and White 2013), though these models only address phonotactic structure in isolation. The stimuli and data from this study could be used to build more robust computational models of phonotactic structure. Moreover, we join the recent rise in theoretically-informed experimental research on phonotactics (Breiss and Hayes 2020; Avcu and Hestvik 2020; Sundara et al. 2022; Kuo 2024, *inter alia*), demonstrating how phonotactic structure is processed in context.

One key assumption of this paper (and of much prior work) is that the processor operates on discrete levels of linguistic input: sub-lexical phonological structure, syntactic structures, discourse, etc. What if the processor does not distinguish different levels of linguistic input, instead computing as much of the signal as it can using top-down and bottom-up information interactively (e.g. Marslen-Wilson and Tyler 1980)? We argue that our results demonstrate that the signal must dynamically prioritize different aspects of the signal over others: the timing and presence of fine-grained representations

[22]Parallels between how people process one level of syntactic embedding and one level of discourse embedding may arise in our experiments due to how our stimuli were constructed: neither the syntactic embedding nor the contextual embedding introduces useful semantic content, as we are looking at how the presence of these structures affects processing of phonotactic distinctions. Accordingly, embedding the non-words in a layer of syntactic or discourse context introduces additional linguistic material for the processor to maintain during the self-paced reading task. Notably, our approach differs from previous experiments that examine how context influences the processor; these studies often introduce discourse contexts that facilitate processing of upcoming linguistic material (Crain and Steedman 1985; Warner and Glass 1987). Such prior work may not observe the same parallels between syntactic and discourse embedding.

of bottom-up phonological information can vary according to top-down information.

A smaller finding of this work is that syntactic embedding and discourse embedding appear to similarly influence how and when phonotactic distinctions arise between non-word targets. Some prior theoretical work has supported similar underlying representations between syntax and discourse, particularly Discourse Representation Theory (DRT) (Givón 1979; Polanyi and Scha 1983; Hopper and Thompson 1984; Mann and Thompson 1988; Taboada and Mann 2006; Kamp 1991; Kamp and Reyle 2013). In DRT, predicates and their referents are represented as Discourse Representation Structures (DRSs), where referents are mapped to their function. For example, the DRS for the sentence *The delegate arrived* would be (10) below (Kamp 1991):

(10) <{x}, {delegate(x), arrive(x)}>

Likewise, should there be more discourse – *The delegate arrived. She ate dinner.* – DRSs simply incorporate the new material into the structure as (11):

(11) <{x}, {delegate(x), arrive(x), ate(x, dinner)}>[23]

As such, syntactic embedding and discourse embedding are accounted for in a similar manner within DRT. While the goal of our experiments was not to probe how to represent discourse, our findings are amenable to theoretical perspectives where people incorporate syntactic and discourse information into similar underlying structures; DRT is one such approach.

Finally, this paper displays how small methodological decisions can greatly impact the presence and timing of distinctions for a well-established linguistic phenomenon. Many psycholinguistic studies test only a small number of structures or place target words in regular carrier phrases. We hope that future psycholinguistic and cognitive research tests their phenomena in a variety of structures and contexts to ensure their results are robust.

8 Conclusion

In this paper, we investigated how differences between low-level representations of sound surface during on-line sentence processing. We found that well-established phonotactic distinctions of non-word targets *generally* surface, though the phonological, syntactic, and discourse contexts within which the targets appear greatly influences how such distinctions arise: one layer of syntactic or discourse embedding affects the timing of the distinctions, multiple layers of embedding eliminates the distinctions, and phonological context affects the presence of distinctions.

Broadly, our results contribute to cognitively-oriented fields in a number of ways. First, we find that distinctions based on phonotactic acceptability, which are well established and have previously been evaluated in isolation from any larger linguistic context (Vitevitch, Luce, Charles-Luce, et al. 1997; Albright 2009; Linzen and Gallagher 2017, *inter alia*), do not always persist when the non-word targets are placed in some

[23]The DRS for this structure has been simplified for readability purposes; usually, dinner would be treated as its own variable y in the DRS.

context. These differences can be influenced by the position of the phonological modification within the word, as well as the level of syntactic and discourse embedding of the target.

Second, our results demonstrate that different aspects of the linguistic signal must be taken into consideration with one another: high-level properties of the linguistic signal, like syntactic and discourse context, interact with low-level properties, like phonotactics. Other work has found interactions between top-down and bottom-up factors during sentence processing – namely Marslen-Wilson and Tyler (1980) and Britt et al. (1992) – and our experiments complement these studies and extend their generalizations to the domain of fine-grained phonological processing during reading.

Third and finally, our findings raise the need to systematically evaluate other kinds of well-established linguistic judgments within different contexts and using different paradigms. For example, many linguistic studies place their target phenomenon in a consistent carrier phrase. However, in this paper, we found that placing the experimental sentences after a one-sentence context alone can greatly influence both the timing and presence of a judgment. As such, prior results may not replicate after manipulation of the contexts within which the target phenomenon is placed. We hope that future psycholinguistic studies conduct rigorous examinations of how different kinds of context – both local and global – affect their results. Such testing allows us to assess how robust our findings are.

References

Acheson, Daniel J and Maryellen C MacDonald (2011). "The rhymes that the reader perused confused the meaning: Phonological effects during on-line sentence comprehension." In: *Journal of Memory and Language* 65.2, pp. 193–207. DOI: https://doi.org/10.1016/j.jml.2011.04.006.

Adriaans, Frans and René Kager (2017). "Learning novel phonotactics from exposure to continuous speech." In: *Laboratory Phonology* 8.1. DOI: https://doi.org/10.5334/labphon.20.

Albright, Adam (2009). "Feature-based generalisation as a source of gradient acceptability." In: *Phonology* 26.1, pp. 9–41. DOI: doi:10.1017/S0952675709001705.

Avcu, Enes and Arild Hestvik (2020). "Unlearnable phonotactics." In: *Glossa: a journal of general linguistics* 5.1. DOI: https://doi.org/10.5334/gjgl.892.

Bader, Markus and Michael Meng (1999). "Subject-object ambiguities in German embedded clauses: An across-the-board comparison." In: *Journal of Psycholinguistic Research* 28, pp. 121–143. DOI: https://doi.org/10.1023/A:1023206208142.

Bailey, Todd M and Ulrike Hahn (2001). "Determinants of wordlikeness: Phonotactics or lexical neighborhoods?" In: *Journal of Memory and Language* 44.4, pp. 568–591. DOI: https://doi.org/10.1006/jmla.2000.2756.

Breen, Mara and Charles Clifton Jr (2011). "Stress matters: Effects of anticipated lexical stress on silent reading." In: *Journal of Memory and Language* 64.2, pp. 153–170. DOI: https://doi.org/10.1016/j.jml.2010.11.001.

Breiss, Canaan (2020). "Constraint cumulativity in phonotactics: Evidence from artificial grammar learning studies." In: *Phonology* 37.4, pp. 551–576. DOI: https://doi.org/10.1017/S0952675720000275.

Breiss, Canaan and Bruce Hayes (2020). "Phonological markedness effects in sentence formation." In: *Language* 96.2, pp. 338–370. DOI: https://dx.doi.org/10.1353/lan.2020.0023..

Britt, M Anne et al. (1992). "Parsing in discourse: Context effects and their limits." In: *Journal of Memory and Language* 31.3, pp. 293–314.

Chomsky, Noam. and Morris Halle (1968). *The Sound Pattern of English*. English. Harper & Row New York, xiv, 470 p.

Christianson, Kiel (2016). "When language comprehension goes wrong for the right reasons: Good-enough, underspecified, or shallow language processing." In: *Quarterly Journal of Experimental Psychology* 69.5, pp. 817–828. DOI: https://doi.org/10.1080/17470218.2015.1134603.

Christophe, Anne et al. (2004). "Phonological phrase boundaries constrain lexical access I. Adult data." In: *Journal of Memory and Language* 51.4, pp. 523–547. DOI: https://doi.org/10.1016/j.jml.2004.07.001.

Crain, Stephen and Mark Steedman (1985). "On not being led up the garden path: the use of context by the psychological syntax processor." In: *Natural Language Parsing: Psycholinguistic, Computational, and Theoretical Perspectives*.

Dankoviþová, Jana et al. (1998). "Phonotactic grammaticality is gradient." In.

Ferreira, Fernanda and John M Henderson (1990). "Use of verb information in syntactic parsing: evidence from eye movements and word-by-word self-paced reading." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16.4, p. 555. DOI: 10.1037//0278-7393.16.4.555.

Ferreira, Fernanda and John M Henderson (1991). “Recovery from misanalyses of garden-path sentences.” In: *Journal of Memory and Language* 30.6, pp. 725–745. DOI: https://doi.org/10.1016/0749-596X(91)90034-H.

Ferreira, Fernanda and Nikole D Patson (2007). “The ‘good enough’approach to language comprehension.” In: *Language and Linguistics Compass* 1.1-2, pp. 71–83. DOI: https://doi.org/10.1111/j.1749-818X.2007.00007.x.

Fodor, Janet Dean (2002). “Psycholinguistics cannot escape prosody.” In: *Proceedings of the 1st International Conference on Speech Prosody*.

Frazier, Lyn (1979). *On comprehending sentences: Syntactic parsing strategies*. University of Connecticut.

Frisch, Stefan A et al. (2001). “Emergent phonotactic generalizations in English and Arabic.” In: *Typological Studies in Language* 45, pp. 159–180. DOI: https://doi.org/10.1075/tsl.45.09fri.

Gibson, Edward, Timothy Desmet, et al. (2005). “Reading relative clauses in English.” In: *Cognitive Linguistics*. DOI: https://doi.org/10.1515/cogl.2005.16.2.313.

Gibson, Edward and H-H Iris Wu (2013). “Processing Chinese relative clauses in context.” In: *Language and Cognitive Processes* 28.1-2, pp. 125–155. DOI: https://doi.org/10.1080/01690965.2010.536656.

Givón, Talmy (1979). “From discourse to syntax: Grammar as a processing strategy.” In: *Discourse and syntax*. Brill, pp. 81–112. DOI: https://doi.org/10.1163/9789004368897_005.

Hall, Kathleen Currie et al. (2018). “The role of predictability in shaping phonological patterns.” In: *Linguistics Vanguard* 4.s2. DOI: https://doi.org/10.1515/lingvan-2017-0027.

Hayes, Bruce and James White (2013). “Phonological naturalness and phonotactic learning.” In: *Linguistic Inquiry* 44.1, pp. 45–75. DOI: https://doi.org/10.1162/LING_a_00119.

Hayes, Bruce and Colin Wilson (2008). “A maximum entropy model of phonotactics and phonotactic learning.” In: *Linguistic Inquiry* 39.3, pp. 379–440. DOI: https://doi.org/10.1162/ling.2008.39.3.379.

Hopper, Paul J and Sandra A Thompson (1984). “The discourse basis for lexical categories in universal grammar.” In: *Language* 60.4, pp. 703–752. DOI: 10.1353/lan.1984.0020.

Hsiao, Franny and Edward Gibson (2003). “Processing relative clauses in Chinese.” In: *Cognition* 90.1, pp. 3–27. DOI: https://doi.org/10.1016/S0010-0277(03)00124-0.

Ishizuka, Tomoko (2005). “Processing relative clauses in Japanese.” In: *UCLA Working papers in Linguistics* 13, pp. 135–157.

Jarvella, Robert J and Steven J Herman (1972). “Clause structure of sentences and speech processing.” In: *Perception & Psychophysics* 11, pp. 381–384.

Just, Marcel A, Patricia A Carpenter, and Jacqueline D Woolley (1982). “Paradigms and processes in reading comprehension.” In: *Journal of Experimental Psychology* 111.2, p. 228. DOI: 10.1037//0096-3445.111.2.228.

Kamp, Hans (1991). “A theory of truth and semantic representation.” In: *Meaning and the Dynamics of Interpretation*. Brill, pp. 329–369. DOI: https://doi.org/10.1515/9783110867602.1.

Kamp, Hans and Uwe Reyle (2013). *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*.

Vol. 42. Springer Science & Business Media. DOI: https://doi.org/10.1007/978-94-017-1616-1.
Keegan, John M (1997). *A reference grammar of Mbay*.
King, Adam and Andrew Wedel (2020). "Greater early disambiguating information for less-probable words: The lexicon is shaped by incremental processing." In: *Open Mind* 4, pp. 1–12. DOI: 10.1162/opmi_a_00030.
Kirby, James P and Alan CL Yu (2007). "Lexical and phonotactic effects on wordlikeness judgments in Cantonese." In: *Proceedings of the International Congress of the Phonetic Sciences*. Vol. 16.
Kluender, Robert and Marta Kutas (1993). "Subjacency as a processing phenomenon." In: *Language and Cognitive Processes* 8.4, pp. 573–633. DOI: https://doi.org/10.1080/01690969308407588.
Ko, Kara L (1998). "The comprehension of main and embedded clauses." PhD thesis. Massachusetts Institute of Technology.
Kuo, Jennifer (2024). "Types of statistical knowledge in alternation learning: insights from artificial grammar learning." In: *West Coast Conference on Formal Linguistics*.
Kuznetsova, Alexandra, Per B Brockhoff, and Rune HB Christensen (2017). "lmerTest package: tests in linear mixed effects models." In: *Journal of Statistical Software* 82, pp. 1–26. DOI: 10.18637/jss.v082.i13.
Linzen, Tal and Gillian Gallagher (2017). "Rapid generalization in phonotactic learning." In: *Laboratory Phonology* 8.1. DOI: 10.5334/labphon.44.
Lord, Carol (2002). "Are subordinate clauses more difficult." In: *Complex sentences in grammar and discourse*, pp. 224–233.
Luce, Paul A and David B Pisoni (1998). "Recognizing spoken words: The neighborhood activation model." In: *Ear and Hearing* 19.1, pp. 1–36. DOI: 10.1097/00003446-199802000-00001.
Lukatela, Georgije and Michael T Turvey (1994). "Visual lexical access is initially phonological: 2. Evidence from phonological priming by homophones and pseudohomophones." In: *Journal of Experimental Psychology* 123.4, p. 331. DOI: https://doi.org/10.1037/0096-3445.123.4.331.
Magne, Cyrille, Reyna L Gordon, and Swati Midha (2010). "Influence of metrical expectancy on reading words: An ERP study." In: *Speech Prosody 2010-Fifth International Conference*.
Mann, William C and Sandra A Thompson (1988). "Rhetorical structure theory: Toward a functional theory of text organization." In: *Text-interdisciplinary Journal for the Study of Discourse* 8.3, pp. 243–281. DOI: https://doi.org/10.1515/text.1.1988.8.3.243.
Marslen-Wilson, William and Lorraine Komisarjevsky Tyler (1980). "The temporal structure of spoken language understanding." In: *Cognition* 8.1, pp. 1–71. DOI: https://doi.org/10.1016/0010-0277(80)90015-3.
McCarthy, John J (2007). "Slouching toward optimality: Coda reduction in OT-CC." In: *Linguistics Department Faculty Publication Series*, p. 74.
McCarthy, John J (2008). "The gradual path to cluster simplification." In: *Phonology* 25.2, pp. 271–319. DOI: 10.1017/S0952675708001486.
McCurdy, Kate, Gerrit Kentner, and Shravan Vasishth (2013). "Implicit prosody and contextual bias in silent reading." In: *Journal of Eye Movement Research* 6.2. DOI: https://doi.org/10.16910/jemr.6.2.4.
McElree, Brian and Teresa Griffith (1995). "Syntactic and thematic processing in sentence comprehension: Evidence for a temporal dissociation." In: *Journal of Exper-*

imental Psychology: Learning, Memory, and Cognition 21.1, p. 134. DOI: https://doi.org/10.1037/0278-7393.21.1.134.
Merity, Stephen et al. (2016). *Pointer Sentinel Mixture Models*. arXiv: 1609.07843 [cs.CL].
Mollin, Sandra (2012). "Revisiting binomial order in English: ordering constraints and reversibility1." In: *English Language & Linguistics* 16.1, pp. 81–103. DOI: https://doi.org/10.1017/S1360674311000293.
Morgan, Emily and Roger Levy (2016). "Abstract knowledge versus direct experience in processing of binomial expressions." In: *Cognition* 157, pp. 384–402. DOI: https://doi.org/10.1016/j.cognition.2016.09.011.
Nooteboom, Sieb G (1981). "Lexical retrieval from fragments of spoken words: Beginnings vs endings." In: *Journal of Phonetics* 9.4, pp. 407–424. DOI: https://doi.org/10.1016/S0095-4470(19)31017-4.
Pimentel, Tiago, Ryan Cotterell, and Brian Roark (Apr. 2021). "Disambiguatory Signals are Stronger in Word-initial Positions." In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Ed. by Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty. Online: Association for Computational Linguistics, pp. 31–41. DOI: 10.18653/v1/2021.eacl-main.3. URL: https://aclanthology.org/2021.eacl-main.3.
Pimentel, Tiago, Brian Roark, and Ryan Cotterell (2020). "Phonotactic complexity and its trade-offs." In: *Transactions of the Association for Computational Linguistics* 8, pp. 1–18. DOI: https://doi.org/10.1162/tacl_a_00296.
Pisoni, David B et al. (1985). "Speech perception, word recognition and the structure of the lexicon." In: *Speech Communication* 4.1-3, pp. 75–95. DOI: 10.1016/0167-6393(85)90037-8.
Ploujnikov, Artem and Mirco Ravanelli (2022). *SoundChoice: Grapheme-to-Phoneme Models with Semantic Disambiguation*. arXiv: 2207.13703 [cs.SD].
Plummer, Patrick and Keith Rayner (2012). "Effects of parafoveal word length and orthographic features on initial fixation landing positions in reading." In: *Attention, Perception, & Psychophysics* 74, pp. 950–963. DOI: https://doi.org/10.3758/s13414-012-0286-z.
Polanyi, Livia and Remco JH Scha (1983). "The syntax of discourse." In: *Text-Interdisciplinary Journal for the Study of Discourse* 3.3, pp. 261–270. DOI: https://doi.org/10.1515/text.1.1983.3.3.261.
Ravanelli, Mirco et al. (2021). *SpeechBrain: A General-Purpose Speech Toolkit*. arXiv:2106.04624. arXiv: 2106.04624 [eess.AS].
Rayner, Keith, Simon Garrod, and Charles A Perfetti (1992). "Discourse influences during parsing are delayed." In: *Cognition* 45.2, pp. 109–139. DOI: https://doi.org/10.1016/0010-0277(92)90026-E.
Scholes, Robert J (2016). *Phonotactic grammaticality*. Vol. 50. Walter de Gruyter. DOI: https://doi.org/10.1515/9783111352930.
Shademan, Shabnam (2006). "Is phonotactic knowledge grammatical knowledge?" In: *Proceedings of the 25th West Coast Conference on Formal Linguistics (WCCFL)*. Vol. 371379. Citeseer.
Shademan, Shabnam (2007). "Grammar and analogy in phonotactic well-formedness judgments." PhD thesis. Citeseer.
Siyanova-Chanturia, Anna, Kathy Conklin, and Walter JB Van Heuven (2011). "Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing

of multiword sequences." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37.3, p. 776. DOI: 10.1037/a0022531.

Smith, Jennifer L (2004). *Phonological augmentation in prominent positions*. Routledge. DOI: 10.4324/9780203506394.

Snedeker, Jesse and John Trueswell (2003). "Using prosody to avoid ambiguity: Effects of speaker awareness and referential context." In: *Journal of Memory and Language* 48.1, pp. 103–130. DOI: https://doi.org/10.1016/S0749-596X(02)00519-3.

Sundara, Megha et al. (2022). "Infants' developing sensitivity to native language phonotactics: a meta-analysis." In: *Cognition* 221, p. 104993.

Taboada, Maite and William C Mann (2006). "Rhetorical structure theory: Looking back and moving ahead." In: *Discourse Studies* 8.3, pp. 423–459. DOI: https://doi.org/10.1177/1461445606061881.

Traxler, Matthew J (2014). "Trends in syntactic parsing: Anticipation, Bayesian estimation, and good-enough parsing." In: *Trends in Cognitive Sciences* 18.11, pp. 605–611. DOI: https://doi.org/10.1016/j.tics.2014.08.001.

Van Son, RJJH, Louis CW Pols, et al. (2003). "How efficient is speech." In: *Proceedings of the Institute of Phonetic Sciences*. Vol. 25. University of Amsterdam, pp. 171–184.

Vitevitch, Michael S and Paul A Luce (1998). "When words compete: Levels of processing in perception of spoken words." In: *Psychological Science* 9.4, pp. 325–329. DOI: https://doi.org/10.1111/1467-9280.00064.

Vitevitch, Michael S, Paul A Luce, Jan Charles-Luce, et al. (1997). "Phonotactics and syllable stress: Implications for the processing of spoken nonsense words." In: *Language and Speech* 40.1, pp. 47–62. DOI: 10.1177/002383099704000103.

Vitevitch, Michael S, Paul A Luce, David B Pisoni, et al. (1999). "Phonotactics, neighborhood activation, and lexical access for spoken words." In: *Brain and Language* 68.1-2, pp. 306–311. DOI: 10.1006/brln.1999.2116.

Warner, John and Arnold L Glass (1987). "Context and distance-to-disambiguation effects in ambiguity resolution: Evidence from grammaticality judgments of garden path sentences." In: *Journal of Memory and Language* 26.6, pp. 714–738. DOI: https://doi.org/10.1016/0749-596X(87)90111-2.

Weber, Andrea and Anne Cutler (2006). "First-language phonotactics in second-language listening." In: *The Journal of the Acoustical Society of America* 119.1, pp. 597–607. DOI: 10.1121/1.2141003.

Zehr, Jeremy and Florian Schwarz (2018). *PennController for Internet Based Experiments (IBEX)*.

9 Appendix A

In this appendix, we report the model output for our summed analyses for all four experiments; these model outputs were not included in the original text to improve the readability of the paper. Tables 6, 7, 8, and 9 reflect the model output for the summed analyses of Experiments 1, 2, 3, and 4, respectively. Significant effects are described in the main text.

	Estimate	SE	*t* value	Pr(>|t|)
(Intercept)	6.831	0.053	129.78	**<0.001** ***
STRUCTURE.Embedded	0.041	0.037	1.116	0.266
STRUCTURE.C-Embedded	0.010	0.037	0.283	0.778
TARGET.Unviable	0.070	0.037	1.893	0.060
TARGET.Real	-0.163	0.037	-4.466	**<0.001** ***
STRUCTURE.Embedded:TARGET.Unviable	0.042	0.052	0.792	0.429
STRUCTURE.C-Embedded:TARGET.Unviable	0.018	0.052	0.339	0.735
STRUCTURE.Embedded:TARGET.Real	0.012	0.052	-0.227	0.821
STRUCTURE.C-Embedded:TARGET.Real	0.048	0.052	0.924	0.357

Table 6: Model output for summed analyses for Experiment 1. Significant effects and interactions are bolded.

	Estimate	SE	*t* value	Pr(>|t|)
(Intercept)	6.838	0.053	129.04	**<0.001** ***
STRUCTURE.Embedded	0.061	0.036	1.707	0.090
TARGET.Unviable	0.114	0.036	3.172	**0.002**
CONTEXT.Random	-0.009	0.031	-0.288	0.77
STRUCTURE.Embedded:TARGET.Unviable	-0.078	0.051	-1.531	0.128
STRUCTURE.Embedded:CONTEXT.Random	-0.048	0.047	-1.029	0.304
TARGET.Unviable:CONTEXT.Random	0.034	0.047	0.730	0.466
STRUCTURE.Embedded:TARGET.Unviable: CONTEXT.Random	0.016	0.066	0.237	0.812

Table 7: Model output for summed analyses for Experiment 2. Significant effects and interactions are bolded.

	Estimate	SE	*t* value	Pr(>|t|)
(Intercept)	6.754	0.065	104.26	**<0.001** ***
STRUCTURE.Embedded	0.009	0.044	0.196	0.846
TARGET.Unviable	0.080	0.044	1.838	0.074
STRUCTURE.Embedded:TARGET.Unviable	-0.008	0.062	-0.122	0.903

Table 8: Model output for summed analyses for Experiment 3. Significant effects and interactions are bolded.

	Estimate	SE	*t* value	Pr(>|t|)
(Intercept)	6.948	0.061	114.512	**<0.001** ***
STRUCTURE.Embedded	0.045	0.035	1.294	0.198
STRUCTURE.C-Embedded	0.043	0.035	1.226	0.222
TARGET.Unviable	0.079	0.036	2.206	**0.029**
TARGET.Real	-0.155	0.035	-4.466	**<0.001** ***
STRUCTURE.Embedded:TARGET.Unviable	-0.025	0.050	-0.490	0.625
STRUCTURE.C-Embedded:TARGET.Unviable	-0.034	0.050	-0.678	0.498
STRUCTURE.Embedded:TARGET.Real	-0.036	0.049	-0.732	0.465
STRUCTURE.C-Embedded:TARGET.Real	-0.046	0.049	-0.931	0.353

Table 9: Model output for summed analyses for Experiment 4. Significant effects and interactions are bolded.

10 Appendix B

In this appendix, we report the results of some post-hoc analyses that incorporate additional variables – orthotactic effects & lexical neighborhood density effects – into our statistical models.

Prior psycholinguistic research has found that phonotactic distinctions frequently interact with orthotactic effects and lexical neighborhood density effects (Luce and Pisoni 1998; Vitevitch and Luce 1998; Vitevitch, Luce, Pisoni, et al. 1999; Bailey and Hahn 2001, *inter alia*), especially in spoken word recognition tasks. However, the main statistical models that we present in this paper do not consider the influence of these other orthotactic or neighborhood factors. As noted by one anonymous reviewer, it is not clear from our current analyses whether the phonotactic distinctions that we do find are a result of phonotactic structure, or instead a result of some contribution of orthotactics or lexical neighborhood density. In this section, we will incorporate these measures into our statistical analyses, demonstrating that our primary results replicate.

10.1 Methods

10.1.1 Orthotactics

We approximated orthotactic effects by estimating character-level bigram probabilities using `wiki-text` (Merity et al. 2016), a corpus of over 100 million tokens from the set of verified *Good* and *Featured* articles on Wikipedia. For each word in our critical regions, we calculated the average bigram probability by adding all the bigram probabilities for the word together and then dividing by the total number of bigrams in the word.

10.1.2 Lexical Neighborhood Density

Within our critical regions, real words were converted to their phonemic representations using the CMU Pronunciation Dictionary (CMUdict); phonemic representations for our non-words (or for real words not found in CMUdict) were determined using the well-established `soundchoice` neural-network grapheme-to-phoneme model by `speechbrain` (Ravanelli et al. 2021; Ploujnikov and Ravanelli 2022). These representations were checked by the first author.

Experiment	Model Type	Fixed Effects
1, 3, 4	*Original*	TARGET*STRUCTURE*Position
	Replacement	ORTHO*NEIGHBOR*STRUCTURE*Position
	Maximal	ORTHO*NEIGHBOR*TARGET*STRUCTURE*Position
2	*Original*	TARGET*STRUCTURE*CONTEXT*Position
	Replacement	ORTHO*NEIGHBOR*CONTEXT*STRUCTURE*Position
	Maximal	ORTHO*NEIGHBOR*TARGET*STRUCTURE*CONTEXT*Position

Table 10: Models that were compared for each experiment. All models predicted log-transformed reading times; all models had by-participant and by-item random intercepts.

Then, we approximated lexical neighborhood density effects by counting the number of words that were within a two-phoneme edit distance of each word. We used two-phoneme edit distance instead of the more popular single-phoneme edit distance because our phonotactic manipulations rely on either replacements (single-phoneme change) or metatheses (double-phoneme change). However, given that two-phoneme edit distance allows for significantly large variation in the total number of neighbors, variation which operates on a different scale than our other independent variables[24] and given that we are interested in the relative relationship between sparse and dense lexical neighborhoods rather than the absolute value of neighbors, we re-scaled the number of neighbors within a two-phoneme edit distance using a log-transformation.

10.2 Results

We now present the results of our statistical modeling for each experiment using our orthotactic and lexical neighborhood measures.

For each experiment, we ran two complementary statistical models in addition to those for our primary analyses. To test whether the orthotactic and lexical neighborhood variables better fit the data than our TARGET condition, the first complementary statistical model excluded the TARGET condition and included the orthotactic and lexical neighborhood variables (and their maximal interactions with our non-TARGET conditions); we title this model the *replacement* model. To test the maximal model that was possible, the second complementary statistical model had the same formulas as their original counterparts, but with the inclusion of the orthotactic and lexical neighborhood measures in their maximal interactions (including the TARGET condition); we title this model the *maximal* model. A summary of the various models and their predictors can be found in Table 10.

Given that the interactions are appreciably more complex and difficult to interpret in these models, we primarily report comparisons between the original models, the replacement models, and the maximal models using likelihood ratio tests for each experiment's set of models. Full model output can be found by running our analysis scripts. In Table 11, we report the results of our model comparisons.

[24]For the 3240 items within our critical regions for Experiment 1, the number of neighbors ranged from 6 to 1862, with a geometric mean of 653 and a standard deviation of 587.

Exp.	Model Type	Num. Params.	AIC	BIC	LL	χ^2	DF	Pr(>χ^2)
1	*Original*	21	**3299.9**	**3426.9**	-1628.6			
	Replacement	27	3393.0	3557.3	-1669.5	0.00	6	1
	Maximal	75	3348.9	3805.2	-1599.5	140.12	48	**<0.001**
2	*Original*	19	**3156.4**	**3270.1**	-1559.2			
	Replacement	35	3177.7	3387.1	-1553.8	10.720	16	0.826
	Maximal	67	3180.9	3581.8	-1523.5	60.744	32	**<0.01**
3	*Original*	11	**1927.1**	**1987.4**	-952.56			
	Replacement	19	1951.3	2055.4	-956.65	0.00	8	1
	Maximal	35	1962.5	2154.2	-946.27	20.76	16	0.188
4	*Original*	21	**2174.0**	**2296.3**	-1066.0			
	Replacement	27	2265.1	2422.5	-1105.6	0.00	6	1
	Maximal	75	2190.6	2627.7	-1020.3	170.51	48	**<0.001**

Table 11: Results of likelihood-ratio tests for each set of statistical models.

10.2.1 Experiment 1

For Experiment 1, the maximal model explains significantly more of the variance than the simpler models (χ^2 = 140.12, p < 0.001), though the AIC and BIC for the original model is the lowest. The replacement model adds no significant value over the original model.

10.2.2 Experiment 2

As in Experiment 1, the maximal model for Experiment 2 shows a significant improvement over simpler models (χ^2 = 60.74, p = 0.0016). We also observe that the AIC and BIC are lowest for the original model, and that the replacement model provides no meaningful improvement over the original model.

10.2.3 Experiment 3

Unlike the model comparisons for Experiments 1 and 2, the maximal model for Experiment 3 does not significantly outperform the smaller models (χ^2 = 20.76, p = 0.19), and the replacement model does not explain more of the variance than the original model. Additionally, the AIC and the BIC are lowest for the original model.

10.2.4 Experiment 4

The model comparisons for Experiment 4 align with those for Experiments 1 and 2: the maximal model shows a highly significant improvement over smaller models (χ^2 = 170.51, p < 0.001), the AIC and BIC are lowest for the original model, and the replacement model does not explain more variance than the other models.

10.3 Discussion

In summary, model comparisons for Experiments 1, 2, and 4 show that the maximal models fit the data better than both the original models and the replacement models. The exception to this trend is Experiment 3, where the added complexity of the maximal model does not provide better fit to the data. Across all four experiments, we find that replacement models do not fit the data better than the original models. Additionally, the differences in AIC & BIC across the models for all four experiments indicate that the original models may generalize better than both the replacement models and the maximal models due to their lower complexity. Broadly, these results indicate that orthotactic and neighborhood-density effects do not explain more variance on their own (as in the replacement models), but that they become meaningful when they interact with the TARGET condition (as in the maximal models). In conclusion, the model comparisons presented in this appendix align with our general argument that phonotactic distinctions between non-words are not robust when such non-words are placed in context.