Syntactic context effects on auditory word recognition

Phoebe Gaston, Ellen Lau, & Colin Phillips

Abstract: Better understanding of word recognition requires a detailed account of how top-down and bottom-up information are integrated. In this paper, we use a combination of modeling and experimental work to investigate the mechanism by which expectations from syntactic context influence the processing of perceptual input during word recognition. The distinction between facilitatory and inhibitory mechanisms for the syntactic category constraint is an important aspect of this problem that has previously been underspecified. Syntactic category is also a relatively simple test case for the issue of context in other domains, and our findings demonstrate an approach that can be scaled up to more complex contextual constraints. We first report simulations in jTRACE that predict that the speed with which changes in lexical activation can be observed in dependent measures should depend on the composition of the set of response candidates allowed by the task. This points to an explanation for conflicts that have occurred between different methods regarding the existence and timing of syntactic constraints on lexical cohort competition, and it is relevant for the design and interpretation of experiments involving cohort competition more broadly. These insights informed a new design for the visual world paradigm that distinguishes facilitatory and inhibitory mechanisms and ensures that activation of words from the wrong syntactic category should be detectable if it is occurring. We demonstrate how failure to ensure this could have obscured such activation in previous work, leading to the appearance of an inhibitory constraint. We find that wrong-category competition does occur, a result that is incompatible with an inhibitory syntactic category constraint.

1 Introduction

1.1 Overview

Auditory word recognition requires identifying stored phonological wordforms that are consistent with incoming speech. Many wordforms share initial phonemes, so a large portion of the lexicon (the "cohort") is understood to be activated as the onset of a spoken word is perceived and processed (Luce & Pisoni, 1998; Marslen-Wilson, 1987; McClelland & Elman, 1986; Norris, 1994). These activated words under consideration are said to "compete" for recognition until just a single match remains and the word being heard can therefore be identified. In some cases, hundreds of milliseconds of auditory input may be necessary before a word is uniquely identifiable, or this uniqueness point may not occur until the entire word has been heard (Luce, 1986). However, the likelihood of a lexical candidate under consideration is influenced not only by how well it matches with what is being heard, but also by how well it fits with the sentence context. Therefore, one obvious way in which the recognition process might become faster or more robust is if information from the context also influences the set of likely candidates, allowing a word to become identifiable prior to the point at which it is differentiable on the basis of bottom-up perceptual input alone.

A great deal of work investigating this possibility has indicated that word recognition is indeed influenced by information from the context. However, despite decades of research, fundamental questions remain about how this occurs. One area of particularly strong disagreement has been whether context can impose constraints so quickly that contextually inappropriate lexical candidates are never considered, or even generated, at all. While potentially very useful, a constraint on generation could also have undesirable consequences, in that a word that is ostensibly impossible in the context might be impossible to recognize if it was not

generated as a candidate. Such a constraint would require that information in the context be processed and transformed extremely rapidly into criteria for wordform exclusion, which would have to interface with the same wordform representations activated by the bottom-up cues, and also be able to override those cues. The central issues here are the nature of the restrictions that can be derived from the context, whether or not a top-down cue can take primacy over a perceptual one, and the speed of this potential influence.

In this paper, we investigate the mechanism by which information from syntactic context influences auditory word recognition, using the visual world paradigm. By focusing on the syntactic constraint, we can generate simple initial hypotheses about whether or not a wordform is a good fit for the context. This allows us to ask: when context makes it likely that the word being heard is, for example, a noun, is this information used to facilitate lexical candidates that could be nouns, or inhibit those that can't be nouns? And how quickly does this occur? We use the term "inhibition" to refer to complete inhibition, fully preventing the activation of wrong-category candidates. This type of inhibition gives priority to syntactic over acoustic information and is the only way for a constraint on generation to occur. A facilitatory constraint, in contrast, would simply boost the activation of correct-category candidates, and would allow for influence from either or both information sources. While it seems clear that an inhibitory constraint could hasten word recognition by restricting the set of candidates, a facilitatory constraint could have a similar effect if recognition is based on an activation threshold, which would be reached faster by correct-category items that have received a boost. Support for or against either of these mechanisms would be an important step forward in our understanding of auditory word recognition.

Previous work on syntactic context, reviewed in the next section, has largely focused on whether there are syntactic effects on word recognition at all, and has yielded conflicting findings. Our approach incorporates two new considerations that help to clarify this landscape. First, we use simulations (reported in Section 2) to explore how the response options available to participants in an experimental task can influence whether and when lexical competition effects are observable. Second, we do not assume that the syntactic constraint must be inhibitory (a frequent assumption, though often not expressed in these terms). Thus, the design of our study incorporates the possibility that wrong-category candidates are activated even when a syntactic constraint is in place and ensures that this competition is detectable.

1.2 Background

Many methods have been used to ask whether or not context can constrain the generation of lexical candidates during auditory word recognition. Evidence from cross-modal priming and gating has consistently indicated that contextually illicit competitors are initially available, arguing against a constraint on generation in the case of both syntactic context (Seidenberg et al., 1982; Tanenhaus et al., 1979; Tanenhaus & Donnenwerth-Nolan, 1984; Tyler, 1984; Tyler & Wessels, 1983) and semantic context (Swinney, 1979; Zwitserlood, 1989). However, subsequent evidence from the visual world paradigm suggested the opposite conclusion (Magnuson et al., 2008; Strand et al., 2018), and data from MEG seems to support aspects of both (Gaston & Marantz, 2018). In the following sub-sections, we describe these methods and their findings in more detail, as well as the design and interpretation issues that make it difficult to extract a coherent picture from this literature. The problem of integrating the different contributions of each relevant study can be simplified by breaking down the question into two parts. The first concerns timing: when can we observe evidence of any kind of influence from the syntactic

context? The second concerns the nature of the influence: are wrong-category candidates activated or not? It is important but unacknowledged in many designs that contextual influence could still be observable even if wrong-category competition is occurring. Failure to account for this possibility in an experimental design amounts to an assumption that any syntactic category constraint must be inhibitory.

1.2.1 Cross-modal priming

In a cross-modal priming study, a sentence containing a prime word is presented (usually) auditorily. At some lag(s) from the onset of the prime word, a target word is presented visually, and the dependent measure is the latency of either lexical decision or naming for that target word. Faster reaction times to the target are expected when it is semantically related to the prime.

In order to evaluate the impact of syntactic context on lexical access with cross-modal priming, early investigators took advantage of different-category, different meaning homophones like “watch”, embedding these homophones as primes in syntactically constraining contexts such as "I bought the watch" and "I began to watch" (Tanenhaus et al., 1979). The target word in a given trial would be semantically related to only one of the homophone meanings, e.g. "look" in the current example. On any account, reaction times for “look” should be faster after hearing the verb form of “watch”; the key question is what happens in the noun-biasing context. If syntactic context could fully block consideration of incompatible lexical items, the verb meaning of “watch” should never be accessed in the noun-biasing context, and faster reaction times should not be observed for “look”. However, if syntactic context does not act as an immediate or complete constraint on generation, some effect on reaction time would be observed in the noun-biasing context.

Tanenhaus et al. (1979) measured the naming latency of the target at three time lags after homophone offset: 0 ms, 200 ms, and 600 ms. They indeed found a priming effect (shorter naming latency) in the category-incompatible context immediately following homophone offset, but this did not occur at the later lags. Tanenhaus et al. (1979) took these results as evidence for exhaustive initial access to the meanings associated with the ambiguous primes, with syntactic context only impacting the process later on. This finding was replicated in a similar design (Seidenberg et al., 1982) and then again with lexical decision rather than naming latency as the dependent measure (Tanenhaus & Donnenwerth-Nolan, 1984).

In an influential review, Tanenhaus and Lucas (1987) concluded that evidence for syntactic context effects in cross-modal priming is lacking, as semantic priming still occurs for the category-incompatible meaning. They also argued that such context effects would actually have little utility, in part because syntactic context is rarely fully deterministic of upcoming syntactic category, and syntactic category knowledge does not lead to high conditional probability of any single lexical item. However, Tanenhaus and Lucas (1987) indicated that a possible exception might be the case in which the expected category is of closed-class words, which is a much more restricted set. And indeed, Shillcock and Bard (1993) later conducted a cross-modal priming study on different-category homophones for which one meaning is a closed-class word and the other is an open-class word (e.g. “would”/“wood”). They find faster reaction times for “timber” after “wood” in a noun-constraining context, but not for “timber” after “would” in a context in which a closed-class verb is expected. This was the first and, to our knowledge, only cross-modal priming evidence for selective access due to syntactic context.

Can we safely conclude from the cross-modal priming results that, with the exception of small-set syntactic categories, syntactic context alone does not immediately impact lexical

access? There are several reasons for caution. First, the lack of evidence for a syntactic category constraint could simply reflect the insufficient power of these cross-modal priming studies to detect a true effect. Collapsing across syntactic and semantic context manipulations, a meta-analysis by Lucas (1999) finds a reliable effect size of ~ 0.2 standard deviations for the difference between contextually appropriate and inappropriate priming effects: that is, contextually inappropriate meanings are generated and lead to semantic priming of their associates, but to a consistently lesser degree than contextually appropriate meanings. In most studies, the interaction that would be necessary to show this small effect is not significant, and therefore the authors usually concluded that there was no immediate context effect at all. But Lucas points out that even 50% power to detect the effect size she observes would require 140 participants, and the median sample size across studies is 35. Therefore, the conclusion that context has no effect in cross-modal priming may simply be a result of lack of power. Furthermore, the meta-analysis suggests that there is evidence of activation of the contextually inappropriate meaning. This pattern is not consistent with an inhibitory constraint. The contrast between contexts is instead more consistent with a facilitatory constraint that boosts the activation of correct-category competitors.

A second issue is that these studies rely exclusively on homophones. The question they answer is whether both the contextually appropriate and contextually inappropriate meanings of a wordform are activated. Homophone-based designs cannot indicate whether a lexical candidate with only a contextually inappropriate meaning also becomes activated when it matches the input. To answer this question, a cross-modal priming study would have to probe for semantic priming from a wrong-category cohort competitor of the target, an approach that Zwitserlood

(1989) successfully uses to ask about effects of semantic context. In that study, Zwitserlood reports that contextually inappropriate candidates are activated.

1.2.2 Gating

Gating is a paradigm in which participants are asked to identify word fragments. The technique was pioneered by Grosjean (1980) as a way to determine exactly how much auditory input is necessary for a word to be correctly identified, and what lexical candidates are considered (across participants) before the correct identification point. Grosjean showed that on average 333 ms of input are needed to identify a word presented in isolation, but that this drops to 245 and 153 ms for words presented in short and long sentence contexts, respectively. These initial data only spoke to how quickly a word could be identified, and not to how context impacted the process of getting to that point.

Tyler (1984) reports that when listeners encounter the first 50 ms of a word in a syntactically biasing context, on some portion of trials their completions are words from the inappropriate syntactic category. Tyler argues that this provides further evidence that syntactic contexts fail to constrain word recognition. But her argument would be stronger if there were a comparison with a syntactically unconstraining baseline condition (McAllister, 1988). In a similar experiment to Tyler's, McAllister (1988) found an increase in correct-category candidates relative to a no-context condition, concluding that syntactic context does, indeed, affect the cohort by the first gate, even if it has not eliminated all wrong-category responses. This could mean wrong-category candidates are only partially inhibited, or that the context effect actually works by facilitating correct-category candidates.

Another important consideration for gating is that error rates with respect to syntactic category should only be assessed when it is clear that the cohort is active. If the rate of incorrect-

first-phoneme guesses does not yet differ from chance, indicating that those guesses are not being driven by an activated cohort, then incorrect-category guesses at that sampling point would not be evidence against a context-constrained cohort. Therefore, the context question should not be asked unless deviations from chance are being assessed for both phonological and contextual error rates, at as early a gate as possible. The chance rate for incorrect-first-phoneme guesses can only be assessed when there is no phonological input, so this would need to be at gate 0, or, from another perspective, in an auditory cloze completion task. Only then could we assess whether Tyler's (1984) 40% incorrect first phonemes at Gate 1 reflect phonological competition or not, and therefore whether asking about the syntactic category error rate is informative or not. Finally, the ideal gate size for an accurate time course estimate has yet to be established.

1.2.3 Visual world paradigm

The visual world paradigm tracks participants' looks to items in a visual display while they listen to spoken sentences. Allopenna, Magnuson, and Tanenhaus (1998) proposed an explicit correspondence between observed fixation probabilities in the visual world and lexical activation of the words corresponding to the pictures. Since then, visual world studies have consistently demonstrated more fixations to pictures whose names are cohort competitors of the word being heard than pictures whose names are unrelated distractors. One clear advantage of the VWP in probing cohort competition, relative to the methods previously described, is that fixations can be recorded continuously from the onset of the word. Fixation probability thus provides an immediate and incremental measure understood to be driven at least to some extent by lexical activation. However, simple noun/verb contextual constraints like "to" and "the," which had been used in the gating and cross-modal priming studies, cannot be used straight-forwardly in the visual world because the method requires lexical candidates to be representable

as relatively simple pictures, and this is more challenging for verbs than for nouns. Therefore, until very recently, the two strongest pieces of evidence from the VWP regarding (morpho-) syntactic contextual constraints on the cohort used substantially different designs (Dahan et al., 2000; Magnuson et al., 2008). Both showed that in a syntactically constraining context a wrong-category competitor is not fixated more than a distractor item. This suggests that comprehenders use the contextual information so quickly and effectively that they are not influenced by the match between the incoming sounds and the name of the wrong-category competitor

Dahan et al. (2000) tested for effects of syntactic constraints using grammatical gender in French, showing that, after hearing a gendered determiner (e.g. “un” (masc.) where the target object turns out to be “bouton” (masc., “button”)), people never fixated gender-inconsistent cohort competitors (e.g., “bouteille” (fem., “bottle”)) more than irrelevant distractors (e.g. “table” (fem., “table”)). This appears to be clear evidence that the gender cue stops listeners from activating the wrong-gender match.

Magnuson et al. (2008) dealt with the difficulty of representing both nouns and verbs in the visual world paradigm in a different way. They taught participants an artificial language composed of texture adjectives and shape nouns, all of which described referents in the visual world. In the test phase they modulated syntactic category expectations for adjectives vs. nouns by leveraging pragmatic constraints. When the four items in the display consisted of two items that shared one shape and two items that shared another shape, an adjective was necessary to uniquely specify any one of the items, and participants should therefore have expected an adjective after “Click on the...”. However, when the display contained four distinct objects, an adjective was not necessary, and participants should therefore have expected a noun after “Click on the...”. When the participant was expecting an adjective, Magnuson et al. (2008) showed

increased fixations to the shape whose adjective descriptor was a cohort competitor of the target adjective (e.g., to the "bupe tedu" when "bupʌ pibe" is the target), but shapes whose noun descriptor was a cohort competitor of the target adjective (e.g. the "tedɛ bupo") did not draw increased fixations. Similarly, when the participant was expecting only a noun, they showed increased fixations to the shape whose noun descriptor was a cohort competitor of the target noun, but shapes whose adjective descriptors were cohort competitors of the target noun did not draw increased fixations. This is exactly consistent with what Dahan et al. (2000) found, but for syntactic category rather than grammatical gender.

Magnuson et al. (2008) take their finding to be evidence for a continuous integration hypothesis. Under this hypothesis, top-down information is used as soon as it is available, which could result in a constraint on generation of candidates in some cases and a constraint on selection of candidates in other cases. However, as we will later discuss, there may be a potential strategy available to participants that could conceal competition from wrong-category candidates, if that competition were occurring. There is also the obvious caveat that these results occur in an artificial language in which participants had limited training.

More recently, Strand, Brown, Brown, and Berg (2018) employed a new visual world design that bypassed the imageability constraint with the inclusion of action pictures as referents for verbs. They presented visual displays with two pictures of objects (e.g. rug, bench) and two pictures of actions (e.g. run, pray), with auditorily presented sentences like "They thought about the rug", in addition to grammatically unconstraining sentences like "The word is rug." They also included trials in which the competitor was of the same syntactic category as the target. In line with Dahan et al. (2000) and Magnuson et al. (2008)'s findings, Strand et al. (2018) found phonological competition effects for syntactically compatible competitors, but not for wrong-

category competitors (i.e., no looks to “run” during “They thought about the rug”). However, as was the case for the Magnuson et al. (2008) design, potential strategies exist that could have prevented fixations to the wrong-category competitor even if it were activated. It is also worth noting that in an additional condition in which Strand et al. cross-spliced the onset of a recording of the competitor word into the target sentence, they did observe looks to the wrong-category competitor. This suggests that sufficiently strong bottom-up input can overwhelm top-down information in some circumstances.

Across three quite different studies, then, the implication from the visual world paradigm is consistent: category constraints can prevent phonological competition. Even more striking is the contradiction this poses with the gating and cross-modal priming results that preceded them. But as with gating, there are important design issues that need to be addressed in order for this interpretation of the results to stand. Furthermore, and independent of these issues, these designs cannot distinguish between inhibitory and facilitatory mechanisms for contextual constraints. The study we present in Section 3 is intended to resolve these problems.

1.2.4 MEG

A final piece of evidence in this literature comes from a MEG study (Gaston & Marantz, 2018). Rather than probe the status of a single wordform to ask whether it is or is not active in the cohort, as occurs in much of the behavioral work described above, this study asks about the overall status of the cohort as reflected in the processing of any single wordform. Gaston and Marantz examine correlations between neural activity in auditory cortex and phoneme-by-phoneme measures that are understood to reflect the probability distribution of the cohort. These measures—phoneme surprisal and cohort entropy—can be calculated for any phoneme in a word given a hypothesized cohort and the lexical frequencies of all items in that cohort. Phoneme

surprisal is the negative log of the conditional probability of a phoneme given the phonemes that preceded it, and cohort entropy reflects how much uncertainty there is in the probability distribution of items in the cohort. A cohort with just one high probability (high frequency) competitor will have low entropy, while a cohort with many equally likely competitors will have high entropy. Both phoneme surprisal and cohort entropy had previously been shown to correlate with neural activity during the processing of single words.

Gaston and Marantz (2018) presented words in small syntactically constraining contexts (e.g., “the clash persisted”, “to gleam brightly”) and constructed hypothesized cohorts constrained by the context in different ways. In one version of the constraint, wordforms that cannot be used in the syntactic category that fits with the context are dropped from the cohort. In a second version, this same constraint on cohort membership is applied, but the frequencies of remaining cohort members are adjusted so that they only reflect the frequency of that wordform in the syntactic category allowed by the context. Using those constrained cohorts to calculate new surprisal and entropy values, they then tested correlation of neural activity with constrained and unconstrained surprisal and entropy.

Gaston and Marantz (2018) found that the earliest detectable effects of cohort dynamics were for the first constrained version of phoneme surprisal, in which the only adjustment is that wordforms that cannot be used in the syntactic category required by the context are removed from the cohort. However, a major caveat for the reported result is that the category-constrained cohort distributions were constructed under the assumption that a category constraint could only work via inhibition, such that items of the wrong category are not in the cohort. The study is not designed to detect the possibility that the category constraint in fact works via facilitation, such that items of the correct category are simply activated more strongly, or have higher probability,

than items of the wrong category. Evidence for this type of constraint would require recomputing the category-constrained cohort measures from a new hypothesized cohort, and re-evaluating for correlation with the neural data.

Support for the possibility of facilitation comes from the fact that, in addition to the early restricted cohort effect, Gaston and Marantz (2018) also found effects of phoneme surprisal calculated from the unconstrained cohort. This could be interpreted as evidence for an unconstrained cohort that is active simultaneously with the constrained cohort. However, if the constraint in fact operates such that good fits are facilitated rather than poor fits inhibited, the constrained and unconstrained cohorts as constructed in this study would each be capturing one aspect of the true cohort, and therefore might each display some degree of correlation.

Because the surprisal effects that Gaston and Marantz report did not begin until the second phoneme, we can only conclude from this data that as early as there is evidence for the cohort, that evidence is consistent with restriction. Brodbeck, Hong, and Simon (2018) also find a lack of neural evidence for active cohort dynamics before the second phoneme. Because visual world studies examining phonological cohort competition generally use stimuli in which both of the first two phonemes are overlapping, current evidence across methods could actually be consistent with the cohort being generated at the second phoneme (or, e.g., at the first vowel), which would make it more likely that Gaston and Marantz (2018)'s findings reflect an immediate constraint.

1.2.5 Summary

In conclusion, although much work has investigated the question of whether syntactic context can constrain the generation of lexical candidates, the answer is still unclear because of conflicts in the results across different methods. Gating and cross-modal priming indicate (with

some caveats) exhaustive initial access of lexical candidates despite syntactically constraining context, while the visual world paradigm indicates immediate context effects such that wrong-category candidates never compete. Evidence from neural activity as measured by MEG cannot rule out early exhaustive access but is potentially consistent with an immediate (facilitatory) constraint.

In this paper, we try to resolve this conflict by asking not simply whether there is wrong-category cohort competition, but *how* it is that auditory word recognition could be impacted by information from syntactic context, and how such impacts can be measured. This requires close examination of the measures we use to investigate these processes, the potential mechanisms for interaction between different sources of information, and the mechanics of both auditory word recognition and syntactic prediction.

1.3 Broader considerations for syntactic context

The syntactic constraint constitutes an unusually clear testing ground for the interaction of top-down and bottom-up information. Testing whether or not a contextual constraint seems to be in place requires a clear hypothesis regarding the effects that constraint should have on the activation levels of each wordform in a set of lexical candidates. In other words, the constraint must be definable in such a way that specific properties of the wordforms can make them more or less compatible with it. How easy or hard it is to spell out such a hypothesis varies considerably among the set of constraints that may apply when word recognition occurs in a sentence. In the case of, for example, semantic or pragmatic constraints, it may be difficult to determine exactly which wordforms should be more or less likely, and on what basis. In the case of syntactic constraints, it is fairly clear that given a certain syntactic parse for the sentence, some wordforms are grammatical in the context and others are not. This makes investigating the

mechanism for the syntactic constraint more tractable, and progress in this area should then aid progress toward understanding more complex constraints.

It is important to acknowledge, however, that even in the relatively simple syntactic case the manner in which the syntactic category of lexical items is represented is not completely straight-forward, with implications for how we think about the constraint. Under a Distributed Morphology view of the lexicon (Halle & Marantz, 1993), root morphemes do not belong to syntactic categories or have features that designate them as such. Instead, category-less roots combine, syntactically, with categorizing affixes that allow them to then operate in the larger syntactic structure as e.g. nouns. Syntactic category, then, is not a stored part of the lexical representation. Vigliocco et al. (2011) draw a related distinction between so-called "combinatorial" and "lexicalist" views of syntactic category, largely with reference to the question of whether nouns and verbs are represented and processed differently in the brain.

For many psycholinguistic questions, it seems unlikely to matter whether syntactic category is a stored feature or is derived. This includes our question, on whether and how the set of wordforms under consideration is influenced by restrictions from the syntactic context. If some roots never occur in some categories, for example, a constraint that is sensitive to these probabilities will be difficult to distinguish from one that operates on stored category features. However, sensitivity to the fact that a root has not been observed to occur in a certain category is not equivalent to a prohibition on it doing so. Thus, we consider the Distributed Morphology view to be more explicitly compatible with the possibility that no item is completely excluded from consideration during word recognition on the basis of syntactic category (i.e., that there is not an inhibitory constraint). Though the terminology used in this paper may at times seem to

indicate an assumption of stored syntactic category information for wordforms, referring to items that can or cannot be nouns, we do not intend for this to be a theoretical commitment.

Finally, many questions remain as to how exactly the syntactic parse arises, which then leads to an expectation for specific categories, but the nature of any syntactic constraints on word recognition will figure importantly in our understanding of the timing and predictive nature of syntactic structure-building. Eventually, we should aim to integrate evidence from these two different problems: the structures and algorithms that give rise to syntactic prediction, and the effects of those predictions on incoming input.

2 Simulations on the role of the response candidate set

In the Introduction, we described a conflict in the results from different methods used to investigate syntactic context effects on cohort competition. Some studies show initial competition from wrong-category items, and others do not. The presence or absence of wrong-category competition is crucial for our understanding of the mechanism of a syntactic category constraint. Therefore, we next explore whether one contributing explanation for the conflict could be the difference between *forced-choice* and *free-choice* tasks. The visual world paradigm is an example of a forced-choice task because its dependent measure is the proportion of fixations to each of a small number of referents on screen. Participants have a limited set of options for responding to the auditory input. In a free-choice task like gating, participants are free to respond to the auditory input with any word from the lexicon. Could the appearance of lexical competition be influenced by the nature of the set of response candidates made available in these different kinds of tasks? This would be important for both the interpretation of prior findings and the design of new studies.

Although our focus in this paper is lexical competition as influenced by top-down syntactic information, in this section we will describe simulations on the role of the response candidate set when lexical competition is driven by bottom-up auditory input alone. From a practical standpoint, this is because the model we use does not allow for the incorporation of syntactic context information. However, we also believe both that top-down effects cannot be understood without reference to bottom-up competition and that any influence of the response candidate set on bottom-up competition dynamics is likely to apply in some analogous fashion to top-down effects, making these simulations a necessary starting point.

Our simulations lead to two main conclusions. The first is that simulated bottom-up lexical competition effects can indeed be observed earlier in forced-choice relative to free-choice scenarios. Future work will be necessary to determine whether the same holds true for top-down effects on lexical competition; if so, this could help explain why, for example, there appears to be an earlier context effect in the visual world paradigm than in gating. Our second conclusion arises from investigating why task-related timing differences are predicted by the model. We see that what matters is in fact not the size of the response candidate set but its composition. Even if an item is activated, competition may be obscured because of the behavior of the other items in the response candidate set. This has important consequences for the design and interpretation of forced-choice experiments, such as the VWP. In Section 3, we explain how we leverage this principle in a new visual world design that increases the detectability of wrong-category competition.

2.1 TRACE model of speech perception

TRACE (McClelland & Elman, 1986) is a well-known connectionist model of speech perception, designed to simulate lexical and phoneme-level activation during the processing of

spoken words. It has been reimplemented in Java as jTRACE (Strauss et al., 2007), which is freely available and designed to be accessible to researchers wishing to evaluate their own hypotheses. In this section we describe key properties of the model that are relevant for the cross-method discrepancies described previously, and that influence our interpretation of the simulations to follow. Particularly important is the model's distinction between activation/response strength and response probability.

As an interactive activation model, TRACE has three layers of processing units (features, phonemes, and words) which are both inter- and intra-connected. By default, TRACE has feed-forward excitatory connections from the feature to the phoneme and the phoneme to the word layers, and feed-back excitatory connections from the word to the phoneme layer. There are also inhibitory connections among units within each layer, but there are no between-layer inhibitory connections. Given specific phonological input, TRACE can simulate lexical activation for all of the words in its lexicon. Simulated lexical activation can then be converted to response probability, or the probability that a given item is e.g. selected, produced, or fixated in an experimental task. Allopenna et al. (1998) were the first to show that TRACE's simulated response probabilities yield a close correspondence with fixation probabilities observed in the visual world paradigm. To compute response probability for a given word, activation values from TRACE are first converted to response strength, per Equation 1. S is response strength, a is activation, and k is a constant.

Equation 1: $S = e^{ka}$

This conversion, according to McClelland and Elman (1986), is intended to ensure that all values are positive, and to give more weight to stronger activations. The default for k in

jTRACE is 7, and this default is used in all simulations reported here. Response probability (L) can subsequently be derived from response strength using the Luce choice rule, as in Equation 2. In this step, response strength S for an item i is divided by the sum of response strengths for all j items in the set of possibilities for making a response (what we refer to as the response candidate set).

$$\text{Equation 2: } L_i = \frac{s_i}{\sum s_j}$$

The proportion in Equation 2 can therefore be taken to indicate the probability of choosing that item out of the alternatives in the response set. Originally, this was used by McClelland and Elman (1986) to simulate word identification responses. In word identification, the response set should be the whole lexicon, and we assume this to also be the case in a task like gating. We refer to these as *free-choice* tasks. In contrast, simulations of the visual world restrict that set of possibilities for making a response to the (usually) four items that are on screen. We refer to these as *forced-choice* tasks.

It is important to note that the response strength for a given word in the numerator of Equation 2 will always be the same regardless of the size of the set of response candidates it occurs in, as it is a simple function of activation level. This reflects an assumption built into the model that underlying lexical activation is unaffected by task. A difference in response probability, in contrast, will arise in forced as compared to free-choice scenarios because of differences in the summed response strength in the denominator, which varies depending on whether one is summing over four items or the entire lexicon. For example, if all items in a 100-word lexicon have the same activation/response strength, response probability for each item in a four-item forced choice should be 25%, while in a free choice from the whole 100-word lexicon

it should be 1%. This is important because when we ask about syntactic context effects, what we are really asking about is the underlying activation level, but the data that we use to evaluate the question are typically filtered through the observable measure of response probability.

2.2 Simulation details

We used jTRACE to investigate whether the dynamics of a forced-choice task among four items are such that the influence of bottom-up phonological information can be observed more quickly than is possible in a free-choice scenario. Specifically, we asked whether the manifestation of cohort competition in response probability was affected by the size of the response candidate set. This was motivated by conflict in the previous literature on syntactic constraints on cohort competition. An effect of the response candidate set on the timing of bottom-up cohort effects (as we explored in these simulations) would suggest the possibility of an analogous influence on the timing of top-down cohort effects (as are at issue in the previous literature).

In order to ensure that direct comparisons with previous work using TRACE and jTRACE were warranted, we chose to replicate and then extend simulations from Dahan, Magnuson, and Tanenhaus (2001) that were executed with the original TRACE model and then later validated by Strauss et al. (2007) with jTRACE. Strauss et al. provide a template that includes the stimuli used in Dahan et al. (2001)'s original experiment, as well as the lexicon used in Dahan et al.'s TRACE simulation. This lexicon was based on the 235-word lexicon made available with the original TRACE implementation, with some additions. We simulated each of the 17 trials and then averaged their results for comparison with the averaged human data and averaged simulations presented by Dahan et al. (2001). We then repeated these simulations in a free-choice scenario.

Each trial included a target, a high-frequency and a low-frequency cohort competitor each with the same onset as the target, and a distractor whose onset did not overlap (e.g. bench, bed, bell, lobster). Forced-choice trials evaluated response probability over time for a response candidate set consisting of only these four items, while the set for free-choice trials was the entire lexicon of 301 words, though we only looked at response probability for these four items. In this type of design, at some short delay following the onset of the target, we generally expect activation and response probability to increase for the target and competitors and to decrease for the distractor. Later in the trial, at some short delay following the phoneme that differentiates the competitors from the target, we expect response probability for the competitors to decrease and for the target to continue to increase.

Therefore, our dependent measure was the cycle at which the response probability of the distractor begins to decrease without later recovery, and then the cycle at which the response probability for the high-frequency competitor begins to decrease without later recovery. We consider the distractor's point of decrease to be the first time at which bottom-up information observably shapes the cohort, and this is our primary concern. However, the high-frequency competitor's point of decrease is also useful to consider because it reflects the speed of an update to the cohort without any special considerations that the onset of the word might bring.

Though not directly related to our question, there are parameter settings in jTRACE that were manipulated by Dahan et al. (2001) and which we are therefore also concerned with. Dahan et al. (2001) compared simulations in which lexical frequency information was added to TRACE by 1) proportionately weighting the resting activation levels of word units, or 2) proportionately weighting the connection strengths between phoneme and word units, or 3) allowing frequency to influence a post-activation decision stage. All three options are available for simulation in

jTRACE. Dahan et al. (2001) found that resting-level and connection-weight implementations of frequency both produce simulated fixation probabilities with extremely good fit to the human data, while frequency in the post-activation decision rule performs less well. Therefore, we ran simulations both ways, one set with the resting-level implementation and one set with the connection-weight implementation[1].

The other parameter for consideration is the baseline resting activation default set to -0.3 by Dahan et al. (2001), rather than the -0.01 standard for TRACE. Dahan et al.'s setting ensured that the maximum to which frequency could scale a resting activation level was still below zero, guaranteeing stable effects of the frequency biases. We ran simulations with both the standard default (-0.01) and Dahan et al. (2001)'s -0.3 in order to clarify the implications of their choice.

2.3 Simulation results

For each simulation, the cycle at which the response probability of the distractor began to decrease is reported in **Table 1**, and the cycle at which the response probability of the high-frequency competitor began to decrease is reported in **Table 2**. Each cycle in a TRACE simulation is argued by Dahan et al. (2001) to correspond to roughly 12 ms. in human behavioral data. The interacting comparisons of free and forced-choice tasks, resting-level (RL) and connection-weight (CW) frequency implementations, and standard and reduced baseline resting

[1] The difference between the two more successful implementations is that, in the simulations, frequency implemented in the resting activation levels produces baseline frequency effects prior to the onset of bottom-up effects that occur in response to the target input. Statistically significant baseline frequency effects are not observed in the human data, leading Dahan et al. (2001) to conclude in favor of a connection-weights implementation of frequency. However, we cannot be sure that there is no frequency effect on lexical activation levels during the baseline period simply because there is no observed difference in fixation proportions. This is because the correspondence between lexical activation and fixations is understood to be driven by the task (in this instance, to click on the picture matching what is being heard). In the absence of linguistic input and therefore in the absence of a task, the linking hypothesis between resting activation levels and fixations is less obvious. Thus, we do not assume that the human fixation data during the baseline period is less consistent with the resting-level than the connection-weight implementations of frequency.

activation defaults resulted in a 2x2x2 manipulation of Dahan et al. (2001)'s basic high/low frequency design.

Table 1. *Cycle at which response probability of distractor begins to decrease (indexing earliest effect of phonological input).*

Default baseline for resting activation:	*-0.01*		*-0.3*	
	Resting-level frequency	Connection-weight frequency	Resting-level frequency	Connection-weight frequency
Forced-choice task	10	14	14	14
Free-choice task	16	14	14	14

Table 2. *Cycle at which response probability of high-frequency cohort competitor begins to decrease.*

Default baseline for resting activation:	*-0.01*		*-0.3*	
	Resting-level frequency	Connection-weight frequency	Resting-level frequency	Connection-weight frequency
Forced-choice task	34	35	30	33
Free-choice task	48	48	41	43

The conditions with the parameter settings used by Dahan et al. (2001) (-0.3 default baseline, forced choice) matched those reported results. With resting-level frequency and the (standard) -0.01 default baseline, we observed that the point at which the distractor's response probability began to decrease was 6 cycles (72 ms.) earlier in the forced-choice task than the free-choice task. In the other three parameter combinations, the distractor diverged at cycle 14 for both free choice and forced choice. In these cases, activation was static and did not change from its starting point until cycle 14. This suggests that there may be a constraint on activation changes in the model that is circumvented when the baseline is very close to zero and the resting-level frequency implementation allows for lateral inhibition. For the high-frequency competitor,

the point of decrease was earlier for forced choice relative to free choice in all parameter combinations, by 10-14 cycles (~120-168 ms.). The consistency of this pattern for the high-frequency competitor lends support to the possibility that task differences for the distractor are obscured by idiosyncrasies of trial onset. For our purposes, it is instructive that there is any parameter setting under which such differences can arise. However, the parameter differences we observed deserve follow-up in future work.

2.4 Set size vs. set composition

In this section, we examine how these timing differences arise between free choice and forced choise. Our findings make clear that in fact it is the composition of the response candidate set, not its size, that is relevant for competition dynamics. We use the timing of response probability changes for the distractor (as shown in **Table 1**) to illustrate our point, but identical logic applies to the timing of the response probability changes for the high-frequency competitor (**Table 2**).

It is important first to understand the time course of changes in underlying response strength (transformed activation) during a trial. Recall that response strength for a given item does not vary by task in the TRACE architecture in the way that response probability does, nor is it affected by the other items in the response candidate set. **Figure 1** shows response strength over the course of a trial in our simulations (using the parameter settings that resulted in a free/forced-choice difference for the distractor). After the first cycle in **Figure 1**, we see that response strength increases for all four items until cycle 24 (288 ms.), when it starts to decrease for the distractor. Response strength can increase for all items from the very beginning of the trial because, first, the resting level frequency implementation means that the different items start with different baseline activation levels. Lateral inhibition then leads higher-frequency items in

the lexicon to increase in activation while lower-frequency items decrease, even without bottom-up input (see Dahan et al. (2001) for further discussion). We believe that the reason all four are increasing here is that they are relatively higher frequency in the lexicon. In any case, the decrease in response probability that we observe for the distractor (as shown in **Table 1**) is much earlier than cycle 24, even though response strength is still increasing**.**

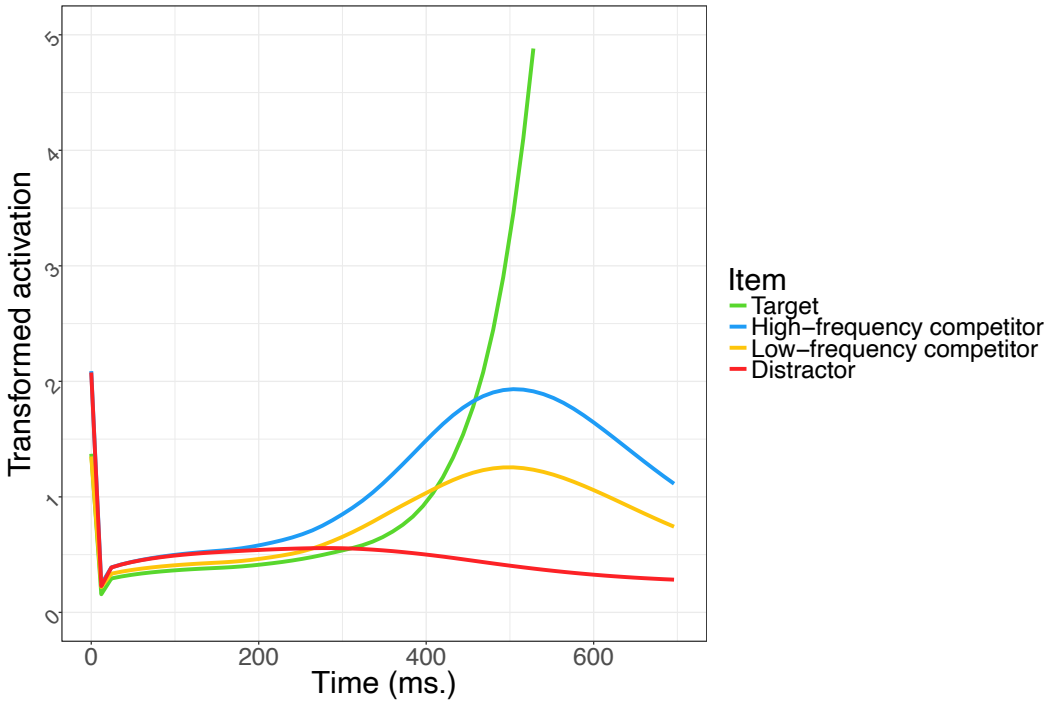

Figure 1. *The time course of underlying response strength (transformed activation) in each condition of the first simulation using the resting level frequency implementation and standard default resting baseline.*

Next, we examine how changes in response strength over time translate to changes in response probability, since response probability is our dependent measure. This relationship rests on the relative rate of increase in response strength from cycle to cycle for the four items in either task. Response probability for the distractor is the response strength of the distractor divided by the summed response strength of all of the items in the response candidate set, which

is either the forced-choice set of four or the entire free-choice lexicon. From one cycle to the next, if the percent increase in response strength is the same for the distractor (the numerator) and the response candidate set "pool" (the denominator), the response probability for the distractor will stay the same. If the distractor's response strength increases proportionally less than the response candidate set pool, response probability for the distractor will instead decrease.

Our free vs. forced-choice timing differences arise because the summed response strengths for the forced-choice and free-choice response candidate set increase at different rates. In the forced-choice trials, there are three items consistent with the first phoneme (the target and two competitors), and one distractor item that is not. In the free-choice trials, there are more than three items consistent with the first phoneme, but many more items that are not. Therefore, when the first phoneme is perceived and all words that are consistent with it experience a boost in activation, the percent increase of the pool is higher in the forced-choice scenario than the free-choice scenario, because a higher proportion of items in the forced-choice set are experiencing a boost. This means that the percent increase in response strength for the forced-choice pool can surpass that of the distractor more quickly, causing the response probability for the distractor to decrease earlier, as illustrated in **Figure 2**.

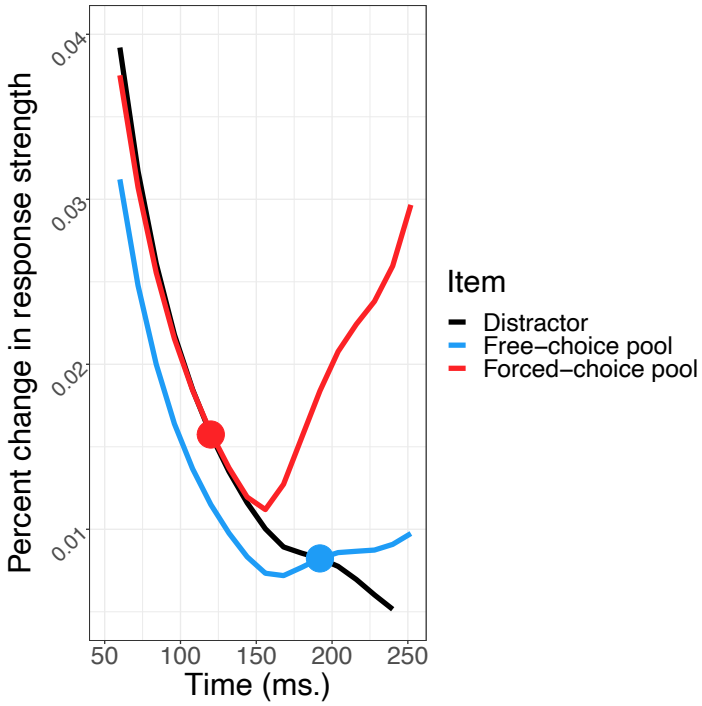

***Figure* 2.** *The time course of percent change in response strength for the distractor, free-choice pool (entire lexicon), and forced-choice pool (target, two cohort competitors, distractor) in the resting level frequency/standard default resting baseline condition. Dots indicate the points at which the percent change in response strength for the forced-choice or free-choice pool surpasses the percent change in response strength for the distractor.*

We conclude that the earlier changes in response probability in the forced-choice scenario are not due strictly to the smaller size of the response candidate set: they are due to the response candidate set being disproportionately composed of items consistent with the input. This means that if the forced-choice candidate set and the overall lexicon had the same distribution of items consistent or not consistent with the input at each position, the timing of changes in response probability would not be expected to differ in free and forced-choice tasks.

2.5 Discussion

The aim of these simulations was to determine whether changes in response probability in response to bottom-up phonological input occurred earlier in forced-choice tasks, as a first

step toward explaining cross-method differences in the timing of top-down syntactic context effects. We did find evidence in favor of the idea that forced-choice tasks can demonstrate changes in cohort competition dynamics earlier than free-choice tasks. Looking at how quickly response probability began to decrease for a distractor item whose first phoneme did not match the input, we saw that under a parameter setting that allowed activation to vary from the onset of the trial, this indicator occurred six cycles (~72 ms) earlier for a forced-choice relative to a free-choice scenario. Looking at how quickly response probability began to decrease for a cohort competitor that ceased to match the input at the third phoneme, we saw that across all parameter settings this indicator occurred 120-168 ms earlier in forced-choice relative to free-choice scenarios.

Describing how these timing differences arise, we noted that the relevant difference between the forced-choice and free-choice scenarios that we compared was the composition of the response candidate sets rather than their size (though the two are related). This means that in addition to the choice of task, stimulus selection *within* the visual world paradigm should be an important determinant of the observed timing of changes in cohort competition. Changes in the response probability of a critical item must then be interpreted in the context of any activation changes expected to occur for other items in the response candidate set. For example, it may take longer for a critical item's response probability to begin changing after its activation has begun to change in a trial where no other item in the set is a full or partial match to the input, as compared to a trial in which all other items in the set are full or partial matches to the input. This principle also highlights that response probability for a critical item may decrease even though its activation is not changing, if the activation of another item in the candidate set is increasing. Conversely, response probability for a critical item may stay constant even if its activation is

increasing, if the activation of other items in the candidate set is also increasing. Therefore, the clearest possible inferences about a critical item's activation occur when the activation of other items in the candidate set should not be changing. If any other items' activation is expected to change due to full or partial matches with the input, we can have little certainty in attributing a change in the critical item's response probability to a change in its own activation. In the next section, we incorporate this understanding of the importance of set composition into a new design in the visual world that prioritizes making wrong-category competition detectable.

Our primary question in this paper is about the activation of category-incompatible lexical candidates during word recognition, an issue surrounded by conflict in the literature. These simulations have not addressed that question directly, but they have provided a new potential framework for understanding cross-method differences, and they have highlighted an important design consideration. Fully testing jTRACE's predictions for human data will require use of the same stimuli and maximally similar set-ups across methods, and direct comparisons of the timing of top-down and bottom-up effects in each dataset. If the simulated cohort dynamics are correct, and if set composition is as important for top-down changes in cohort competition as it is suggested to be for bottom-up changes, one potential factor in the observed conflict is that, while the same underlying activation changes occur in all tasks, visual world studies might allow top-down effects to become observable in behavior exceptionally quickly. This could be expected to occur if the proportion of items in the forced-choice candidate set that are impacted by contextual information is higher than that proportion in the free-choice set.

3 A visual world study on facilitation vs. inhibition as mechanisms for the syntactic constraint on cohort competition

In the previous section, we used simulations to better understand the importance of the response candidate set in designing and interpreting behavioral studies that measure cohort competition. Those findings may help resolve conflicts in the literature related to syntactic context effects on the cohort competition process. In this section, we report an experiment in the visual world paradigm that is set up to distinguish the effects of a facilitatory and an inhibitory syntactic category constraint, with a design informed by our simulations. The mechanism for the constraint is an issue that has not previously been addressed, and is an important step toward a full characterization of auditory word recognition.

Our study examines fixations to pictures in the visual world paradigm whose names have the same onset as the word being spoken, and either are or are not consistent with the syntactic context. A lack of fixations to wrong-category items in previous visual world results suggests that wrong category words are not activated by auditory input, hence lending support to an inhibitory mechanism. However, we argue that previous designs actually made such fixations difficult or impossible to detect. If, in a more sensitive design, we find evidence that the constraint is in fact facilitatory, this would present another potential factor to aid in explaining observed conflict in the prior literature. Such an explanation would not be in conflict with the timing-related explanation we explored in our simulations; both are potentially at work. If top-down effects do manifest more quickly in the visual world than in other methods, this should be true regardless of the mechanism for the syntactic constraint.

Our visual world design has several key features. In contrast to previous visual world experiments on syntactic constraint, one notable design choice we made in order to increase the

sensitivity of our measure was that, while participants are looking for an item related to what they are hearing on every trial, critical trials do not actually include a target. We also did not include distractor items whose activation could be expected to increase for any reason. These choices were made to ensure that the cohort competitor would be the only item whose activation should change due to the auditory input, as this specific activation change was the crucial object of study. For example, when the auditory input was '*He chose the battleship*', the display would include an image of cohort competitor *balcony*, but no battleship, and no other items that matched the input. As discussed in Section 2, using response probability as a dependent measure for making inferences about changes in lexical activation requires extreme care in causal attribution. The response probability for any given item is jointly determined by its own activation and by the activations of the other members of its candidate set. Therefore, a change in response probability can only be unambiguously attributed to a given item if it is the only item whose activation could be changing due to the input. Huettig and McQueen (2007) also argue that target-less trials may serve as a more sensitive measure for evaluating cohort competition because looks are not being split between the target and the competitor (see also Brock & Nation (2014)).

The other difficult design choice we were faced with was how best to implement a syntactic context manipulation involving noun and verb contexts, given that there are inherent differences in the imageability of verbs and concrete nouns. One approach in past work has been to compare fixations to verb referents (action pictures) and noun referents (object pictures) in, for example, a verb context. However, this means that visual correlates of syntactic category could drive fixations independent of the unfolding of the target word (an issue we consider further in the Discussion section). To avoid this possibility, our displays only contained pictures of objects,

and we compared fixations to the same object in noun- vs. verb-constraining auditory contexts. For example, in a trial whose critical competitor picture was of a balcony, we asked whether looks to that balcony would increase relative to baseline in a verb context ("to bask") in the same way that they were expected to in a noun context ("the battleship"). Filler trials ensure that objects are potential referents for verbs across the experiment.

These two design choices help ensure that the patterns in fixation probability that we observe in our data are the clearest possible reflection of lexical activation that the visual world paradigm allows.

3.1 Predictions

Our predictions with respect to the mechanism for the category constraint are as follows. We assume in all cases that bottom-up input leads to an increase in activation over baseline for the competitor but not the distractors.

In the verb context ("to bask"), a category constraint acting via total inhibition of wrong-category candidates should stop activation of wrong-category, noun-only "balcony" as well as the other (wrong-category, noun-only) distractors, such that fixations do not increase relative to a baseline for any of them. In contrast, a category constraint acting via facilitation of correct-category, verb-compatible candidates in the verb context should not affect the activation of wrong-category, noun-only "balcony" or the distractors, meaning that fixations should still increase relative to a baseline.

However, in the noun context ("the battleship"), an inhibitory category constraint for wrong-category candidates would be irrelevant for the four correct-category, noun-only pictures, leading to the typical increase in fixations for the competitor (balcony). A facilitatory category constraint for correct-category, noun-compatible candidates, in the noun context, would increase

the activation of all four noun-only pictures equally and so should also not cause any additional changes in response probability other than the change caused by the phonological input. Thus, we predict normal cohort competition for *balcony* in the noun context condition, under either type of constraint, while in the verb context such competition should only occur under facilitation.

The facilitatory constraint predicts a cohort competition effect (an increase in fixations for the competitor relative to baseline) in both noun and verb contexts, and no difference in this competition effect between the contexts. An inhibitory constraint predicts a cohort competition effect in the noun context but not the verb context, and thus a difference in this competition effect between the contexts. Because the prediction for a facilitatory constraint is indistinguishable from a null result in which there is no difference for competition between the two contexts, we included a second contrast, to be described below, in which fixations to the competitor should differ between the two contexts regardless of the constraint being inhibitory or facilitatory. This expected difference would also yield information about the timing of the constraint.

We acknowledge that in naturally occurring language, it is not the case that noun-only items are fully ruled out following infinitival “to” or that verb-only items are fully ruled out following “the.” This is true, however, in the repetitive structure of our stimuli, and we believe it to be a reasonable heuristic in our design. We leave a finer-grained accounting of probabilistic syntactic prediction to future work.

3.2 Methods

3.2.1 Assumptions about the visual world paradigm

Our study relies on basic phonological cohort competition effects, in which fixations to a phonological onset competitor typically increase in the first 200 to 400 ms after target word onset (cf. Allopenna et al. (1998), and many others following). The specific effects expected under any given design in the visual world paradigm depend on one's assumptions about the way in which stimulus-driven fixations arise. Magnuson (2019) provides a thorough account of different potential linking hypotheses for the visual world, and their relative merits. In this section, we explicitly outline our own assumptions about the paradigm, aligned most closely with Magnuson's second linking hypothesis.

1. Participants have been instructed that they will be completing a task which will require them to have perceived and processed the pictures presented to them visually. This task need not be as obvious as to "Look at the…" or "Click on the…" picture; more passive engagement is also sufficient. Huettig and McQueen (2007) show that phonological competition arises even without any task at all, though we prefer assurance of at least a basic level of attention.
2. Pictures are presented to the participant at least 1000 ms before auditory input begins.
3. Participants scan the visual scene, and perceive the pictures. Lexical items associated with the pictures are activated, and then phonological representations corresponding to the lexical items are activated. This set of phonological representations is actively maintained or focused in short-term memory.
4. Auditory input begins. Perception of the first phoneme of a word leads to the activation of all phonological representations that are consistent with the input. Each subsequent phoneme further facilitates representations that are still consistent with the input.

Activation of phonological representations eventually leads to activation of corresponding lexical representations, but the exact timing of this process is not relevant for the current considerations.

5. Per Huettig and McQueen (2007)'s "phonological hypothesis," phonological representations have been retrieved or activated both on the basis of visual (picture) information as well as on the basis of auditory (speech) information. Attention is shifted to a picture if there is a match between the representations activated via the two routes.
6. Scanning of the visual scene continues for the duration of the trial. The probability of fixating on any given picture at any given time point is determined by its relative activation levels. TRACE (McClelland & Elman, 1986) proposes the Luce Choice Rule for conversion from activation to fixation probability, where activation (after an exponential transformation) is divided by the summed activation for the set of items in the visual display.
7. If the auditory input changes the activation levels of a phonological representation that matches with one of the visually activated representations, these changes in activation lead to changes in fixation proportion. It is estimated to take roughly 200 ms for a saccade to be planned and executed (Rayner et al., 1983), but cancellation of planned saccades may be somewhat faster. We expect changes in auditory input to manifest as changes in fixation proportion with at least a 100 ms lag.

An important component of our assumed linking hypothesis is that, given sufficient preview time prior to the onset of auditory input, participants implicitly name the visually presented pictures. As soon as the subsequent auditory input also activates phonological representations, attempted matching between the two inputs can occur. A common alternative

hypothesis is that the pictures are not implicitly named. Instead, only conceptual representations associated with the pictures are activated. Auditory input activates phonological, then lexical, then conceptual and possibly visual representations, and fixations are driven by matches with the conceptual representations activated by the pictures or simply their visual features. This hypothesis is central to the assumption of independence between language processing and visual processing made by Allopenna et al. (1998).

Because we assume that attempted matching occurs at every level of representation of the visual and auditory inputs, our implicit naming hypothesis does not rule out that conceptual/visual matching also occurs once these levels of representation have been activated from the auditory input. However, if phonological representations have already been activated by the pictures when auditory input begins, the first match that can be made is between the visually and auditorily activated phonological representations.

We assume implicit naming is occurring because of Huettig and McQueen (2007)'s finding that cohort competition effects are eliminated when there is not sufficient time to preview the visual scene before the onset of auditory input. If typical cohort competition effects arise via conceptual/visual matching rather than phonological matching, the amount of preview time provided should matter very little. The visual or conceptual features of the pictures are readily available when the scene is initially scanned, and there is no additional time needed for activation of alternative levels of representation. However, if advance visual input is required for cohort competition effects to arise, it must be that time is required for the pictures to activate their corresponding phonological representations, so that the match at that level can drive fixations. If this has not occurred in advance, the phonological competitor may no longer be consistent with the word being heard by the time attempted matching at the conceptual/visual

level occurs. This explains why Huettig and McQueen still found semantic and shape competition effects in the absence of preview (and why such effects occur at all).

A more extensive, incremental manipulation of preview time would help differentiate these two hypotheses, and could be used to shed light on the timing of the various sub-processes involved in auditory word recognition. Magnuson (2019) argues that the implicit naming hypothesis is implausible because it is unlikely that people activate the names for every object and visual scene that they encounter in normal life. If it is the case that they don't do so in normal life but they do engage in implicit naming in the visual world paradigm due to task demands, Magnuson notes the risk that language processing as measured in the visual world is then highly distorted. However, it is not known whether or to what extent the activation of phonological representations from the pictures actually distorts the activation of phonological representations from the auditory input. Implicit naming may enable a convenient window onto phonological representations without actually influencing them.

3.2.2 Design details

3.2.2.1 Noun-only competitors

For our first primary contrast, we used 30 sets of four noun-only pictures (i.e., pictures whose names could only be used as nouns, according to the SUBTLEX-US corpus (Brysbaert & New, 2009)). See Appendix for a full list of stimuli (also available for download on OSF). Within each set, no picture names shared an onset. For example: balcony, moustache, curtain, wheelbarrow. The four pictures were presented in the corners of a 3x3 grid. We counter-balanced the auditory context they were presented with such that half of the displays occurred with a noun context for Group A participants and a verb context for Group B participants, and half of the displays occurred with the opposite category. For example, a sentence containing a noun-only

auditory target (battleship) would be “He chose the battleship for his birthday.” One picture (here, balcony) was a phonological onset competitor of the auditory target, as determined by the CMU pronouncing dictionary (Weide, 1994). The remaining pictures were considered distractors. Note that in typical visual world designs, in which the word being heard is also pictured in the display, this word is called the “target”. In our design, the word being heard is not pictured in the display. We continue referring to this word that is heard as the “auditory target”. The picture in the display whose name has an overlapping onset with the auditory target is referred to as the competitor.

We measured the proportion of fixations to each of the four pictures following the onset of “battleship.” Looks to the balcony were expected to increase relative to a baseline, in a time window roughly 200 to 400 ms after the onset. The task for participants was to indicate via button press after each trial whether they had seen anything on the screen related to what they were hearing. The verb context version of each item used the same visual grid and contained the same pronoun and main verb and a target with the same two-phoneme onset. For example: “He chose to bask in the sun.” The noun and verb auditory targets were not matched for length. We did not employ a fully within-subjects design here because we did not want the same participant to see identical grids twice, which we expected would make competitor status more predictable.

In many visual world designs, the same picture occurs as target, competitor, and/or distractor (potentially counter-balanced across participants) to ensure there are no differences between these pictures (other than presence or absence of onset overlap) that could explain any observed effects. Our crucial comparison involves identical visual stimuli occurring in different auditory contexts, so we did not have this concern, and thus we did not fully rotate pictures between the competitor and distractor roles. However, because of the limited size of the set of

pictures that met our criteria (described below), we did have to repeat individual pictures. Competitors were each used twice as distractors in other grids. Half of the competitors also served (twice) as the targets in filler trials with noun context sentences. This ensured that repetition of a picture did not make its condition (as competitor or distractor) predictable. Because of the pictures that appeared three times in the experiment as critical items (once as competitor, twice as filler target), we broke up the presentation list into three blocks, and allowed only one repetition per block. The order of these three blocks was counter-balanced (interacting with the counter-balancing of contexts for a total of 24 lists), and the order of trials within these blocks was pseudo-randomized. Each block was then broken in half for presentation of 20 trials at a time. Participants were able to take a break after each half-block of 20 trials. Distractor items that never appeared as competitors/targets were repeated five times across the course of the experiment.

3.2.2.2 Noun/verb-ambiguous competitors

As described above, a facilitatory implementation of the category constraint does not make different predictions for the noun and verb contexts when noun-only competitors are used, and would be indistinguishable from a scenario in which there is no category constraint. In the noun context, we would expect the constraint to cause the same proportional boost in activation to occur for all four of the noun-only word-forms associated with the pictures, leading to no observable difference in response probability due to context. This would be indistinguishable from the verb context in which the constraint does not affect any of the items (because it only boosts verbs) and so no response probability difference occurs.

To aid in interpreting such a result, we included a second contrast aimed at distinguishing a facilitatory constraint and a lack of constraint, using competitors that were noun/verb category

ambiguous (e.g. “soap”). Given some assumptions about the implementation of multiple-category lexical items, both the inhibition and the facilitation account would predict a difference due to context in this second contrast.

As above, four pictures were presented in the corners of a 3x3 grid. A competitor picture with a noun/verb ambiguous name (e.g. a bar of soap) was included in addition to three other distractor pictures with noun-only names. All category designations were according to the SUBTLEX-US part-of-speech tags. In most trials (27 out of 30), the frequency of the target name was biased towards noun usage, according to SUBTLEX-US. Noun bias ranged from 69.98% to 99.84%, with a mean of 93.1%. Items were counterbalanced across participants to appear in either noun contexts (“He neglected the sofa in the playroom”) or verb contexts (“He neglected to socialize the puppies when they were young”). As above, our hypotheses were evaluated against fixation proportions to the soap roughly 200-400ms after the onset of the critical word (“socialize” or “sofa”).

Because the competitor is noun/verb ambiguous, both syntactic contexts should lead to changes in activation, under either version of the category constraint. In all cases, the change in activation for the noun/verb ambiguous competitor should be different from the change in activation for the noun-only distractors, making it visible as a change in response probability that would not be expected if there were no constraint. The extent of competitor/distractor difference should vary between the noun and verb contexts, which is, crucially, what leads us to expect a difference in response probability for the competitor between the two contexts.

More specifically, a category constraint acting via inhibition of wrong-category candidates should inhibit activation of “soap” in the context of “to socialize” more than it should inhibit activation of “soap” during the context of “the sofa,” because soap is used more often in

noun contexts than verb contexts, and we assume that such a category constraint would operate proportionally with respect to frequency. Distractors, with activation already at or near baseline, should see little change in activation whether the constraint applies to them or not. We would then expect to see more fixations to "soap" during "sofa" than during "socialize." A category constraint acting via facilitation of correct-category candidates should, we suggest, have nearly indistinguishable effects from the inhibitory constraint. "Soap" should be facilitated in the context of "to socialize" just as in the context of "the sofa," but simply to a lesser extent, based on its frequency imbalance. Distractor activation should increase in the noun context and not the verb context because all distractors are noun-only. This could somewhat neutralize the advantage in the noun context due to the noun-biased frequency, but we would still expect more competition in noun than verb contexts. Therefore, if our assumptions about the representation of category-ambiguous words hold, this second contrast leads us to expect a context effect under either implementation of the constraint, and so could it provide positive evidence that participants apply some form of category constraint. This would be useful in the case that the first contrast yields competition in both noun and verb contexts and therefore does not distinguish facilitation and no constraint. We note that the exact details of expected activation and proportion changes vary considerably with one's assumptions about the implementation of facilitatory and inhibitory constraints.

Because this second contrast predicts a difference in response probability for the noun/verb ambiguous competitor in the noun and verb contexts under both facilitation and inhibition, the timing of this difference would provide information about the timing of the category constraint. The first contrast would only provide timing information under inhibition, as there should be no context difference under facilitation.

We were only able to find 27 usable stimulus triplets (picture word, noun context competitor word, verb context competitor word) in which the noun frequency was higher than the verb frequency. Three with higher verb than noun frequency were included in the experiment so that the number of trials per condition would not be skewed, but these three were not included in the analysis. For these 27 useable competitor words, the mean SUBTLEX frequency per million as a noun was 22.03 and as a verb was 1.28. The mean frequency of the noun-only distractors was 7.92. Recall that because our comparison is between the same display of pictures in different auditory contexts, frequency differences between pictures in the display are not inherently problematic.

The set of distractor pictures used was the same as the set used for the first contrast, but re-shuffled among grids. Half of the competitor pictures appeared twice as targets in filler trials with verb context sentences. These pictures that were used as both competitors and targets were also used twice as distractors in filler trials with noun-context sentences, so that it was not the case that any time a noun-verb picture appeared, it was guaranteed to be a target or competitor.

3.2.2.3 Filler trials with targets

In addition to the two critical contrasts described, we also included filler trials in which the display contained a referent for the auditory target, so that in half of the trials one of the picture names would actually be mentioned in the sentence. These filler trials also allowed us to verify the presence of the most basic type of visual world effect in our data: increased fixations to the picture matching the auditory target. In 30 noun-context filler trials, the display contained the target and three distractors. For example, the four pictures might be: balcony, sword, tractor, clock. The auditory sentence would be “He chose the balcony with a view of the ocean.” In 30 verb-context filler trials, the display also contained the target, but in this case it was a noun-verb

ambiguous item (e.g. a bar of soap with the auditory sentence ("He neglected to soap his hands thoroughly.").

Participants were warned in the initial instructions that they might see instances in which an action in the sentence was related to an object on screen. This was made clear with a practice trial that included a picture of a shovel along with the sentence "He refused to shovel the snow." Participants seemed to find these instances completely straightforward and they answered "yes" (that they had seen something on the screen related to what they were hearing) without issue. These trials were necessary to ensure that participants had motivation to look at the pictures when the sentence context indicated a verb. Notably, we also included noun-verb pictures as distractors in the noun-context filler items (e.g. "clock" above) so that it would not be the case that in all occurrences of a noun-verb picture, that picture was a cohort competitor or target.

3.2.3 Stimulus creation

To construct our set of picture stimuli, we started with all color line drawings of objects available from the BCBL MultiPic (Duñabeitia et al., 2018) and SVLO (Rossion & Pourtois, 2004) repositories, as well as a small number of supplemental clip-art drawings. Items from MultiPic are standardized for name agreement in British English, and name agreement data is also available for the SVLO items, though we did not use this information. Many visual world experiments use black and white line drawings, but we retained color in order to preserve the predictably elicited names, and because our key contrast was between two conditions presenting identical visual stimuli. Color differences between individual pictures were thus irrelevant. We reviewed the full set of possible pictures and excluded any pictures for which multiple names seemed possible, the picture showed a scene or more than one object, the object in the picture was not immediately recognizable, or the picture called to mind a verb before a noun.

We then restricted this set to those whose names were listed with pronunciations and morphological parses in the English Lexicon Project (Balota et al., 2007) and appeared in SUBTLEX-US (Brysbaert & New, 2009) listed as usable only as nouns or only as nouns or verbs. We removed pictures whose names had fewer than 3 phonemes. We used phoneme parses as indicated by the CMU pronouncing dictionary (Weide, 1994). For selecting critical items that would serve as phonological competitors of auditory targets, we removed pictures for which the second phoneme in the name was a vowel and the third phoneme was /m/, /n/, or /r/, to avoid allophonic changes on the vowel. We then took the subset of the remaining pictures that could form a triple with a noun-only and verb-only onset competitor (the auditory targets) listed in SUBTLEX-US (Brysbaert & New, 2009). Distractor-only items were not subject to the constraint on the third phoneme and did not have to form such a triple.

The sentence frames we used (e.g. “She wanted to/the…”, “They chose to/the…”) were adapted from Fox and Blumstein (2016). We always used the past tense, and equally rotated pronoun use across conditions. The remainder of the sentence following the critical item was not subject to any specific considerations other than felicitousness. Our critical items always had a two-phoneme onset overlap with the corresponding/counter-balanced critical item and the corresponding picture in the display (as in Allopenna et al. (1998) and many following studies, in which two phonemes have been shown to be sufficient for eliciting cohort competition). The sentences were recorded by a female native speaker of English who read from a list of the 120 items whose order had been fully randomized. Two recordings were made and the clearer recording of each item was used. After recording we used Praat (Boersma & Weenink, 2014) to normalize the stimulus files at 65 dB and resample to 22.5 kHz.

Using naturally recorded full sentences meant that our context words (to, the) contained some co-articulation potentially providing advance knowledge of the onset identity of the critical word to follow (see Salverda, Kleinschmidt, & Tanenhaus (2014) for thorough consideration of this issue). Though this might lead to cohort competition effects occurring up to 70 ms earlier than if we had used neutral tokens of "to" and "the," this is not an issue in our design, since the onset identities of our critical words are necessarily matched within pairs, such that co-articulation cues in the noun and verb contexts should be the same. We also intended this experiment to reflect the conditions of natural speech as closely as possible, so cross-spliced neutral tokens would have been undesirable in this respect.

A different co-articulation concern might be that the differing third phoneme between e.g. *battleship* and *bask* with respect to the third phoneme in *balcony* might cause the timing of the termination of cohort competition to vary between the two conditions. Any such differences should not be correlated with syntactic category, so they should not introduce a confound in our design. However, to mitigate this, we made the third phoneme the same in the noun and verb competitor words whenever possible, so that co-articulation during the onset would be different in the same way from the target.

Finally, following recording, we used the Montreal Forced Aligner (McAuliffe et al., 2017) to identify word and phoneme boundaries inside the audio files.

3.3 Procedure

We used a tower-mounted SR Research Eyelink 1000 eye-tracker, which has a sampling rate of 1000 Hz, to record eye movements. For the majority of participants, only the right eye was tracked. 14 participants had the left eye tracked instead due to technical issues. Participants were seated comfortably with their chin resting on the chin rest of the tower mount. The setup

was such that the distance from participants' eyes to the center of the (23 inch) monitor presenting the visual stimuli was 38.4 inches. Participants heard mono recordings of the stimulus sentences via speakers located next to the monitor.

We performed a nine-point calibration, and then presented four practice trials and allowed participants to ask questions about the task and request adjustments to the volume before starting the first block of the study. The experiment was presented using SR Research's Experiment Builder software. In each trial, a 3x3 grid appeared on the screen with a picture in each of the four corners. The grid was displayed for 1000 ms before the sentence started playing and disappeared when the sentence ended (since the auditory target was mid-sentence, this yielded more than 1000 ms of preview, which was more than sufficient for phonological competition to arise via implicit naming of the pictures). After a 300 ms blank screen, the task question appeared on the screen ("Did you see anything on the screen related to what you were hearing?") and remained until the participant's button press. After the button press, a dot appeared in the center of the screen for drift correction, after which there was a 300 ms blank and then the next grid was presented. After every block of 20 trials, participants were given the option to take a break. A mandatory break occurred every two blocks, when re-calibration was performed. We also re-calibrated after any additional voluntary breaks. Each of the six blocks of 20 trials lasted three to four minutes. The entire experiment took roughly 25 minutes.

3.4 Participants & statistical power

To determine the appropriate sample size for this study, we started with Huettig and McQueen (2007)'s phonological competition effect for noun pictures in noun context. There was no existing design or analysis that exactly matched our design and planned analysis, and we considered the Huettig and McQueen competition effect to be as comparable as possible to what

we expected to observe in our study because their design also omitted a target. However, the design in that study differed from ours in that it included a semantic and a shape competitor. The analysis also differed from our planned analysis in that it computed the ratio of the competitor fixation proportion to the summed competitor and distractor fixation proportions and compared this value to 0.5, while we compared competitor fixations to the competitor's own baseline (see next section). Huettig and McQueen also tested averages within 100 ms. time windows, while we used temporal cluster tests. Thus, our sample size estimate for achieving 80% power is only an approximation, given the information available to us.

The Huettig and McQueen phonological competition effect is present in several time windows, but the smallest *t*-value with which it manifests is 3.3. With a sample size of 30, this constitutes an effect size (Cohen's d_z) of 0.602. However, because effect sizes in published studies are systematically over-estimated (recently discussed by e.g. Vasishth, Mertzen, Jäger, & Gelman (2018)), we used a more conservative estimate of 0.4. Power analysis using G*power (Faul et al., 2009) indicates that for a repeated measures one-tailed *t*-test with desired alpha of 0.05 and desired power of 80%, the necessary sample size for a simple phonological competition effect is 41. However, our intention was to test for the equivalent of an interaction that would indicate whether or not the phonological competition effect differed between context conditions.

Conventional wisdom indicating that interaction designs should maintain the same N per cell would dictate doubling the sample to 82 participants. However, because the nature of the expected effect was a "knock-out" interaction in which the simple effect is present in one condition but absent in the other, the interaction effect size can be estimated at roughly half the magnitude of the main effect. Therefore, "knock-out" interactions are considered to require four times as many participants, yielding a recommended sample size of 164 (see e.g. Simonsohn

(2015)). We collected data from 165 participants, but 21 datasets were excluded due to technical issues or because the participant reported learning a language other than English before age 7, so only 144 were included for analysis. Thus, with 144 rather than 164 participants, we in fact had power of 76%, not 80%, to detect the expected knock-out interaction based on the Huettig and McQueen phonological competition effect.

For the contrast in which we used noun-verb ambiguous competitors, we expected the fixation increases due to phonological overlap to be larger because of the higher frequency of the picture names. But we also did not expect the interaction between contexts for this contrast to be a "knock-out" interaction; the effect in verb context was predicted to be attenuated but not completely eliminated, which would make the effect harder to detect. Therefore, we could not calculate or even approximate the necessary sample size for 80% power in this case.

All participants were recruited from the University of Maryland community and were at least 18 years of age. All gave informed consent and were compensated for their time with course credit or cash ($12/hour, and sessions typically lasted one hour). The mean age for the 144 participants included in the analysis (58 men, 86 women) was 20.7 years.

3.5 Analysis

Because the task for participants was to indicate whether they had seen anything on the screen that was related to what they were hearing, and relatedness is subjective, we did not use task response accuracy as a filter for dataset inclusion. The data were processed by first removing all samples labeled by the eye-tracking software as saccades. Our threshold for the number of samples that could be missing from a trial was 2000, i.e., 2 seconds of data; any trials exceeding this threshold were excluded. Out of the 144 datasets included for analysis, 128 had no trials excluded. For the 16 datasets with excluded trials, the average number of excluded trials

was 2. For each fixation sample, we coded whether it was to a competitor/target or to one of the distractors. Using onset information from the output of the Montreal Forced Aligner (McAuliffe et al., 2017), we also coded what word in the sentence the sample had occurred during. We did not include for analysis the three items from the critical noun-verb ambiguous condition that were verb-biased rather than noun-biased.

We then extracted only the fixations that occurred during the context word and the critical item, taking a 400 ms window time-locked from the onset of the context word and a 1000 ms window (which could overlap) time-locked from the onset of the critical word. For each participant and condition, for each time point in these two relative time courses, we calculated the proportion of instances of this time point across trials for which the fixation was to the competitor. Here and following, we will use the term “competitor” to refer to the picture name whose onset overlaps with the word being heard. In filler trials, this competitor turns out to fully match the auditory target. This gave us a single time course of proportions of looks to the competitor for each participant and condition, in the context and critical windows. Finally, for each participant and condition we computed the mean proportion of fixations to the competitor in the first 100 ms of the context word; this served as the baseline for that participant/condition. We subtracted this baseline from the proportion of fixations at each time point in the critical window to create a "competitor advantage" reflecting any increase in the proportion of fixations to the competitor relative to a time window when looks could not have been driven by the difference between conditions. We then smoothed the data using a 20 ms Hamming window. The time course of the competitor advantage for each participant was submitted to temporal cluster tests.

For filler trials in which the competitor (the critical item) turns out to fully match the auditory target, the temporal cluster tests were one-sample *t*-tests against zero to determine when

there was a reliable competitor advantage, i.e., when fixation proportions were greater for the mentioned item. This was done separately for noun-only competitors and noun-verb ambiguous competitors. For the critical trials, the temporal cluster tests were related-measures *t*-tests to determine whether there was a difference in the time course of the competitor advantage between the noun and verb contexts. This was done separately for noun-only competitors and noun-verb ambiguous competitors. We then conducted one-sample *t*-tests against zero asking when there was a reliable competitor advantage within each context.

Temporal cluster tests were always one-tailed and conducted with 10,000 permutations and a threshold of $p < .05$ for forming clusters. For fillers, we conducted tests in a single window from 100 to 1000 ms, because the competitor advantage was expected to increase systematically throughout the entire window. For critical trials, we conducted tests in two separate windows, based on the results of Strand et al. (2018) and Huettig and McQueen (2007). Both found that fixations to competitors began to decrease slightly after 500 ms, so we broke the epoch into two equal-sized windows that would capture this. The first window, from 100 to 550 ms, was where we expected competition to occur robustly (at least in the correct context). In the second window, from 550 to 1000 ms, we expected weaker and tapering competition from (at least) the noun-only competitors. We used the same time windows for the noun-verb ambiguous competitors.

The choice to evaluate the proportion of competitor fixations relative to its own baseline is not completely standard. Many visual world analyses instead consider when fixations to the competitor exceed fixations to the distractor items. However, it is difficult to ensure that fixations to the competitor and distractors at the beginning of the trial are perfectly matched; they may vary due to slight differences in frequency, visual salience, or simply chance. Since what we

are interested in assessing is when (and to what extent) the activation of the competitor exceeds what its activation level was before auditory input, we maintain that the increase in the proportion of fixations to the competitor is the most directly relevant dependent measure. This comparison is possible in our design exactly because the competitor is the only item whose activation can be expected to change, making it straight-forward to reason from changes in its proportion of fixations. In a typical design that includes a target, the change in the proportion of fixations to the competitor is also influenced by changes in activation for the target. Comparison between the competitor and the distractors is therefore preferable in these designs because the distractor proportions are also influenced by the target.

3.6 Results

For the noun-only filler trials (**Figure 3**) in the window of 100-1000 ms, the competitor advantage was found to be significantly different from zero starting 175 ms after auditory target onset and persisting for the remainder of the epoch ($p < 0.001$). Note, again, that we use the term "competitor advantage" for competitors in filler trials as well as for competitors in our critical trials; it refers to the increase in fixations relative to baseline for the item whose onset overlaps with the auditory target, and in filler trials the rest of the word is a match as well. For the noun-verb ambiguous filler trials (also plotted in **Figure 3**), the competitor advantage was found to be significantly different from zero for the entire analysis window ($p < .001$). These results indicate that participants did, indeed, look more at the pictures whose names they were hearing. The competition effects are somewhat earlier than is often observed but given the available co-articulatory cues on the context word, this is not surprising. In the second half of the trial, the proportion of looks to the target appears to reach a slightly higher maximum for the noun-only relative to the ambiguous condition, though we had no hypotheses regarding such a difference.

This could potentially indicate that participants had slightly less confidence in the noun-verb ambiguous pictures as referents in verb context than they did in the noun-only pictures as referents in noun context.

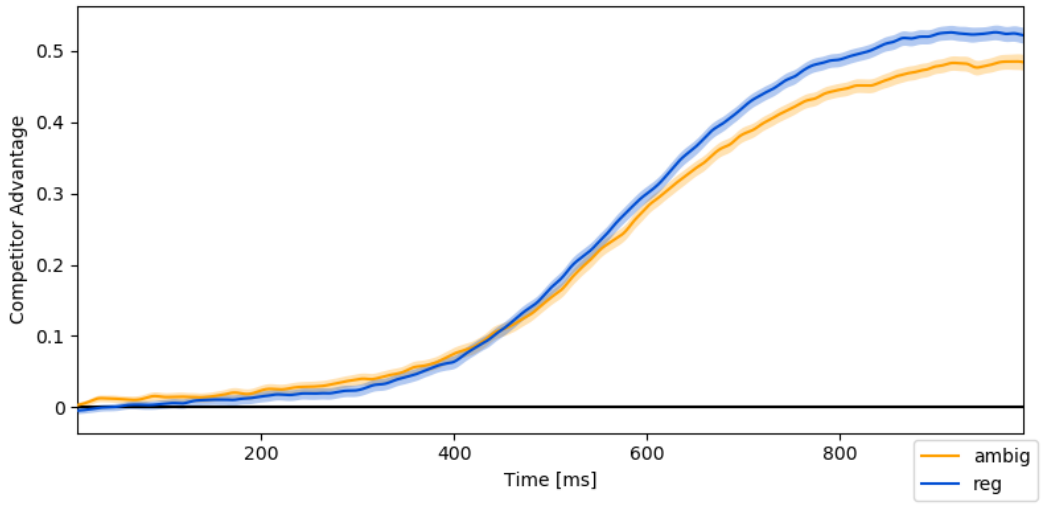

Figure 3. *For filler trials, the smoothed time course of competitor advantage relative to baseline following auditory target onset, for noun-only ('reg') and noun-verb ambiguous ('ambig') competitors, which are fully consistent with the auditory target. Shading indicates one standard error.*

For our first critical contrast, which focused on cohort competitors whose names were category-unambiguous nouns, there were no clusters indicating a difference in the competitor advantage between the noun and verb contexts **(Figure 4**). One-sample *t*-tests in each context in the early window (100-550 ms) indicate significant clusters in which the competitor advantage differs from zero in both the noun context, from 263 to 550 ms ($p < .01$), and in the verb context, from 316 to 453 ms ($p < .05$). In the later time window, we found significantly increased fixations to the noun-only competitor in the noun context (550-706 ms and 749-928 ms, both $p < .05$) but not in the verb context. The magnitude of the competition effect (a difference in fixation

proportions of ~0.04, relative to baseline) is comparable to or slightly larger than what can be observed in the plots from Huettig and McQueen (2007) and Strand et al. (2018), though neither study reports exactly this measure. An obvious concern about our failure to find a difference between the noun and verb contexts is that this could have been due to lack of power, despite our efforts to ensure a sufficient sample size. However, the knock-out interaction that we expected was such that there would be no competition in the verb case, and we found a significant effect in this condition. Thus, while there may be a smaller difference between the conditions that we are unable to detect, we do have positive evidence against the inhibitory effect as it was hypothesized.

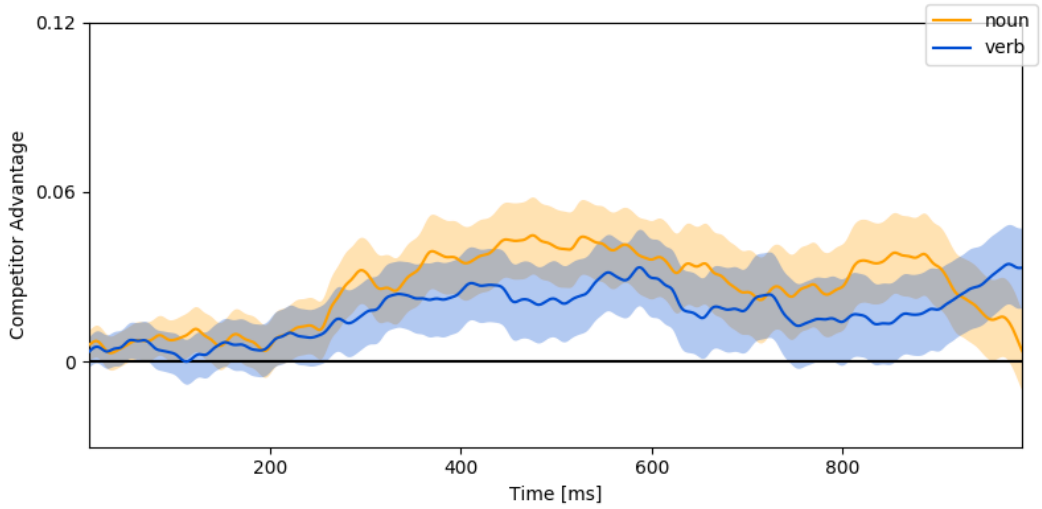

Figure 4. *For the noun-only competitor in critical trials, the smoothed time course of competitor advantage relative to baseline following auditory target onset, in noun and verb contexts. Shading indicates one standard error.*

For our contrast between noun-verb ambiguous competitors (**Figure 5**), we also observed standard cohort competition effects, but again there were no clusters indicating a difference in the competitor advantage between the noun and verb contexts. Competitor advantage effects

were numerically bigger than in the noun-only contrast, likely due to the higher overall lexical frequency of the competitor names. One-sample *t*-tests in each context in the early window indicate significant clusters in which the competitor advantage differs from zero in the noun context, from 197 to 550 ms ($p < .001$), and in the verb context, from 270 to 550 ms ($p < .01$). In the later time window we found significant competition in both contexts, ending at 801 ms in the noun context ($p < .01$) and 963 ms in the verb context ($p < .001$).

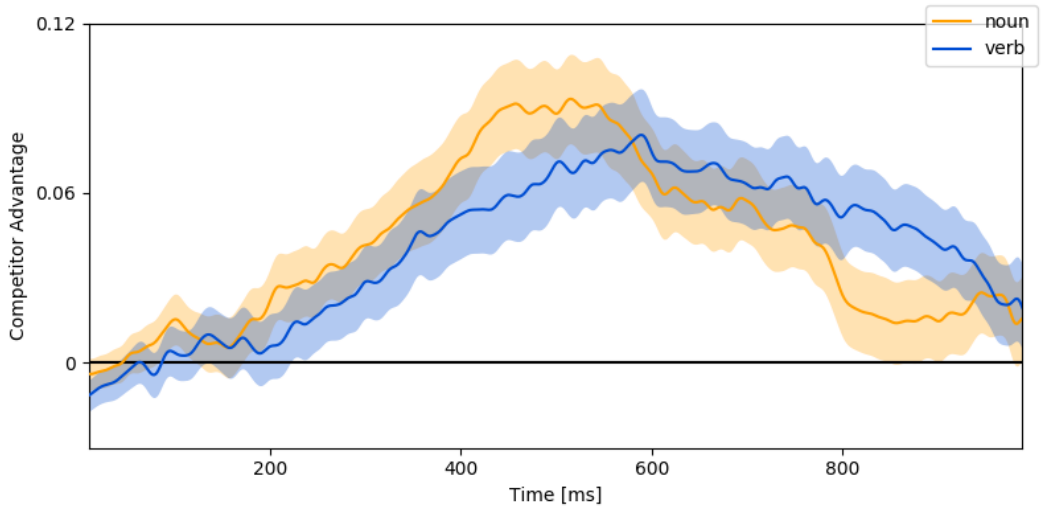

Figure 5. *For the noun-verb ambiguous competitor in critical trials, the smoothed time course of competitor advantage relative to baseline following auditory target onset, in noun and verb contexts. Shading indicates one standard error.*

3.7 Discussion

In this study, we used the visual world paradigm to ask whether and how syntactic category expectations constrain lexical activation. In syntactic context, is it the case that the activation of wrong-category candidates is prevented, that the activation of correct-category candidates is boosted relative to wrong-category candidates (which still compete), or that activation is unaffected by category status? Previous studies have conflicted in whether or not

they find evidence for the activation of wrong-category candidates. This lack of coherence in the literature may be due in part to designs that consider only the expected effects of an inhibitory mechanism or are not sufficiently sensitive to detect wrong-category competition. Our study, in contrast, was specifically intended to distinguish between facilitatory and inhibitory mechanisms for the syntactic context constraint. Participants heard auditory targets that they could expect to be either nouns or verbs on the basis of the syntactic context. We then measured fixations to pictures of cohort competitors that could only be nouns or were strongly noun-biased. Our design was set up to maximize the chance that the activation of wrong-category competitors would be detectable if it were occurring. Our simulations in jTRACE had demonstrated the potential influence of the response candidate set on whether and when changes in a critical item's activation become visible in response probability. Therefore, we did not include pictures in the display whose activation changes might obscure the activation changes of our critical items. We also ensured that the pictures in the display could not be classified as noun or verb referents based solely on their visual properties, so that context effects on fixations could only be driven by the syntactic category information associated with the word-forms.

Given these design parameters, we found immediate cohort competition from wordforms that were incompatible with their syntactic context, with no evidence that the size of the effect differed between the wrong-category and correct-category conditions. This contrasts with previous work in the visual world paradigm (Magnuson et al., 2008; Strand et al., 2018) but it is consistent with earlier evidence for wrong-category competition in cross-modal priming (Seidenberg et al., 1982; Tanenhaus et al., 1979; Tanenhaus & Donnenwerth-Nolan, 1984) and gating (Tyler, 1984; Tyler & Wessels, 1983). We observe the same pattern of significant but

indistinguishable cohort competition from noun-biased cohort competitors in noun and verb contexts.

Our finding of competition in both noun and verb contexts for the noun-only competitors is inconsistent with the prediction from an inhibitory account that we should see normal cohort competition in the noun context and no cohort competition in the verb context, as Strand et al. (2018) and Magnuson et al. (2008) had found for the analogous comparisons in their designs. Our finding is consistent with both a facilitatory account and an account in which syntactic context did not constrain lexical activation at all. The simulations we reported in Section 2 had suggested the possibility that apparent effects of an inhibitory constraint observed in prior visual world studies were not in conflict with other methods but were simply visible earlier due to the nature of forced-choice tasks. Our visual world study does not support an inhibitory constraint, but the task-related timing differences predicted by our simulations may still apply, and should be investigated in future work.

Our contrast between noun-verb ambiguous competitors was aimed at distinguishing these two remaining alternatives, but the results were somewhat equivocal. We expected that a constraint that operates in proportion to frequency in the category, whether faciliatory or inhibitory, would lead to greater activation and competition in the noun context than the verb context, because our noun-biased items would be either more facilitated or less inhibited. Without a constraint, the noun-verb ambiguous competitors should appear the same in both contexts. We found a numerical difference between the contexts, but the difference was not significant, suggesting either that the syntactic constraint does not apply to the competition process or that we lacked the power to detect a relatively small modulation in the competition

effect. Because our power analysis did not extend to this condition, we cannot know what our power was in this case.

3.7.1 Considerations for the inhibitory account

Our findings do not support an inhibitory constraint in which wrong-category competitors are not activated. We believe an assumption that constraints on cohort competition are inhibitory is implicit in many experimental designs, because the effect of interest is simply the presence or absence of wrong-category competition. Testing for a facilitatory constraint, in contrast, requires examining the modulation of correct-category competition.

Why did we observe wrong-category competition when two previous visual world studies (Magnuson et al., 2008; Strand et al., 2018) did not, and so appear to support the inhibitory account? In each case, the appearance of such a constraint could have arisen because of design properties that led participants not to fixate on wrong-category competitors even though they were activated.

In the Magnuson et al. (2008) study, the experimental context may have impacted the way in which images were implicitly labeled by participants. As discussed in the Introduction, this study used an artificial language consisting of texture adjectives and shape nouns that could be used to describe the items presented on screen. When four different shapes were presented, participants could expect that they would hear only a shape noun after "Click on the..", because using an adjective would be unnecessarily specific. When two items in the display had the same shape, participants could expect that an adjective would be used to differentiate them, and therefore that they would hear an adjective after "the." This was the basis of the expected syntactic constraint on cohort competition. Wrong-category competitors were included in the display to test whether they would be fixated more than distractors. These wrong-category

competitors consisted of, for example, shapes whose texture adjectives were cohort competitors of the target's shape noun when only a shape noun was expected. Failure to look at these competitors more than distractors was taken to be evidence that their phonological representations were not actually competing in the cohort, due to the syntactic constraint.

However, for observable cohort competition to occur requires that there be a match between the name for the picture and lexical representations activated by the auditory input. We assume that this match is what drives fixations. We want to know whether wrong-category phonological representations are activated by the auditory input, and in order for their activation to be detectable in fixations, there would need to be a picture name with which a match could occur. The logic of this study therefore relies on the assumption that participants use the same implicit label, of the form “adjective-noun”, for each image on every trial, regardless of the pragmatic context, so that in e.g., trials where a noun is expected, if adjectives are nevertheless activated by the auditory input, they can still match with the picture name and drive fixations, because the picture name contains the adjective. But if the pragmatic context for the noun trials led participants to implicitly label the images with their shape (noun) name only, then even if adjectives are activated via the auditory input (which is our question) there will be no opportunity for them to match, and therefore no way for them to drive fixations. This would make wrong-category cohort competition undetectable.

In the case of Strand et al. (2018), the issue is instead related to potential strategies arising from regularities in the stimuli. Strand et al. (2018) include pictures of actions in their visual world design, so that verbs can also have referents and they can plausibly ask whether participants fixate verb competitors during noun contexts and vice versa. This solves a major problem for the visual world paradigm. Unfortunately, though, it also introduces a new problem.

Because action pictures depict fundamentally different entities than object pictures do, they tend to be visually quite distinct. In fact, it is generally easy to surmise from a line drawing whether its intended referent is a noun or a verb. Most verb pictures involve a human or animal in motion, and most noun pictures involve a simple object. Because Strand et al. (2018)'s design is such that a verb target always has a verb picture as a referent and a noun target always has a noun target as a referent, we suspect that it was possible for participants to know after the context word, before the target word has started, which two of the four pictures on screen could be the referent and therefore which they should look at. In other words, they could know not to look at the wrong-category competitor even before the critical word had started. So this, again, is a case where the activation of wrong-category competitors becomes undetectable, because the link that would allow it to drive behavior is overcome by a more powerful strategy.

An additional concern in the Strand et al. (2018)'s design is that the presence of correct-category distractors has the potential to obscure changes in activation from wrong-category competitors. The logic of studies of this kind is that increased looks to a wrong-category competitor, relative to a distractor, is evidence of increased activation. But if the distractor itself might receive an activation boost, by virtue of being associated with the same category that is supported by the context, then this might mask any increased looks to the wrong-category competitor.

Beyond these visual world studies, as reviewed in the Introduction, other previous results are more consistent with the absence of an inhibitory constraint. Findings from gating have never been consistent with an inhibitory constraint, because they have shown that wrong-category options are initially proposed. Cross-modal priming is restricted to homophony and only demonstrates an inhibitory-appearing constraint 200 ms after word offset, which is too late to

bear on the question of initial cohort competition. We conclude that the previous literature, in combination with our finding of wrong-category competition, does not support the possibility of an immediate inhibitory constraint that prevents initial wrong-category competition.

3.7.2 Considerations for the facilitatory account

We observed more competition from our noun-biased noun/verb ambiguous competitors in noun context than verb context, consistent with the presence of a syntactic constraint, but this difference was not statistically significant. An a priori estimate of the effect size for this contrast was not possible. However, if there is a true effect that we failed to detect due to lack of power, even with 144 participants, this raises questions for the viability of the visual world paradigm in answering questions along these lines, unless experimenters are willing to drastically surpass the typical sample size. Even with large sample sizes, while there may be other potential visual world designs that could provide positive evidence for a facilitatory constraint, the design considerations that we have discussed in this paper considerably restrict the space of possibilities. We have described why noun-only and verb-only competitors should not be compared in the same display, both because activation of only one item in the display should be manipulated at one time for clear inferences about response probability, and because of the action/object picture confound. However, once all four items in the display have the same syntactic category, a proportional increase that applies to all four equally, due to the context, will be impossible to see as a change in response probability. Comparisons of competition in and out of context are also potentially problematic because of the possibility that there is some degree of default expectation for nouns as targets in the visual world paradigm, since it relies on easily imageable referents. Future work will therefore need to seek alternative evidence for or against the facilitatory constraint likely outside the visual world paradigm.

Support for a faciliatory constraint would mean that findings of wrong-category competition in cross-modal priming and gating are not evidence against the existence of the constraint, as they have frequently been taken to be, and as they would be under inhibition. Wrong-category competition is expected in this scenario, and in order to bear on the presence or absence of a constraint on lexical activation, these methods would need to establish whether or not there is increased competition for words from the correct category relative to the incorrect one, or for the correct category relative to no context. McAllister (1988) finds that there is such a pattern in gating, where the presence of context leads to a change in the percentage of incorrect-category items proposed but does not eliminate those incorrect-category items entirely. This has not been shown in the relevant cross-modal priming design, but it is exactly the type of effect that Lucas (1999) identifies in her meta-analysis of (primarily) semantic context effects in cross-modal priming. She notes that most of the studies she considers are not adequately powered to detect such an effect, as we argue may also have occurred in our visual world study.

What would a facilitatory constraint mean for our understanding of lexical representation, word recognition, and lexical access? The presence of a constraint at all, and especially an immediate one, would indicate the status of syntactic category as the type of lexical feature that can be used as a cue to alter wordform activation. Furthermore, a facilitatory constraint means that the effect of auditory input on the cohort is not overridden or outweighed by top-down syntactic information. An inhibitory constraint could make it impossible to recognize a word if the context was misheard or the word is simply very unexpected in its context. Preserving bottom-up information, as a faciliatory constraint would, could have significant utility. However, among the different types of information that sentence context provides, we would argue that syntactic category expectations lead to some of the clearest predictions for what should or

shouldn't be considered next. Category restrictions are harder to violate than plausibility restrictions, for example, as violations can lead to ungrammatical phrases or sentences. Therefore, syntactic category information should be a good candidate for an inhibitory constraint, if one exists. That not even syntactic category appears to operate this way makes it perhaps less likely that other types of information would do so, and raises questions about potential implementational issues with such a mechanism.

3.7.3 The no-constraint account

If we take our null effect for noun-verb ambiguous competitors to mean that there is indeed no effect of syntactic context on cohort competition, we must contend with the evidence from previous work that a syntactic category constraint does apply to word recognition, and consider alternative explanations. We have explained at length how the mistaken appearance of a constraint may have occurred in the visual world paradigm. In cross-modal priming, early probe points do not show context effects, though a high-powered study would be necessary to more confidently rule them out. The later constraints observed in cross-modal priming have no obvious explanation under an account where syntactic context has no impact at all. The context effect found in gating (McAllister, 1988) is also problematic for the no-constraint account unless gating itself is disregarded because of the possibility of conscious reflection before a participant proposes a competitor. The no-constraint account is less well supported than the facilitatory account, but the two cannot be distinguished more definitively without additional data points from multiple methods.

3.7.4 Caveats

A final explanation for our findings that we must acknowledge is that particular aspects of our own study might have resulted in context information being simply ignored or not being usable quickly enough, in a way that is unrepresentative of processing in natural situations. Although possible, we find this explanation hard to reconcile with the fact that Strand et al. (2018) found that syntactic context impacted looks in a similar visual world design with very similar sentences. While we suspect that wrong-category competition was not detectable in that design because visual properties of the pictures tracked syntactic category, allowing an alternative strategy for avoiding wrong-category visual referents, this strategy still itself relied on the early availability of context information. Nevertheless, a study in which more time is available between the context cue and the onset of the target word, or in which the sentence context is more involved and might lead to stronger predictions, would help clarify this issue.

We also note that our second contrast, designed to distinguish facilitation and a lack of constraint, depended on a separate set of assumptions about how homophony is mentally represented and accommodated during lexical access. In order to ask whether a constraint was in place, we tested whether noun-verb ambiguous (but noun-biased) items competed more in noun context than verb context, assuming that the constraint would lead them to compete in proportion to their frequency in the category. If our assumptions are misplaced regarding the representation of homophony or the way it specifically is influenced by context, it could be that we should not expect such an effect even if a syntactic constraint is applying. However, it is difficult to imagine designs that avoid this problem, and general understanding of the interaction between homophony and category representation is not so developed that there is an obvious alternative prediction for our design.

Finally, as we acknowledged at the outset of this paper, we have restricted our consideration of the syntactic constraint to the contrast between total inhibition (by which activation levels are reduced to zero, or to baseline) and facilitation. Partial inhibition, in which activation of wrong-category candidates is not completely reduced, yields an outcome that would be extremely difficult to distinguish from the outcome of a facilitatory constraint: more activation for the correct category than the wrong category. It might be possible to make this distinction in a design comparing competition in and out of context.

3.7.5 Further questions

Our study has focused on a relatively simple test case for the question of whether and how context affects word recognition. Investigating syntactic context, specifically, makes it relatively easy to say which lexical items should be considered compatible or incompatible with the context, and therefore which should be subject to the contextual constraint (depending on its implementation). However, it is often not the case that syntactic context unambiguously predicts or rules out entire syntactic categories. Though a repetitive experimental setting may make contextual cues fully or nearly deterministic, syntactic context is rarely deterministic in natural language (even for the simple examples we used in this study), and category predictions are therefore likely to be probabilistic. Furthermore, as employed and explored in our second contrast, many wordforms are compatible with more than one syntactic category. The extent to which the different category usages of the same wordform map onto the same or overlapping meanings is also quite variable. Each of these areas of uncertainty is relevant for a complete understanding of how syntactic context functions as a constraint on word recognition. Another obvious area needing more consideration is the recognition of multimorphemic words, whose complexity we have largely avoided here. Finally, how exactly syntactic category expectations

arise from syntactic structure-building and prediction, and in what form, is an open question. Scaling from syntactic to other even less deterministic forms of context only complicates these matters further, but a thorough mechanistic account of even the simplest type of context effect is a necessary first step.

4 Conclusion

Our visual world study and the jTRACE simulations that informed it provide new perspectives on the interaction between top-down and bottom-up input in language processing, as well as on the measures we use to study that interaction. An understanding of the cognitive mechanisms by which syntactic context influences cohort competition has so far been lacking, and previous evidence on whether a constraint applies at all has been conflicted. Investigating that cross-method conflict, we showed with simulations in jTRACE that the manifestation of cohort competition effects in behavioral measures can be expected to depend on the composition of the response candidate set for the task. This can potentially be extended to our understanding of the timing of top-down influences on the cohort, as a means of explaining differences between tasks with different response candidate sets. We then used this understanding of the importance of the response candidate set to inform a visual world eye-tracking study that could clarify a different and important factor in how syntactic context effects manifest. Our design was intended to distinguish whether the potential category constraint operates by facilitating good fits for the context while allowing wrong-category competition, or by inhibiting wrong-category competition altogether. We found that wrong-category competition does occur, which is incompatible with an inhibitory constraint on word recognition due to syntactic category expectations.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grants DGE-1449815 and BCS-1749407. PG was also supported by a Flagship Fellowship from the University of Maryland. We thank Margaret Kandel, Anna Namyst, Nadiya Klymenko, Oliver Bentham, Reese Alpher, and Lalitha Balachandran for assistance with stimulus preparation and data collection. We also thank Naomi Feldman and Bill Idsardi for feedback on earlier versions of this paper. Some of the results presented in this paper were previously reported in the Proceedings of the 11th International Conference on the Mental Lexicon 2018.

Appendix: Complete list of stimuli

Condition	List 1 sentence	List 2 sentence	Critical Picture	Distractor 1	Distractor 2	Distractor 3
Critical: noun-only competitor	He chose the battleship for his birthday.	He chose to bask in the sun.	balcony	moustache	curtain	wheelbarrow
Critical: noun-only competitor	He demanded the reptile as a birthday gift.	He demanded to represent his brother in court.	rectangle	suspenders	cactus	elbow
Critical: noun-only competitor	He loved the drummer of the band.	He loved to dramatize his stories.	drawer	grapes	pineapple	mitten
Critical: noun-only competitor	He preferred the macaroni from the blue box.	He preferred to maximize his profits.	mattress	scarf	accordion	needle
Critical: noun-only competitor	He wanted the baggage for his upcoming trip.	He wanted to baptize the child at the same church.	basket	cabin	sword	gorilla
Critical: noun-only competitor	She hated the cobras at the zoo.	She hated to coerce the witness.	coconut	lips	guitar	magnet
Critical: noun-only competitor	She needed the glaciers in the background of the photo.	She needed to glean more information from her sources.	glove	calendar	shrimp	apron
Critical: noun-only competitor	She offered the bricks for their project.	She offered to browse a section of the bookstore.	bracelet	fruit	kitchen	notebook
Critical: noun-only competitor	She preferred the sweetener from natural sources.	She preferred to swaddle the baby.	sweater	ballerina	camel	ankle
Critical: noun-only competitor	She wanted the cafeteria to be open.	She wanted to captivate the audience.	calculator	greenhouse	microscope	fingerprint
Critical: noun-only competitor	They hated the jesters because of their hats.	They hated to jeopardize their chances.	jellyfish	slippers	tractor	compass
Critical: noun-only competitor	They offered the scrapbook as an apology.	They offered to sculpt a new statue.	scorpion	balloon	dolphin	kiwi

Critical: noun-only competitor	They remembered the salamander that was hiding behind the rock.	They remembered to satisfy the program requirements.	saxophone	desk	acorn	island
Critical: noun-only competitor	They tried the helmets before returning them.	They tried to hesitate before speaking up.	hedgehog	lipstick	wagon	cube
Critical: noun-only competitor	They tried the machines at the gym.	They tried to muffle the noise from the radio.	mosquito	pliers	coffin	vulture
Critical: noun-only competitor	He expected to calculate what he had spent.	He expected the cashier to arrive early.	cabin	bracelet	pomegranate	eyebrow
Critical: noun-only competitor	He forgot to submit the assignment.	He forgot the celebrity from the movie.	suspenders	mattress	umbrella	triangle
Critical: noun-only competitor	He prepared to categorize the new inventory.	He prepared the casserole for his neighbor.	calendar	balcony	scissors	lighthouse
Critical: noun-only competitor	He prepared to gratify his mother's wishes.	He prepared the gravestone for the new plot.	greenhouse	drawer	violin	medal
Critical: noun-only competitor	He refused to budge from his seat.	He refused the buckets of candy.	balloon	sweater	lamp	rhino
Critical: noun-only competitor	He remembered to decorate the office for Mary's birthday.	He remembered the deaths of his grandparents.	desk	basket	funnel	cigar
Critical: noun-only competitor	She chose to plagiarize rather than write the essay.	She chose the plates for the banquet.	pliers	scorpion	dominoes	ruler
Critical: noun-only competitor	She liked to baffle her parents with trivia.	She liked the backyard at the new house.	ballerina	calculator	earring	dragonfly
Critical: noun-only competitor	She wanted to liberate the animals at the zoo.	She wanted the liquor at the back of the cabinet.	lipstick	jellyfish	trophy	fireplace
Critical: noun-only competitor	They hated to greet rude visitors.	They hated the gremlin in the haunted house.	grapes	coconut	stool	megaphone

Critical: noun-only competitor	They knew to scour the old pan.	They knew the skydiver who was in the accident.	scarf	rectangle	ambulance	clarinet
Critical: noun-only competitor	They liked to freshen the flowerse.	They liked the frogs in the pond.	fruit	glove	helicopter	controller
Critical: noun-only competitor	They prepared to liquefy the helium gas.	They prepared the liver for transplant.	lips	hedgehog	salad	glasses
Critical: noun-only competitor	They remembered to slacken the rope.	They remembered the slogan from the commercial.	slippers	mosquito	toothbrush	chameleon
Critical: noun-only competitor	They wanted to matriculate in the spring semester.	They wanted the molasses for gingerbread cookies.	moustache	saxophone	keyhole	globe
Critical: noun-verb ambiguous competitor	He expected the throat to be inflamed.	He expected to thrive in the new environment.	thread	bracelet	cactus	fireplace
Critical: noun-verb ambiguous competitor	He neglected the sofa in the playroom.	He neglected to socialize the puppies when they were young.	soap	drawer	camel	triangle
Critical: noun-verb ambiguous competitor	He preferred the witches rather than the ghosts.	He preferred to wilt the spinach.	whistle	balcony	dolphin	clarinet
Critical: noun-verb ambiguous competitor	He prepared the blender for shipping.	He prepared to bless the worshippers.	blanket	sweater	guitar	chameleon
Critical: noun-verb ambiguous competitor	He tried the mackerel but it wasn't very fresh.	He tried to madden the teacher with his antics.	mask	basket	tractor	rhino
Critical: noun-verb ambiguous competitor	She chose the brownie for her snack.	She chose to brighten the room with fresh paint.	bread	scorpion	kitchen	dragonfly
Critical: noun-verb ambiguous competitor	She chose the platypus as her essay topic.	She chose to placate the students with extra recess.	plant	mattress	acorn	globe
Critical: noun-verb ambiguous competitor	She declined the textile in favor of exposed brick.	She declined to testify in the trial.	telephone	saxophone	microscope	ruler

Critical: noun-verb ambiguous competitor	She liked the myths that she was reading for Latin class.	She liked to mystify her students.	milk	hedgehog	curtain	cigar
Critical: noun-verb ambiguous competitor	She tried the fridge in the lunch room.	She tried to frighten her brother.	frame	rectangle	coffin	eyebrow
Critical: noun-verb ambiguous competitor	They liked the puddle that formed every time it rained.	They liked to publish new blog posts on Mondays.	puzzle	coconut	sword	glasses
Critical: noun-verb ambiguous competitor	They offered the palette that most clients preferred.	They offered to pacify the child with a walk.	package	mosquito	shrimp	lighthouse
Critical: noun-verb ambiguous competitor	They preferred the rodent to the insect.	They preferred to rove the hallways without a nurse.	rope	glove	pineapple	megaphone
Critical: noun-verb ambiguous competitor	They prepared the dresser for the move.	They prepared to drown out the noise.	drop	jellyfish	accordion	controller
Critical: noun-verb ambiguous competitor	They tried the sheet that fit the bed more tightly.	They tried to sheathe the sword before taking it onstage.	shield	calculator	wagon	medal
Critical: noun-verb ambiguous competitor	He chose to dedicate a song to his daughter.	He chose the decks of the ship that were less crowded.	desert	suspenders	keyhole	fingerprint
Critical: noun-verb ambiguous competitor	He chose to squander his allowance on candy.	He chose the schoolhouse as the backdrop for the photos.	skateboard	fruit	violin	gorilla
Critical: noun-verb ambiguous competitor	He refused to swelter in the heat.	He refused the suede for the costume.	swing	desk	pomegranate	vulture
Critical: noun-verb ambiguous competitor	He tried to scamper away before being noticed.	He tried the skunks but they didn't like the new enclosure either.	screw	calendar	trophy	needle
Critical: noun-verb ambiguous competitor	He tried to tether the ball to the pole.	He tried the textbook from the famous professor.	telescope	cabin	dominoes	wheelbarrow
Critical: noun-verb ambiguous competitor	He wanted to notify the parents quickly.	He wanted the notebook with the blue cover.	nose	lipstick	toothbrush	island

Critical: noun-verb ambiguous competitor	She liked to cling to her mother.	She liked the clinic at the new hospital.	clock	moustache	scissors	notebook
Critical: noun-verb ambiguous competitor	She loved to translate poems in her spare time.	She loved the tracksuit that they found in the closet.	trumpet	ballerina	salad	kiwi
Critical: noun-verb ambiguous competitor	She preferred to flout the rules.	She preferred the florists at the old shop.	flag	pliers	earring	mitten
Critical: noun-verb ambiguous competitor	She wanted to popularize her views on immigration.	She wanted the pottery that she saw on the shelf.	pocket	greenhouse	ambulance	magnet
Critical: noun-verb ambiguous competitor	They expected to boggle the listeners.	They expected the bobbin to break.	bottle	slippers	lamp	cube
Critical: noun-verb ambiguous competitor	They preferred to transform the extra bedroom into an office.	They preferred the trapezoid for the company logo.	truck	balloon	umbrella	compass
Critical: noun-verb ambiguous competitor	They refused to plod along the path.	They refused the player from the opposing team.	plug	grapes	funnel	ankle
Critical: noun-verb ambiguous competitor	They tried to mobilize older voters.	They tried the molars but they were also very sensitive.	motorcycle	lips	stool	apron
Critical: noun-verb ambiguous competitor	They wanted to magnetize the sheet of metal.	They wanted the mast from the disassembled sailboat.	map	scarf	helicopter	elbow
Filler: noun-only target	He chose the balcony with a view of the ocean.	He chose the balcony with a view of the ocean.	balcony	sword	tractor	clock
Filler: noun-only target	He expected the cabin to be much bigger.	He expected the cabin to be much bigger.	cabin	notebook	trophy	soap
Filler: noun-only target	He forgot the suspenders that he was supposed to wear.	He forgot the suspenders that he was supposed to wear.	suspenders	dragonfly	rope	glasses
Filler: noun-only target	He loved the balcony at the old apartment.	He loved the balcony at the old apartment.	balcony	gorilla	microscope	pocket

Filler: noun-only target	He preferred the mattress with the foam top.	He preferred the mattress with the foam top.	mattress	coffin	ambulance	soap
Filler: noun-only target	He preferred the suspenders to a belt.	He preferred the suspenders to a belt.	suspenders	cube	package	ruler
Filler: noun-only target	He prepared the calendar before meeting with his boss.	He prepared the calendar before meeting with his boss.	calendar	island	elbow	telephone
Filler: noun-only target	He remembered the calendar on the wall in his office.	He remembered the calendar on the wall in his office.	calendar	wagon	vulture	bottle
Filler: noun-only target	He requested the mattress that he had tried at the hotel.	He requested the mattress that he had tried at the hotel.	mattress	helicopter	toothbrush	package
Filler: noun-only target	He wanted the cabin with two bedrooms.	He wanted the cabin with two bedrooms.	cabin	mitten	ankle	bottle
Filler: noun-only target	She chose the calculator that would work for both classes.	She chose the calculator that would work for both classes.	calculator	notebook	lighthouse	swing
Filler: noun-only target	She chose the lipstick on display in the window.	She chose the lipstick on display in the window.	lipstick	stool	milk	keyhole
Filler: noun-only target	She chose the pliers from the tool box.	She chose the pliers from the tool box.	pliers	wagon	swing	lamp
Filler: noun-only target	She hated the coconut but didn't mind the pineapple.	She hated the coconut but didn't mind the pineapple.	coconut	dominoes	shrimp	bread
Filler: noun-only target	She knew the ballerina from last night's performance.	She knew the ballerina from last night's performance.	ballerina	coffin	globe	truck
Filler: noun-only target	She liked the ballerina in the first row.	She liked the ballerina in the first row.	ballerina	pineapple	kitchen	milk
Filler: noun-only target	She needed the coconut for the curry she was making.	She needed the coconut for the curry she was making.	coconut	mitten	rhino	truck

Filler: noun-only target	She needed the pliers for her next project.	She needed the pliers for her next project.	pliers	megaphone	skateboard	chameleon
Filler: noun-only target	She offered the bracelet to her friend.	She offered the bracelet to her friend.	bracelet	guitar	wheelbarrow	motorcycle
Filler: noun-only target	She offered the sweater to the shivering child.	She offered the sweater to the shivering child.	sweater	medal	cactus	bread
Filler: noun-only target	She preferred the sweater that her sister had borrowed.	She preferred the sweater that her sister had borrowed.	sweater	cube	trophy	frame
Filler: noun-only target	She requested the bracelet that matched her sister's.	She requested the bracelet that matched her sister's.	bracelet	island	violin	map
Filler: noun-only target	She wanted the calculator for her exam.	She wanted the calculator for her exam.	calculator	triangle	rhino	skateboard
Filler: noun-only target	She wanted the lipstick but was not allowed to buy it.	She wanted the lipstick but was not allowed to buy it.	lipstick	funnel	pocket	scissors
Filler: noun-only target	They hated the grapes grown at the new vineyard.	They hated the grapes grown at the new vineyard.	grapes	controller	pomegranate	telephone
Filler: noun-only target	They hated the jellyfish that were floating in the water.	They hated the jellyfish that were floating in the water.	jellyfish	sword	clarinet	map
Filler: noun-only target	They liked the grapes on the cheese plate.	They liked the grapes on the cheese plate.	grapes	earring	frame	cigar
Filler: noun-only target	They remembered the jellyfish that often appeared in the summer.	They remembered the jellyfish that often appeared in the summer.	jellyfish	pineapple	eyebrow	motorcycle
Filler: noun-only target	They remembered the saxophone that had been stolen last year.	They remembered the saxophone that had been stolen last year.	saxophone	guitar	toothbrush	rope
Filler: noun-only target	They requested the saxophone but received a clarinet.	They requested the saxophone but received a clarinet.	saxophone	gorilla	pomegranate	clock

Filler: noun-verb ambiguous target	He chose to skateboard to school.	He chose to skateboard to school.	skateboard	dragonfly	compass	magnet
Filler: noun-verb ambiguous target	He liked to swing for hours at the playground.	He liked to swing for hours at the playground.	swing	gorilla	clarinet	acorn
Filler: noun-verb ambiguous target	He loved to skateboard along the sidewalks.	He loved to skateboard along the sidewalks.	skateboard	mitten	globe	dolphin
Filler: noun-verb ambiguous target	He neglected to soap his hands thoroughly.	He neglected to soap his hands thoroughly.	soap	pineapple	tractor	fireplace
Filler: noun-verb ambiguous target	He refused to swing without someone to push him.	He refused to swing without someone to push him.	swing	island	rhino	camel
Filler: noun-verb ambiguous target	He tried to soap the baby's arms and legs.	He tried to soap the baby's arms and legs.	soap	helicopter	microscope	glasses
Filler: noun-verb ambiguous target	She chose to bread the chicken before frying it.	She chose to bread the chicken before frying it.	bread	megaphone	vulture	curtain
Filler: noun-verb ambiguous target	She chose to frame her diploma.	She chose to frame her diploma.	frame	stool	compass	dolphin
Filler: noun-verb ambiguous target	She declined to telephone the anxious client.	She declined to telephone the anxious client.	telephone	funnel	shrimp	magnet
Filler: noun-verb ambiguous target	She forgot to milk the goats.	She forgot to milk the goats.	milk	triangle	shrimp	chameleon
Filler: noun-verb ambiguous target	She liked to clock how long it took to walk to the office.	She liked to clock how long it took to walk to the office.	clock	helicopter	ambulance	ruler
Filler: noun-verb ambiguous target	She liked to milk the cows before breakfast.	She liked to milk the cows before breakfast.	milk	dragonfly	cactus	acorn
Filler: noun-verb ambiguous target	She neglected to clock her overtime hours.	She neglected to clock her overtime hours.	clock	medal	vulture	umbrella
Filler: noun-verb ambiguous target	She offered to telephone the patient's wife.	She offered to telephone the patient's wife.	telephone	stool	kitchen	needle

Filler: noun-verb ambiguous target	She tried to bread the cauliflower for the new recipe.	She tried to bread the cauliflower for the new recipe.	bread	triangle	wheelbarrow	kiwi
Filler: noun-verb ambiguous target	She tried to frame the photo for her uncle.	She tried to frame the photo for her uncle.	frame	wagon	elbow	camel
Filler: noun-verb ambiguous target	She tried to pocket three cookies from the bake sale.	She tried to pocket three cookies from the bake sale.	pocket	coffin	ankle	needle
Filler: noun-verb ambiguous target	She wanted to pocket the change from the transaction.	She wanted to pocket the change from the transaction.	pocket	earring	violin	fingerprint
Filler: noun-verb ambiguous target	They expected to bottle the beer as soon as it was ready.	They expected to bottle the beer as soon as it was ready.	bottle	notebook	eyebrow	cigar
Filler: noun-verb ambiguous target	They expected to bottle the milk all at once.	They expected to bottle the milk all at once.	bottle	funnel	elbow	keyhole
Filler: noun-verb ambiguous target	They hated to rope the cattle.	They hated to rope the cattle.	rope	earring	microscope	accordion
Filler: noun-verb ambiguous target	They needed to map the city's transit system.	They needed to map the city's transit system.	map	dominoes	trophy	apron
Filler: noun-verb ambiguous target	They offered to package the shoes for shipment.	They offered to package the shoes for shipment.	package	megaphone	cactus	lamp
Filler: noun-verb ambiguous target	They offered to truck the furniture at no cost.	They offered to truck the furniture at no cost.	truck	guitar	ankle	curtain
Filler: noun-verb ambiguous target	They preferred to motorcycle rather than walk.	They preferred to motorcycle rather than walk.	motorcycle	dominoes	toothbrush	scissors
Filler: noun-verb ambiguous target	They preferred to rope the chairs together overnight.	They preferred to rope the chairs together overnight.	rope	medal	kitchen	apron
Filler: noun-verb ambiguous target	They preferred to truck the material across town.	They preferred to truck the material across town.	truck	cube	lighthouse	accordion

Filler: noun-verb ambiguous target	They remembered to package the delicate items separately.	They remembered to package the delicate items separately.	package	controller	tractor	fingerprint
Filler: noun-verb ambiguous target	They tried to motorcycle across the country.	They tried to motorcycle across the country.	motorcycle	sword	pomegranate	kiwi
Filler: noun-verb ambiguous target	They wanted to map the route before the hike.	They wanted to map the route before the hike.	map	controller	wheelbarrow	salad

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/doi:10.3758/bf03193014

Boersma, P., & Weenink, D. (2014). *Praat: Doing phonetics by computer [Computer software]*. Retrieved from http://www.praat.org/.

Brock, J., & Nation, K. (2014). The hardest butter to button: Immediate context effects in spoken word identification. *The Quarterly Journal of Experimental Psychology*, *67*(1), 114–123. https://doi.org/10.1080/17470218.2013.791331

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. *Current Biology*, *28*(24), 3976-3983.e5. https://doi.org/10.1016/j.cub.2018.10.042

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time Course of Frequency Effects in Spoken-Word Recognition: Evidence from Eye Movements. *Cognitive Psychology*, *42*(4), 317–367. https://doi.org/10.1006/cogp.2001.0750

Dahan, D., Swingley, D., Tanenhaus, M. K., & Magnuson, J. S. (2000). Linguistic Gender and Spoken-Word Recognition in French. *Journal of Memory and Language*, *42*(4), 465–480. https://doi.org/10.1006/jmla.1999.2688

Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology*, *71*(4), 808–816. https://doi.org/10.1080/17470218.2017.1310261

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fox, N. P., & Blumstein, S. E. (2016). Top-down effects of syntactic sentential context on phonetic processing. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(5), 730–741. https://doi.org/10.1037/a0039965

Gaston, P., & Marantz, A. (2018). The time course of contextual cohort effects in auditory processing of category-ambiguous words: MEG evidence for a single "clash" as noun or verb. *Language, Cognition and Neuroscience*, *33*(4), 402–423. https://doi.org/10.1080/23273798.2017.1395466

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, *28*(4), 267–283. https://doi.org/10.3758/BF03204386

Halle, M., & Marantz, A. (1993). Distributed morphology and the pieces of inflection. In K. Hale & S. J. Keyser (Eds.), *The view from building 20* (pp. 111–176). MIT Press.

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language*, *57*(4), 460–482. https://doi.org/10.1016/j.jml.2007.02.001

Lucas, M. (1999). Context effects in lexical access: A meta-analysis. *Memory & Cognition*, *27*(3), 385–398. https://doi.org/10.3758/BF03211535

Luce, P. A. (1986). A computational analysis of uniqueness points in auditory word recognition. *Perception & Psychophysics*, *39*(3), 155–158. https://doi.org/10.3758/BF03212485

Luce, P. A., & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and Hearing*, *19*(1), 1–36. https://doi.org/10.1097/00003446-199802000-00001

Magnuson, J. S. (2019). Fixations in the visual world paradigm: Where, when, why? *Journal of Cultural Cognitive Science*, *3*(2), 113–139. https://doi.org/10.1007/s41809-019-00035-3

Magnuson, J. S., Tanenhaus, M. K., & Aslin, R. N. (2008). Immediate effects of form-class constraints on spoken word recognition. *Cognition*, *108*(3), 866–873. https://doi.org/10.1016/j.cognition.2008.06.005

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1), 71–102. https://doi.org/10.1016/0010-0277(87)90005-9

McAllister, J. M. (1988). The use of context in auditory word recognition. *Perception & Psychophysics*, *44*(1), 94–97. https://doi.org/10.3758/BF03207482

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017, August). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Proceedings of the 18th Conference of the International Speech Communication Association*.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189–234. https://doi.org/10.1016/0010-0277(94)90043-4

Rayner, K., Slowiaczek, M. L., Clifton, C., & Bertera, J. H. (1983). Latency of sequential eye movements: Implications for reading. *Journal of Experimental Psychology: Human Perception and Performance*, *9*(6), 912–922. https://doi.org/10.1037/0096-1523.9.6.912

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's Object Pictorial Set: The Role of Surface Detail in Basic-Level Object Recognition. *Perception*, *33*(2), 217–236. https://doi.org/10.1068/p5117

Salverda, A. P., Kleinschmidt, D., & Tanenhaus, M. K. (2014). Immediate effects of anticipatory coarticulation in spoken-word recognition. *Journal of Memory and Language*, *71*(1), 145–163. https://doi.org/10.1016/j.jml.2013.11.002

Seidenberg, M. S., Tanenhaus, M. K., Leiman, J. M., & Bienkowski, M. (1982). Automatic access of the meanings of ambiguous words in context: Some limitations of knowledge-based processing. *Cognitive Psychology*, *14*(4), 489–537. https://doi.org/10.1016/0010-0285(82)90017-2

Shillcock, R. C., & Bard, E. G. (1993). Modularity and the processing of closed-class words. In G. Altmann & R. C. Shillcock (Eds.), *Cognitive models of speech processing: The second sperlonga meeting* (pp. 163–185).

Simonsohn, U. (2015). [17] No-way Interactions. *The Winnower*. https://doi.org/10.15200/winn.142559.90552

Strand, J. F., Brown, V. A., Brown, H. E., & Berg, J. J. (2018). Keep listening: Grammatical context reduces but does not eliminate activation of unexpected words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(6), 962–973. https://doi.org/10.1037/xlm0000488

Strauss, T. J., Harris, H. D., & Magnuson, J. S. (2007). jTRACE: A reimplementation and extension of the TRACE model of speech perception and spoken word recognition. *Behavior Research Methods*, *39*(1), 19–30. https://doi.org/10.3758/BF03192840

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, *18*(6), 645–659. https://doi.org/10.1016/S0022-5371(79)90355-4

Tanenhaus, M. K., & Donnenwerth-Nolan, S. (1984). Syntactic context and lexical access. *The Quarterly Journal of Experimental Psychology Section A*, *36*(4), 649–661. https://doi.org/10.1080/14640748408402184

Tanenhaus, M. K., Leiman, J. M., & Seidenberg, M. S. (1979). Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of Verbal Learning and Verbal Behavior*, *18*(4), 427–440. https://doi.org/10.1016/S0022-5371(79)90237-8

Tanenhaus, M. K., & Lucas, M. M. (1987). Context effects in lexical processing. *Cognition*, *25*(1–2), 213–234. https://doi.org/10.1016/0010-0277(87)90010-2

Tyler, L. K. (1984). The structure of the initial cohort: Evidence from gating. *Perception & Psychophysics*, *36*(5), 417–427. https://doi.org/10.3758/BF03207496

Tyler, L. K., & Wessels, J. (1983). Quantifying contextual contributions to word-recognition processes. *Perception & Psychophysics*, *34*(5), 409–420. https://doi.org/10.3758/BF03203056

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151–175. https://doi.org/10.1016/j.jml.2018.07.004

Vigliocco, G., Vinson, D. P., Druks, J., Barber, H., & Cappa, S. F. (2011). Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience & Biobehavioral Reviews*, *35*(3), 407–426. https://doi.org/10.1016/j.neubiorev.2010.04.007

Weide, R. (1994). *CMU pronouncing dictionary*. http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition*, *32*(1), 25–64. https://doi.org/10.1016/0010-0277(89)90013-9