ABSTRACT

Title of Dissertation:	THE ROLE OF SYNTACTIC PREDICTION IN AUDITORY WORD RECOGNITION
	Phoebe Gaston Doctor of Philosophy, 2020
Dissertation directed by:	Professor Colin Phillips & Professor Ellen Lau Department of Linguistics

Context is widely understood to have some influence on how words are recognized from speech. This dissertation works toward a mechanistic account of how contextual influence occurs, looking deeply at what would seem to be a very simple instance of the problem: what happens when lexical candidates match with auditory input but do not fit with the syntactic context. There is, however, considerable conflict in the existing literature on this question. Using a combination of modelling and experimental work, I investigate both the generation of abstract syntactic predictions from sentence context and the mechanism by which those predictions impact auditory word recognition.

In the first part of this dissertation, simulations in jTRACE show that the speed with which changes in lexical activation can be observed in dependent measures should depend on the size and composition of the set of response candidates

allowed by the task. These insights inform a new design for the visual world paradigm that ensures that activation can be detected from words that are bad contextual fits, and that facilitatory and inhibitory mechanisms for the syntactic category constraint can be distinguished. This study finds that wrong-category words are activated, a result that is incompatible with an inhibitory syntactic category constraint.

I then turn to a different approach to studying lexical activation, using information-theoretic properties of the set of words consistent with the auditory input while neural activity is recorded in MEG. Phoneme surprisal and cohort entropy are evaluated as predictors of the neural response to hearing single words when that response is modeled with temporal response functions. This lays the groundwork for a design that can test different versions of surprisal and entropy, incorporating facilitatory or inhibitory syntactic constraints on lexical activation when the stimuli are short sentences.

Finally, I investigate a neural effect in MEG previously thought to reflect syntactic prediction during reading. When lexical predictability is minimized in a new study, there is no longer evidence for structural prediction occurring at the beginning of sentences. This supports the possibility of a tighter link between syntactic and lexical processing.

THE ROLE OF SYNTACTIC PREDICTION IN AUDITORY WORD RECOGNITION

by

Phoebe Elizabeth Gaston

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor Colin Phillips, Co-Chair
Professor Ellen Lau, Co-Chair
Professor Naomi Feldman
Professor William Idsardi
Professor Jan Edwards, Dean's Representative

Foreword

This work was supported by the National Science Foundation under grants DGE-1449815 and BCS-1749407.

Acknowledgements

There is no way that I can adequately express all of my gratitude here for the many people who have been a part of my life in Maryland, but I'll do my best.

Before I even got to grad school, I had had so many incredible teachers and advisors who helped me find this path. Max Gabrielson, who as my high school Latin teacher also taught me to read Homeric Greek early in the mornings before school, was the person who first sparked my fascination with the structure of language. Raffaella Zanuttini, Bob Frank, and Maria Piñango showed me how being fascinated with the structure of language could be a career. They taught me linguistics at Yale, gave me my first research experiences, and encouraged me at every step. Lars Meyer was amazingly devoted as a mentor, in a way that I very much appreciated at the time but didn't know was quite unusual until much later. He taught me many things that turned out to be crucial for the work I would go on to do (including the importance of standardized file structures, and how to write a for loop), and he gave me a lot of confidence. Alec Marantz took me on as a lab manager after I graduated, and working with him has had a massive influence on the direction of my research. He always seemed to trust that I knew what I was doing, which sometimes felt bewildering but gave me a lot of room to learn and think. The lab structure and community that Alec Marantz and Liina Pylkkänen have built has always been an inspiration.

Since the moment I arrived at UMD, Colin and Ellen have been my unwavering advocates. I am incredibly lucky to have had both of them as advisors. Thank you to Colin for re-explaining to me why my research was interesting every time I lost faith (honestly, that must have gotten old). Thank you for caring about me as a whole

person. Thank you for letting me spend my first 2.5 years here mostly just thinking, and for not telling me that you thought the problem I wanted to pursue was already solved. Thank you for taking me seriously at the LSA Summer Institute in 2011 when I knew so little. Thank you for the many, many hours of discussion. Thank you for convincing me at the beginning of my second year that I should stick it out, and things would get better. Thank you for working late into the night on my drafts, and also somehow not creating the expectation that I work late into the night on my drafts. Thank you for not letting me off the hook on the important things.

Thank you to Ellen for thinking through every single step of every idea with me. I could never be sure if anything really, truly made sense unless we'd been through it together. Thank you for stunningly thoughtful, useful, constructive feedback on everything I wrote. Thank you for making real life just as important as research. Thank you for laughing at me when I was being ridiculous. Thank you for being there for me even from Paris. Thank you for validating my frustrations. Thank you for understanding when I was struggling, without us having to talk about it. Thank you for your magical ability to listen to my vague conjectures and repeat them back as ideas that make sense. Thank you for being realistic above all else, and helping me focus my attention on what actually mattered. Thank you for the discussions we had just because they were interesting and not because they were going to lead anywhere.

Thank you to Naomi Feldman, Jan Edwards, and Bill Idsardi for making up the rest of my committee. I'm really grateful for the perspectives you added.

Thank you to my brilliant, funny, caring cohort: Paulina Lyskawa, Max Papillon, Suyoung Bae, Chia-Hsuan Liao, and Annemarie van Dooren. I couldn't have chosen

a better group to go through the PhD with. Thanks for all the fun and for the dozens of ways we were there for each other over the years. I'm sorry that we couldn't spend our last few months in Maryland together.

During my first year in the PhD program, attending CNL lab meetings on Fridays, I became convinced I was going to have to drop out rather than eventually present. I thought there was no way I could stand up in front of that group. Thank you all for proving me wrong. Thank you for engaging so intensely and caring so fiercely. Thank you to those of you who said nice things after my "surprise 888" and gave me hope that my ideas might work out.

Thank you to Jeff Lidz and Alexander Williams for their consistently insightful and supportive commentary over the years. Thank you, similarly, to both the real Omer Preminger and the imaginary Omer who made comments about what a word is while I wrote my talks. Thank you to Norbert Hornstein for caring in the way that he has. Thank you to Peggy Antonisse and Tonia Bleam for teaching me how to teach. Thank you to all of the faculty in our department for rooting for me, and all of us, so genuinely. I learned a lot from each and every one of you.

Thanks to the daily lunchroom crowd for the entertainment, distraction, and camaraderie.

Thanks to those of you who humored me in showing up for Rejection Club. Those discussions helped me a lot and I hope they helped you too.

Thank you to William Matchin, Shota Momma, Zoe Schlueter, Anton Malko, Allyson Ettinger, Lara Ehrenhofer, Laurel Perkins, Nick Huang, Mina Hirzel, Tyler

Knowleton, Jackie Nelligan, and Adam Liter for being wonderful role models, colleagues, and friends.

Thank you to Kim Kwok, for solving every problem and always being so kind. Thank you to Anna Namyst for the very patient MEG guidance. Thank you to Maggie Kandel and Lalitha Balachandran for their help in many practical matters.

Thanks to Nick Huang and Jackie Nelligan for being the best possible project-mates in Computational Linguistics.

Thank you to Michael Robinson and Robert Metzger, the physical therapists who were crucial in making it possible for me to write a dissertation and exist in academia despite having limited use of my dominant arm a lot of the time.

Thank you to everyone who, when I was sick for much of 2019, cooked food I could actually eat, never made me feel like a burden, and helped me feel like myself even when I felt terrible. Thank you to Colin and Ellen for understanding how limited my capacity sometimes was, never making it a problem, and reminding me that my value as a person goes far beyond how much work I can get done in a single day or week or month.

Thank you to the Language Science Center for broadening my horizons and for the many human connections I would not otherwise have made (and also, for funding me and my work). Thank you to those of you who were part of LSLT and Winter Storm. I have really valued being part of this larger community. Thank you to Shevaun Lewis, Caitlin Eaves, and Tess Wood for their roles in the LSC.

Thanks to Laurel Perkins, Kasia Hitczenko, and then Hanna Muller for making our office such a compassionate, inspiring, productive, and also laughter-filled environment. I was so happy to see you every day.

Thank you to the Language Science Gorilla for the honor of your presence at both my final LSLT and my defense celebration.

Thank you to Aura Cruz Heredia for making those long days at the MNC bearable, for dropping everything when I needed help in any realm of work or life, for surprising me over and over again with chai from Vigilante, and for unnecessarily treating me like a star.

Thank you to the students in my LING 449 seminar, who showed me that I do love teaching after all. Thank you to Nadiya Klymenko and Reese Alpher for their hard work as research assistants, and for helping me (start to) learn how to be a mentor. Thanks also to Oliver Bentham, Stephanie Pomrenke, Fen Ingram, Daphne Amir, and Macie McKitrick for their help creating and running experiments.

Thank you to Hanna Muller, Jon Burnsky, Zoe Ovans, Tyler Knowleton, Maggie Kandel, Lalitha Balachandran, Chia-Hsuan Liao, and Aura Cruz Heredia for being extremely fun CUNY house-mates and sharing in my greatest Airbnb find of all time.

Thank you to Anton Malko, Hanna Muller, Adam Liter, Kathleen Oppenheimer, and Ian Phillips for the statistics group discussions, which I enjoyed immensely.

My inexpressible gratitude to Paulina Lyskawa, who was always there for me, Christian, and Finn, especially at the very worst moment. I think Paulina is the most generous person I know.

Thank you to the Bias in Linguistics working group, for teaching me so much about the world, providing an outlet for my anger about gender bias, and giving me an opportunity to do an entirely different kind of research. I'm so proud of what we've done together.

Thank you to Hanna Muller and to Kasia Hitczenko for the unconditional friendship, for allowing me to vent all the time, for boosting me higher in the good moments, and for the many hours of conversation and companionship. I get to be my whole self when I'm with you.

Thank you to Jim Magnuson and Andrea Martin, whose extraordinary kindness and desire to help and listen while I was on the job market, even though they barely knew me, was almost unfathomable. I hope to be able to do that for someone else someday. I was reminded again and again throughout this process that we're all just human in an imperfect system.

Thank you to the quarantine version of our lab check-in group for making these last few months of isolation a lot more bearable: Colin Phillips, Hanna Muller, Masato Nakamura, Rosa Lee, Chia-Hsuan Liao, and Cassidy Wyatt. I certainly didn't expect to write the second half of my dissertation at my kitchen table, under lockdown, but you helped a lot.

Thank you to Hanna Muller, Aura Cruz Heredia, and Jackie Nelligan for the most intensely affirmational group text thread I've ever been a part of.

Thank you to Finn, who was the most perfect companion for five and a half years, and who I miss every day. Thank you to Moose and Wilson for showing me how life goes on after so much sadness, in their own silly, joyful, loving ways. Thank

you to everyone in our lives who has helped make Wilson's medication schedule work—not minding when I cancelled meetings or left early or asked to meet for dinner at strange times. Thank you to Friendship Hospital for Animals for getting us through every emergency. Thank you to Amy Rogers for making it possible for me and Christian to stay on campus for afternoon meetings.

Thank you to the Hyattsville crew both narrowly and loosely defined (Anton Malko, Allyson Ettinger, Lara Ehrenhofer, Kasia Hitczenko, Paulina Lyskawa, Natalia Lapinskaya and many others, at various points) for the dinners, hangouts, brunches, cake-baking, game nights, egg hunts, kayak adventures, and increasingly elaborate surprise parties. I have been so lucky to have so many amazing friends within a 10 minute walk of my house.

Thank you to Love Yoga in Hyattsville and Numi Yoga in College Park for giving me another community, many friendships, and a lot of strength. Thanks in particular to Monica Corry, Michael Levin, Asia Vianna, Jan Edwards, Sara Crosby, CiCi English, and Rachel Debuque (who proved to me I had been strong enough to do headstands all along, probably a good metaphor for many other things). Thank you to Ken Carter for creating my favorite neighborhood spot. Hyattsville has been a wonderful place to live and I will miss it dearly.

Thank you to Nihal Kayali, Courtney Kaplan, Sophie Broach, and Sinead O'Brien for being there for the good, the bad, the all of it in the last eleven years, and for knowing me deeply throughout.

Thank you to my parents and my grandparents, for making this possible at all and for believing in me no matter what. Thank you to Zoë, Keiley, and Sophie for

being the best friends and support crew you could imagine growing up with. I'm so lucky to have the family that I do.

And finally, thank you to Christian for being my partner in every way, for making my dreams just as important as his, for serving as my in-house MEG consultant, and for cooking dinner most of the time. I know that as a team we can handle anything.

Table of Contents

List of Tables

List of Figures

Chapter 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). 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 before the end of the word (Luce, 1986). However, the likelihood of any given lexical candidate 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. The aim of this dissertation is to develop a mechanistic model of exactly how sentence context impacts the word recognition process, using computational and experimental tools.

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 much of this dissertation, I explore these issues specifically as they relate to syntactic category expectations and the mechanism by which those expectations influence auditory word recognition. Using syntactic context provides us with quite 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? In this dissertation, I 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 still be

a context effect but would simply boost 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. The previous literature on the timing and nature of syntactic context effects, reviewed in the next section, is mixed in its conclusions. Abandoning the assumption that a top-down constraint must be inhibitory (which is pervasive, though often not expressed in these terms) makes the puzzling variation in previous findings easier to account for. Consideration of the linking hypotheses between underlying and observed measures, as I explore in **Chapter 2**, may also aid in unifying this literature.

This area 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 strong 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 (Halle & Marantz, 1993) view of the lexicon, 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 my 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, I 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 dissertation may at times seem to indicate an assumption of stored syntactic category information for wordforms, referring to items that can or cannot be nouns, I do not intend for this to be a theoretical stance.

Finally, many questions remain as to how exactly the syntactic parse arises, that 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, I 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. To that end, in **Chapter 5** I describe efforts to isolate syntactic from lexical prediction in sentence processing.

1.2 Syntactic constraints on word recognition: a methodological conflict

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, I 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.

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 very 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 original three experiments (Seidenberg et al., 1982; Tanenhaus et al., 1979; Tanenhaus & Donnenwerth-Nolan, 1984) probed only as early as the acoustic offset of the ambiguous prime word. The ideal method for this question would probe continuously and as early as acoustic information begins to be processed.

The lack of evidence for a syntactic category constraint could also 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, which 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 final 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. Zwitserlood (1989) successfully uses this approach to ask about effects of semantic rather than syntactic context.

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 we would expect to see incorrect-category guesses at that sampling point as well, and it would not be evidence against a context effect. 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 (1984)'s 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) have proposed an explicit correspondence between observed fixation probabilities in the visual world and lexical activation of the words corresponding to the pictures. One clear advantage of the VWP over the methods previously described is that fixations can be recorded continuously from the onset of the word, providing an immediate and incremental measure 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 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 are using 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.)), people never fixate gender-inconsistent cohort competitors (e.g., “bouteille” (fem.)) more than irrelevant distractors (e.g. “table”). This appears to be clear evidence that the gender cue stops listeners from activating the wrong-gender match.

Dahan et al’s finding is relevant to the syntactic question, but it is not obvious that grammatical gender and syntactic category operate identically within the architecture of the word recognition system. In fact, different results for these two kinds of contextual information have been observed in gating. Tyler (1984) found that competitors of the wrong syntactic category are indeed produced, while Grosjean, Dommergues, Cornu, Guillelmon, and Besson (1994) find that competitors of the wrong gender are never produced.

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 a novel language composed of texture adjectives and shape nouns, all of which describe 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 consist of two items that share one shape and two items that share another shape, an adjective is necessary to uniquely specify any one of the items, and participants should therefore expect an adjective after “Click on the...”.

However, when the display contains four distinct objects, an adjective is not necessary, and participants should therefore expect a noun after "Click on the...". When the participant is expecting an adjective, Magnuson et al. (2008) show increased fixations to the shape whose adjective descriptor is a cohort competitor of the target adjective (e.g., to the "bupe tedu" when "bupʌ pibe" is the target), but shapes whose noun descriptor is a cohort competitor of the target adjective (e.g. the "tedɛ bupo") do not draw increased fixations. Similarly, when the participant is expecting only a noun, they show increased fixations to the shape whose noun descriptor is a cohort competitor of the target noun, but shapes whose adjective descriptors are cohort competitors of the target noun do 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 this 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 I will discuss in **Chapter 3**, there may be a potential strategy available to participants that could conceal competition from wrong-category candidates, if it were occurring. There is also the obvious caveat that these results occur in an artificial language in which participants have 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-splice 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 I present in **Chapter 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 (as reviewed in **Chapter 4**).

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 note that it is impossible to say whether this supports the idea that the *generation* of candidates is being constrained by category. We can only conclude 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, which would make it more likely that Gaston and Marantz (2018)'s findings reflect an immediate constraint.

1.2.5 Summary of the literature

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 dissertation, I 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. This requires close examination of the measures we use we 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 Dissertation outline

Chapter 2 describes simulations with the jTRACE model of auditory word recognition (Strauss et al., 2007) showing the potential influence of an experimental task's response candidate set on how quickly changes in lexical activation can translate to changes in response probability. **Chapter 3** reports a visual world experiment that provides evidence that wrong-category syntactic candidates do compete during word recognition, inconsistent with an inhibitory constraint. **Chapter 4** investigates MEG effects of phoneme surprisal and cohort entropy during auditory single-word recognition, laying the groundwork for a new study of context effects on the cohort whose design I also describe. **Chapter 5** reports a MEG study that fails to find evidence for syntactic prediction in the absence of lexical predictability. **Chapter 6** presents summary, discussion, and future directions.

Chapter 2: Simulations on the relation between lexical activation and experimental measures

2.1 Introduction

In **Chapter 1**, we described a conflict in the results from different methods used to investigate syntactic context effects on cohort competition. In this chapter, we argue that one factor in accounting for these conflicting results may be the difference between *forced-choice* and *free-choice* tasks. The limited set of available referents in the visual world paradigm leads to a forced-choice task in which the dependent measure is a proportion of fixations among a small number of referents on screen. With so few response candidates, it may be that bottom-up auditory information and/or contextual constraints can be acted on earlier than in a free-choice task like gating where participants are free to answer with any word from the lexicon. The size and composition of that *response candidate set* are therefore important properties of an experiment.

We use jTRACE (Strauss et al., 2007) to model forced-choice and free- choice scenarios and show that, indeed, bottom-up effects of cohort membership occur earlier when there are fewer response candidates available. This occurs under nearly all potential combinations of the parameter settings we manipulated, including different implementations of frequency. However, we note that the timing difference is seen in the simulated dependent measure (response probability), and *not* in the underlying simulated activation levels that are the true quantity of interest. The

activation levels are the same regardless of the task. If top-down influences of category also lead to timing differences in cohort effects, it could mean that gating and visual world results are not actually in conflict.

It is not yet possible to directly simulate top-down category effects in jTRACE, as syntactic category is not represented in the model. As a first step in this direction, we simulate these word recognition scenarios using two different frequency distributions for the lexicon, meant to approximate category-restricted and unrestricted comprehension. These results further substantiate the distinction between forced-choice and free-choice paradigms.

2.2 TRACE model of speech perception

TRACE is a well-known connectionist model of spoken word recognition (McClelland & Elman, 1986). 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 **Figure 1**, we reproduce Strauss et al. (2007)'s schematic of the TRACE architecture. 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 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

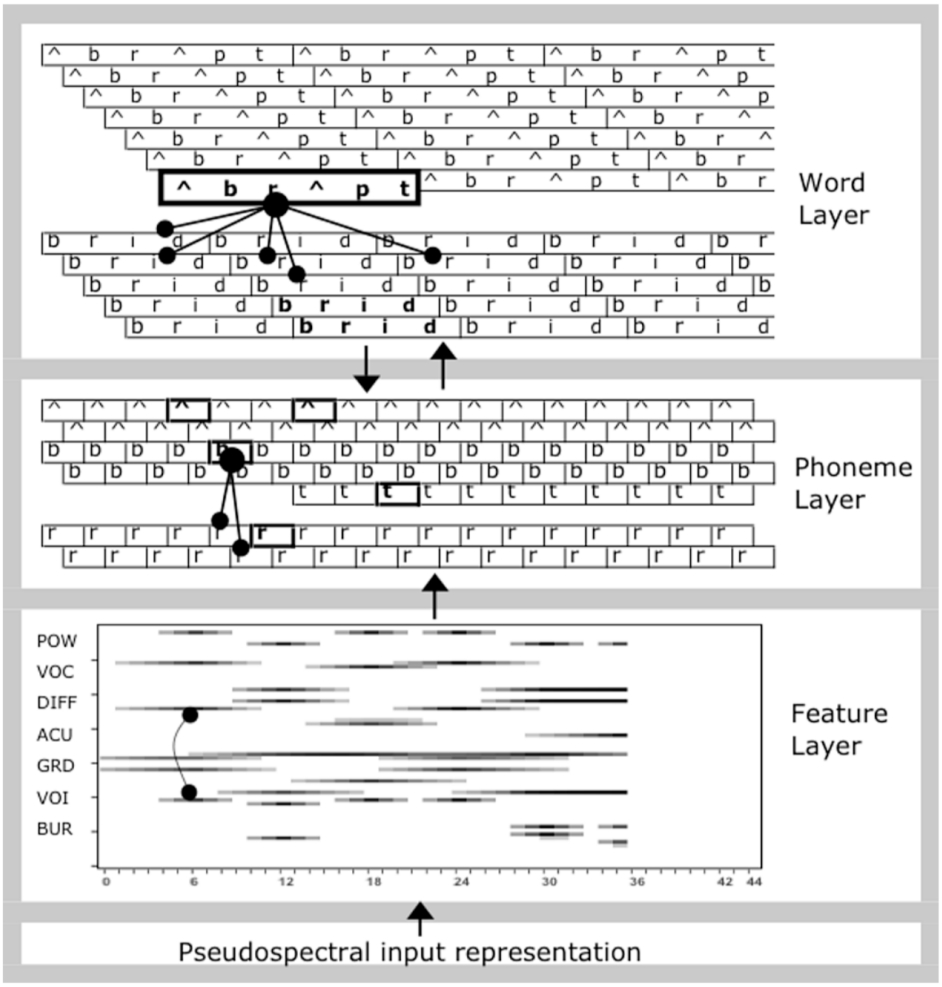


Figure 1. *Strauss et al. (2007)'s Figure 1 illustrates the layers and connections in the TRACE architecture. Arrows indicate excitatory connections, while lines ending in circles indicate within-layer inhibitory connections. Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Behavior Research Methods, jTRACE: A reimplementation and extension of the TRACE model of speech perception and spoken word recognition, Strauss, Harris, and Magnuson, Copyright 2007 Psychonomic Society, Inc., 2007.*

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 and response probabilities for all of the words in its lexicon.

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.

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

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 in this chapter. 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.

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

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. However, a difference in response *probability* for that word 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 will be 25%, while in a free choice from the whole 100-word lexicon it will 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.

2.3 Forced-choice task effects

The first question we investigated with jTRACE was 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.

2.3.1 Simulation details

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. They provide a template that includes the stimuli used in Dahan et al. (2001)'s original experiment, as well as the lexicon used in their TRACE simulation. This makes it possible to simulate each of the 17 trials and then average their results for comparison with the averaged human data and averaged simulations presented by Dahan et al. (2001).

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. 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, and then the cycle at which the

response probability for the high-frequency competitor begins to decrease. We consider the distractor divergence to be the first point at which bottom-up information observably shapes the cohort, and this is our primary concern. However, the high-frequency competitor's divergence point 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. Dahan et al. (2001) tested for differences in fixation probability averaged over the 0-200 ms and 200-500 ms time windows within a trial, which yielded a rather coarse temporal resolution. However, the figures they provide allow rough estimates of the divergence points in their data, for comparison.

2.3.2 Frequency & activation

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, 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) find 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. This 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.

2.3.3 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 activation

[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. 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, the correspondence between lexical activation and fixations to pictures in the visual world paradigm is understood to be driven by the task at hand (usually, to click on or look at the intended referent, sometimes in order to answer a question about it). In the absence of linguistic input and therefore in the absence of a task, a linking hypothesis between resting activation levels and fixations is less obvious, and potentially neutralizes the expected difference between the resting-level and connection-weight implementations of frequency.

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

Table 1. *Cycle of divergence for distractor (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 of divergence for high-frequency cohort competitor.*

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, and in all conditions the expected frequency effect on cohort competition was observed (more cohort competition from the high-frequency competitor). 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. For the high-frequency competitor, the divergence point was earlier for forced

choice relative to free choice in all parameter combinations, by 10-14 cycles (~120-168 ms.). This raises the possibility that the isolation of the forced-choice/free-choice difference to one condition for the distractor might follow from some other property of the model affecting early response strength, which obscures a forced-choice advantage except when the standard baseline default and resting-level frequency implementation are in place.

In **Figure 2** we show the underlying response strengths that are transformed into response probability for the condition that showed a free/forced-choice difference for the distractor (standard default baseline and resting level frequency), to make the overall dynamics of the trial clear. Recall that response strength is underlyingly the same, regardless of the task. In the forced-choice task, only the response strengths of the four plotted items factor into response probability. In the free-choice task, it is the response strengths of the four plotted items as well as the rest of the lexicon. After the first cycle in **Figure 2**, 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 is increasing for all items from the very beginning of the trial because 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 all four are increasing here because they are relatively higher frequency in the lexicon. In any case, the decrease in response probability that we observe for the distractor is much earlier than cycle 24, even though response strength is still increasing**; Figure 3**

shows the point in the trial at which response probability for the distractor begins to decrease in either task (cycle 10 for forced choice and cycle 16 for free choice). To understand how this pattern arises from the response strength, we must look at the relative rate of increase in response strength from cycle to cycle for the four items in either task.

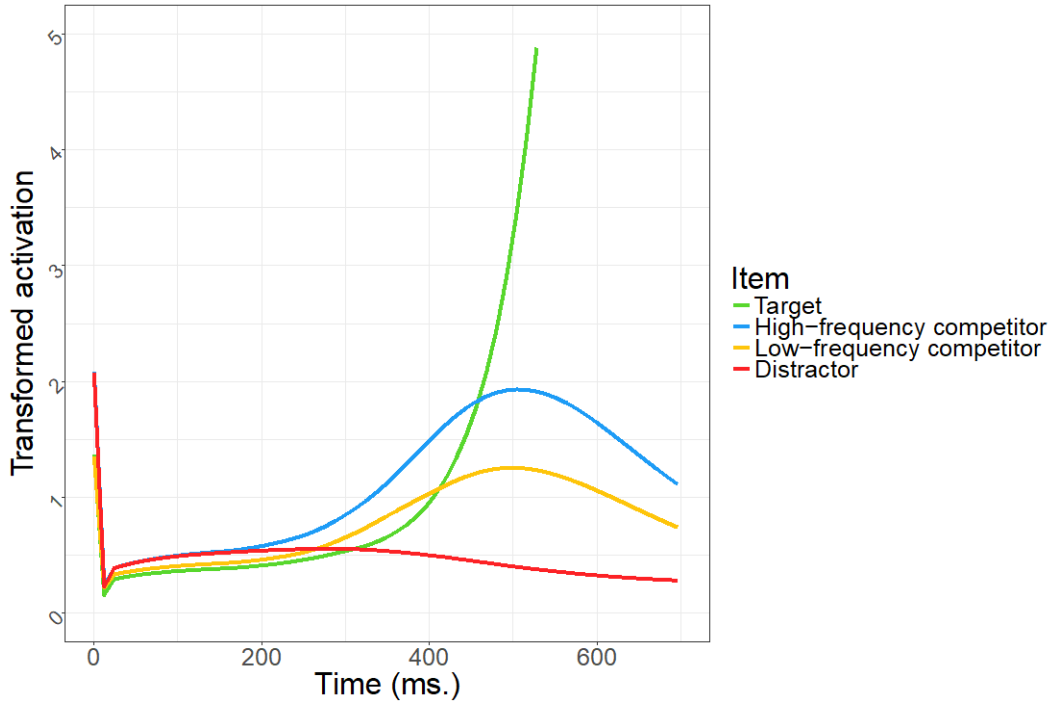


Figure 2. *The time course of underlying response strength (transformed activation) in each condition of the first simulation using the resting level frequency implementation and standard default resting baseline.*

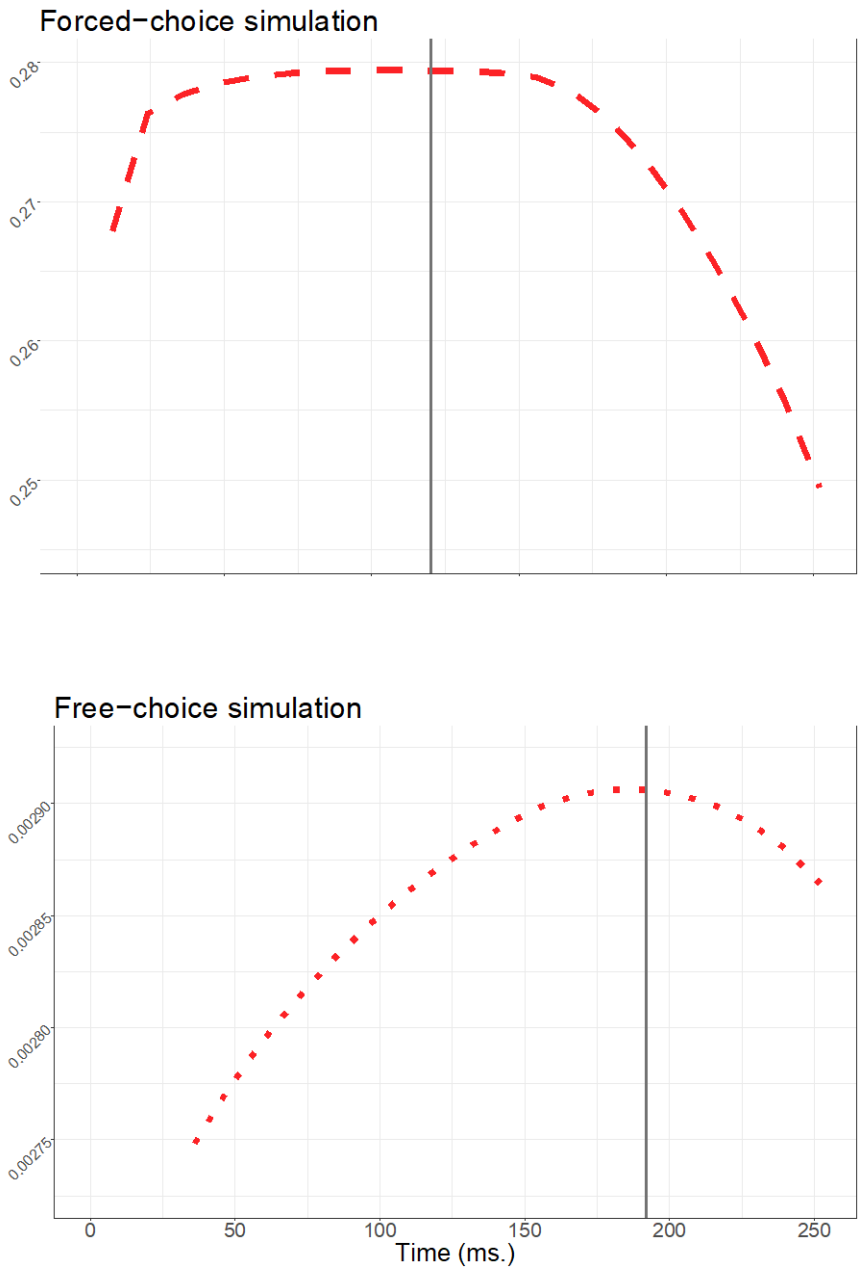


Figure 3. *The time course of response probability for the distractor in forced-choice (top) and free-choice (bottom) tasks in the resting level frequency/standard default resting baseline condition, marking the time that response probability begins to decrease.*

Response probability for the distractor is the response strength of the distractor divided by the summed response strengths of the forced-choice set of four or summed response strength of the free-choice lexicon. Each of these three quantities is increasing at the beginning of the trial (as described above), but at different rates. In the forced-choice trials, there are three items consistent with the first phoneme, and one distractor item that is not. In the free-choice trials, there are at least three items consistent with the first phoneme, as well as 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 percentage increase in summed response strength for the forced-choice trial is higher than the percentage increase in the summed response strength for the free-choice trial, because a higher proportion of items in the forced-choice set are experiencing a boost.

Whenever the percent increase in the response strength of the distractor from cycle to cycle is larger than the percent increase in the summed response strength of the response set, response probability for the distractor will also increase; if the percent increases are equal, response probability will stay the same from cycle to cycle (even though response strength is increasing). The point at which the percent increase in the response strength for the distractor (the numerator) becomes smaller than the percent increase in the summed response strength for the response set (the denominator) is the point at which the response probability for the distractor begins to decrease; this will occur at some point following the boost in activation that applies to the cohort competitors and not the distractor. Because the summed response strength in the forced-choice trials (the forced-choice "pool") is increasing proportionally

much faster (due to its composition) than the summed response strength in the free-choice trials (the free-choice "pool"), it surpasses the percent change in response strength of the distractor more quickly, and therefore forced-choice trials show an earlier divergence of the distractor, as illustrated in **Figure 4**. This is not due strictly to the smaller size of the response candidate set: it is due to the response candidate set being disproportionately composed of items whose first phoneme is consistent with the input. This means that if the forced-choice candidate set could have the same distribution of first phonemes that the overall lexicon had, the timing of the distractor divergence would not be expected to differ from the free-choice task.

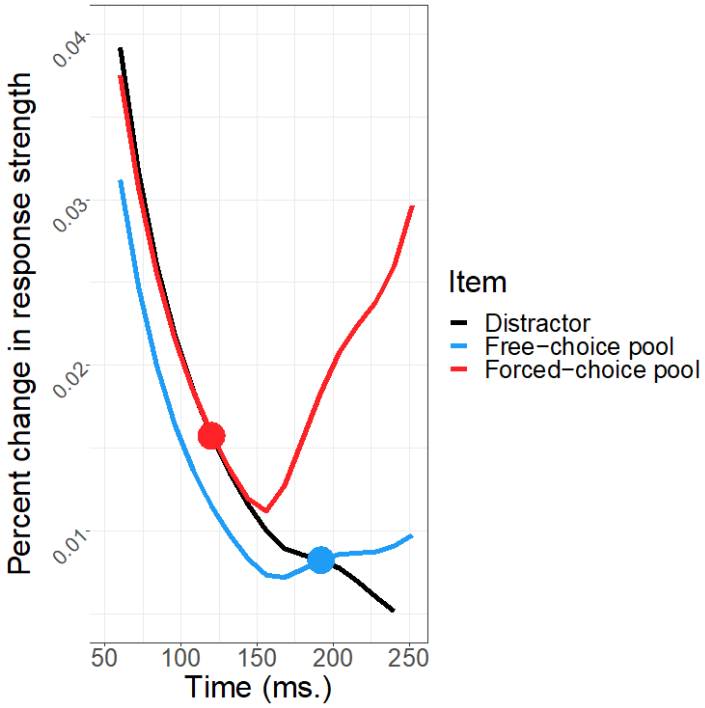


***Figure* 4.** *The time course of percent change in response strength for the distractor, free-choice pool, and forced-choice pool in the resting level frequency/standard default resting baseline condition. Dots indicate the points at which the percent change in response strength for the forced-choice or free-choice pool surpasses the percent change in response strength for the distractor.*

A final implication of the free/forced manipulation in these simulations holds for any threshold-based account of lexical access. In a forced-choice scenario under the standard baseline/resting-level frequency settings, the response probability for the target first crosses 50%, for example, at cycle 43. In a free-choice scenario, this occurs at cycle 68, a predicted difference of ~300 ms. In Dahan et al. (2001)'s forced-choice human data, the target crosses the 50% response probability threshold at roughly 550 ms (which would be cycle 46 in a simulation, remarkably close to the model's prediction of cycle 43).

2.3.4 Summary

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. We found some evidence in favor of this idea. When looking at how quickly response probability was diminished for a distractor whose first phoneme does not match the input, we found that under certain parameter settings (standard default baseline activation, and a resting-level implementation of frequency), this impact occurred six cycles (~72 ms) earlier for a forced-choice relative to a free-choice scenario. When looking at how quickly response probability was diminished for a cohort competitor with an inconsistent third phoneme, we found that this occurred 120-168 ms earlier in forced-choice relative to free-choice scenarios across all parameter settings. These results suggest that task could be an important determinant of the observed timing of influences (whether top-down or bottom-up) on cohort competition.

2.4 Category effects

The next set of simulations was intended to explore the effect of a syntactic category restriction on the distinction between forced and free-choice scenarios. Ideally this would be implemented through a substantial change to the architecture of the TRACE model itself, by adding a category layer such that category expectations could impose top-down influence on word activations. However, as any modification to the TRACE/jTRACE architecture is a complex endeavor with cascading ramifications, we adopted a rough approximation in the current study that allowed us to explore this question with the existing model. Category restriction was implemented as a manipulation of the frequencies associated with each lexical item, such that each item's activation was weighted by its frequency of occurrence in the expected category. Therefore the type of category restriction we approximated is necessarily one that applies immediately, and prior to the onset of the target, such that its effects on the lexicon are already complete. As in the previously described simulations, Experiment 1 from Dahan et al. (2001) was used as a starting point. The design, of course, did not include a condition that would allow assessment of whether a category-inappropriate cohort competitor would initially be considered a candidate or not. Instead, the goal was to establish a rough set of expectations regarding the potential influence of a constrained lexicon on the basic dynamics of a trial. Though we consider this set of simulations exploratory, one prediction is that cohort effects might manifest more quickly in a category-constrained lexicon, as an instance of context speeding word recognition. For now, we are concerned only with the effects

of the category constraint within our simulations, and not with their mapping on to human data

2.4.1 Simulation details

Investigating the frequency counts used by Dahan et al. (2001) and provided in jTRACE revealed several inconsistencies. Both measures were obtained from Francis and Kučera (1982), but for some words the counts were for noun usage only, while for others the counts were collapsed across categories. To remedy this, overall frequency counts for each word in the simulation lexicon were extracted from the SUBTLEX-US database (Brysbaert & New, 2009), along with noun-specific counts (Brysbaert et al., 2012a). Two lexicons were then created: one with overall frequency counts and one with noun-specific counts.

2.4.2 Results

The same 2x2x2 manipulation of free/forced-choice scenario, resting-level and connection-weight frequency, and standard and reduced baseline default activation was then evaluated via simulation, using the noun-specific lexicon and then the general lexicon. The results for the distractor are in **Table 3**, and for the high-frequency competitor are in **Table 4**. We note that in **Table 3**, the divergence point for the distractor in the -0.3 default baseline/resting level condition, with a general lexicon, is unaccountably early.

Table 3. *Cycle of divergence for distractor.*

	Default baseline for resting activation:	*-0.01*		*-0.3*	
		Resting-level frequency	Connection-weight frequency	Resting-level frequency	Connection-weight frequency
Noun-specific lexicon	Forced-choice task	13	14	14	14
	Free-choice task	18	14	14	14
General lexicon	Forced-choice task	3	14	1	14
	Free-choice task	17	14	1	14

Table 4. *Cycle of divergence for high-frequency cohort competitor.*

	Default baseline for resting activation:	*-0.01*		*-0.3*	
		Resting-level frequency	Connection-weight frequency	Resting-level frequency	Connection-weight frequency
Noun-specific lexicon	Forced-choice task	37	34	31	32
	Free-choice task	54	44	41	40
General lexicon	Forced-choice task	35	34	30	32
	Free-choice task	50	44	41	40

Otherwise, as in the previous simulations, we saw a forced-choice advantage in all conditions for the high-frequency competitor, and in the standard default baseline/resting level frequency condition for the distractor. In the current simulations, it was also this specific parameter combination that showed a difference due to the change of lexicon. What we observe, for these scenarios, is that the

divergence points are actually somewhat slower with the noun-specific lexicon than the general lexicon (also see **Figure 5** and **Figure 6**). This is surprising given prevailing notions that context speeds and aids in word recognition. Our simulations

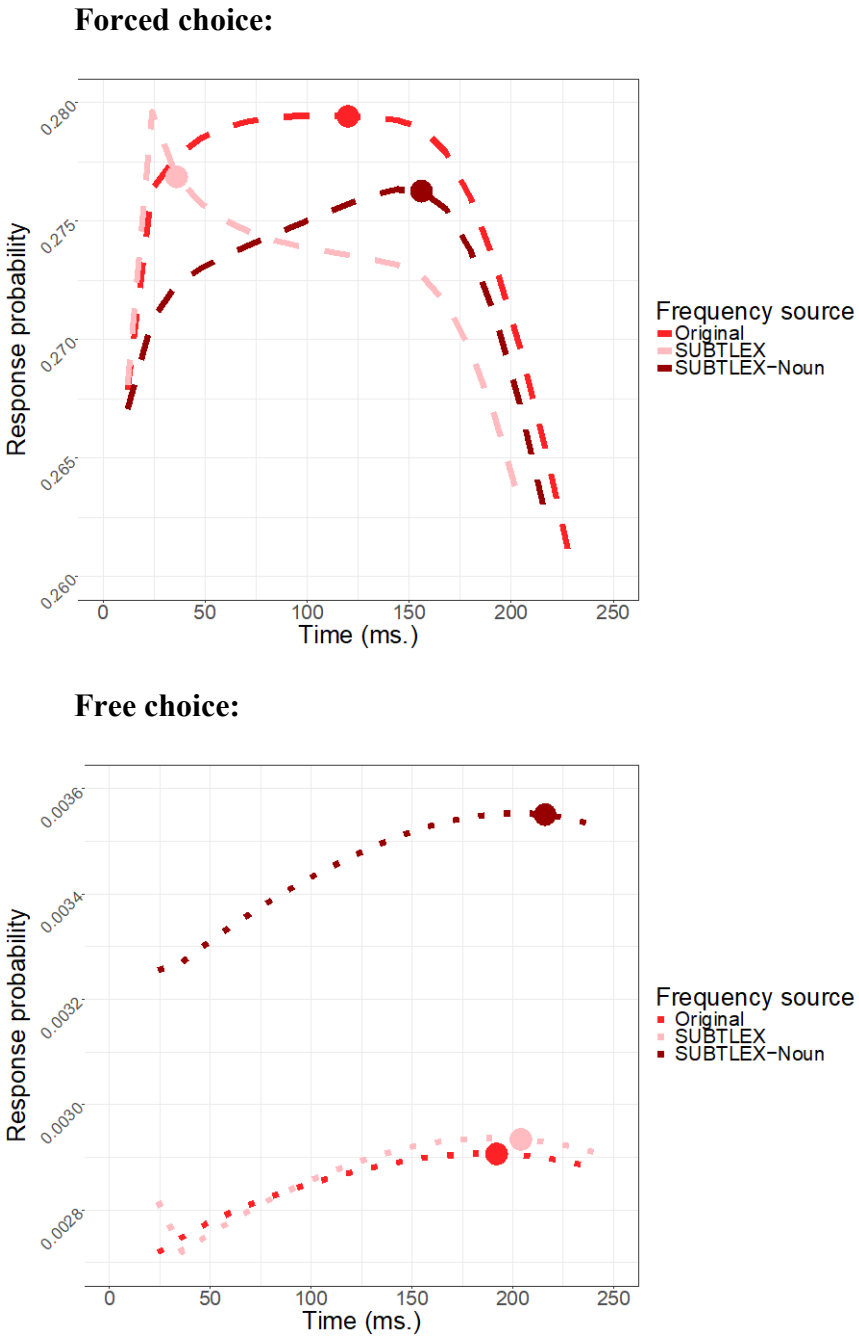
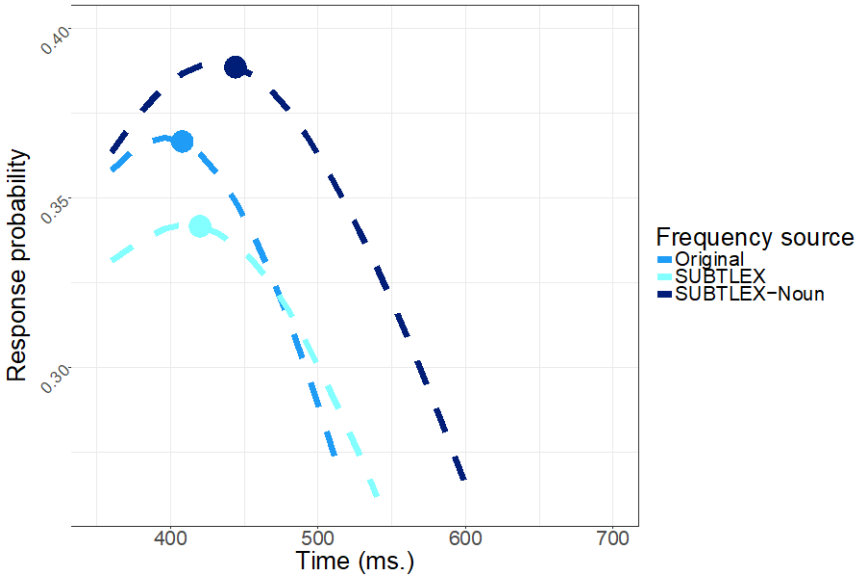


Figure 5. *The time course of response probability for the distractor in forced (top) and free-choice (bottom) tasks in the resting level frequency/default resting baseline condition, marking (with a dot) the time that response probability begins to decrease.*

Forced choice:



Free choice:

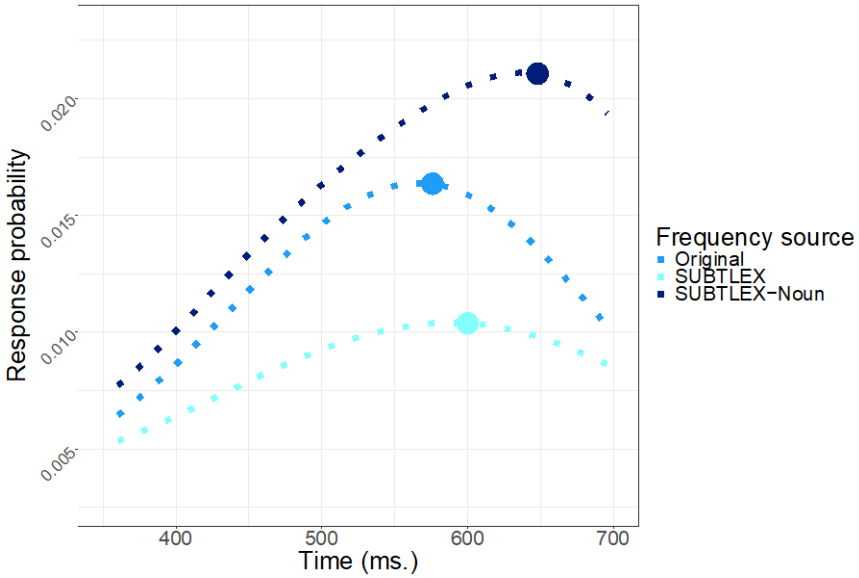


Figure 6. *The time course of response probability for the high-frequency competitor in forced (top) and free-choice (bottom) tasks in the resting level frequency/default resting baseline condition, marking (with a dot) the time that response probability begins to decrease.*

could support the possibility that the way in which context speeds or aids the process is *not* faster cohort competition. Conversely, if faster cohort competition were understood to be a reliable consequence of contextual constraint, these simulations would point to a different implementation of the constraint, and not one in which items from the incompatible syntactic category are effectively not in the lexicon.

2.4.3 Summary

A category-restricted lexicon unexpectedly leads to slower distractor and competitor divergence than a general lexicon does. Exactly what properties of the model lead to the unexpected direction of this effect deserves further scrutiny, and comparison with matched human data would be illuminating. Our results suggest that this approach could be used to shed light on both the mechanism for the contextual constraint (in humans) and the way in which it alters word recognition. These results also reinforce the forced-choice advantage observed with the original frequency values in our first set of simulations. This advantage again occurs across the board when measured for the high-frequency competitor, but only with a specific combination of parameter settings when measured for the distractor.

2.5 Discussion

The goal of these simulations was to begin exploring the hypothesis that context effects appear earlier when the response candidate set is very small (in the visual world paradigm) than when it is unrestricted (in, for example, gating). Our findings were a first step toward quantifying the effect of the size and composition of the response candidate set on typical response dynamics. Examining simple bottom-up phonological competition effects in jTRACE simulations of Dahan et al. (2001),

we found that the response probability of an item ruled out by the current input begins to decrease systematically earlier when the response candidate set is restricted. One important lesson from these simulations was that these differences caused by the response candidate set manifest only in response probability, and not in underlying lexical activation.

For a researcher aiming to assess changes in lexical activation, the visual world might then provide a more faithful estimate of its time course than methods with larger response candidate sets. This idea runs counter to the intuition that a small candidate set might enable unnaturally early effects on the cohort, but is a simple clarification once activation and response probability are properly distinguished. Importantly, the set-size effects we observe are not effects of the size of the set, per se, but of its composition, or more specifically its distribution of items that are or are not affected by the input. The link with size is because small sets are more likely to have distributions that are skewed relative to the unrestricted lexicon. Thus, the contrast we observe between simulated response probability profiles in free and forced-choice tasks illustrates that inferences about activation from response probability depend heavily on set composition. Changes in the response probability of a critical item must be interpreted in the context of any activation changes occurring for other items in the response candidate set.

Another, testable prediction arises from this insight: a dependent measure reflecting lexical activation directly, without the mediating influence of a behavioral response, should align more closely with the time course derived from the visual world than the time course derived from gating. MEG and EEG, allowing neural

activity in response to speech to be recorded with fine-grained temporal resolution during passive listening, are likely to be the best candidates, and could provide more accurate timing estimates that do not rely on assumptions about the link between activation and fixations.

My primary question 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 revealed a new potential framework for understanding cross-method differences, and they have highlighted an important design consideration for any future work. Evidence that a forced-choice task makes it possible for bottom-up phonological effects to occur earlier in response probability than would be possible in free choice, as our simulations have suggested, is of course logically independent of evidence for or against the constrained generation of candidates. Definitively resolving the conflict will require using the same stimuli and maximally similar set-ups in the visual world, gating, and a neural measure, and testing TRACE's predictions fully will require direct comparisons of the timing of top-down and bottom-up effects in each dataset. But it is likely that changes in activation due to top-down influences, like syntactic category, would be subject to the same influence of the response candidate set in the conversion to response probability. The relative delay between forced-choice and free-choice effects would not necessarily be the same, because it would depend on exactly how distorted the distribution of categories is within the forced-choice set as compared to the free-choice set. And, of course, the categories in this case would be syntactic rather than phonological. But if TRACE is correct, and if set composition

has a remotely similar effect on top-down cohort activation changes as it does on bottom-up changes, one potential explanation for the observed conflict is that the delayed syntactic category effect in gating arises from changes in activation that are the same as those causing earlier effects in the visual world; the visual world would simply be reflecting them in behavior more quickly.

In the next chapter, we report a new experiment capitalizing on these insights from our simulations. This study was set up to specifically distinguish facilitation and inhibition as mechanisms for a category constraint, while resolving design issues in some previous experiments asking whether wrong-category cohort competition is occurring.

Chapter 3: A visual world study on facilitation vs. inhibition as mechanisms for the syntactic constraint on cohort competition[2]

3.1 Introduction

3.1.1 Overview

We are concerned, in this chapter, with whether or not lexical items compete for recognition as the word being heard even when their syntactic category is incompatible with the context. Different methods for asking this question have yielded different outcomes, as reviewed in **Chapter 1**. In **Chapter 2**, we showed that the size and composition of response candidate sets might help explain that variation, and are important in experimental design and interpretation. Our simulations suggest the possibility (which will require considerable follow-up) that visual world designs allow activation changes to be transmitted to response probability more quickly than designs with larger response candidate sets. In this chapter, we report a new experiment in the visual world paradigm that is set up to distinguish the effects of a facilitatory and an inhibitory syntactic category constraint, an issue that has not previously been addressed[3].

---

[2] Margaret Kandel, Anna Namyst, Nadiya Klymenko, Oliver Bentham, Reese Alpher, and Lalitha Balachandran assisted with stimulus preparation and data collection for Experiment 1. Some of the results of this study were previously reported in the Proceedings of the 11th International Conference on the Mental Lexicon 2018.

Our interest is in the set of words that are under consideration as auditory input unfolds, and whether that set is influenced by syntactic category; our focus in the visual world is therefore on fixations to pictures whose names have the same onset as the word being spoken, and either are or are not consistent with the syntactic context. Though a lack of fixations to wrong-category items in previous visual world results makes the constraint appear inhibitory, we argue that previous designs actually made such fixations difficult or impossible to detect. If we find evidence that the constraint is in fact facilitatory, this would present a second possible explanation for observed conflict in the literature. Our simulations allow for the possibility that the constraint is indeed inhibitory, as it appears in the visual world, but delayed in some tasks. Evidence for facilitation in this new experiment would instead align the visual world with cross-modal priming and gating under a facilitatory account. Of course, it is possible both that the constraint is facilitatory and that set size/composition matter for effect timing.

Our visual world design has several important 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 critical trials do not actually include the target (e.g., when the auditory input was '*He chose the battleship*', the display would include an image of cohort competitor *balcony*, but no battleship). As discussed in **Chapter 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. In our design, therefore, we omitted the target so 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. 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, 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 of this chapter). 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").

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.

3.1.2 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 are predicting 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.

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.

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 manipulation in which a context effect predicts a difference in fixations to the competitor regardless of the constraint being inhibitory or facilitatory. This expected difference would also yield information about the timing of the constraint.

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 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 Linking Hypothesis 2.

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.

1. Pictures are presented to the participant at least 1000 ms before auditory input begins.
2. 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.
3. 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.
4. Per Huettig and McQueen (2007)'s "phonological hypothesis," phonological representations have been retrieved or activated 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 such that representations are activated via both routes.

5. Scanning of the visual scene continues for the duration of the trial. Pictures will be fixated if there is a match between the visually and auditorily activated representations. The proportion of fixations among the pictures is determined by their relative activation levels. TRACE (McClelland & Elman, 1986) proposes the Luce Choice Rule for conversion from activation to fixation probability.
6. 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 takes roughly 200 ms for a saccade to be planned and executed, 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, though conventional wisdom is that the lag is closer to 200 ms.

The assumption that phonological representations of visually presented pictures are activated in the absence of auditory input is sometimes referred to as the "implicit naming hypothesis" and is not universally accepted. An alternative hypothesis to implicit naming is that cohort competition effects arise via matching at the level of conceptual or visual features. In this scenario it is not necessary for the visual route to activate up to phonological representations. Instead, auditory input activates phonological, then lexical, then conceptual and possibly visual

representations, and fixations are driven by matches with conceptual representations activated by the pictures or simply their visual features. This is the essence of the assumption of independence between language processing and visual processing made by Allopenna et al. (1998).

The logic in implicit naming, in contrast, is that when participants are given preview time, the visual input has a head start in activating up to the phonological level, so matching can occur as soon as the auditory input activates to the phonological level. We assume, in addition to implicit naming, that attempted matching occurs at every level, and thus conceptual/visual matching will also occur 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, this is the first match that can be made.

We assume implicit naming 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 presence or absence of preview time should be irrelevant. The visual or conceptual features of the pictures are readily available with no time needed for activation of alternative levels of representation; it therefore does not matter how long the pictures are available before the auditory input begins. 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 to what extent the activation of phonological representations from the pictures actually distorts the activation of phonological representations from the auditory input, which is what we aim to measure in this paradigm. For example, it could be the case that what is maintained in short-term memory for each picture is a pointer to the phonological representation. The "match" driving fixations would be whether or not a phonological representation activated by the auditory input has a pointer from the short-term memory representation of the visual display. Implicit naming would thus enable a convenient window onto phonological representations without actually influencing them.

3.2.2 Design details

3.2.2.1 Noun-only competitors

Our first manipulation 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)). 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," and 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.” 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 will occur 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 employ repetition of 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 those 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. Since all pictures in the display have noun labels, a constraint that boosts activation for all nouns should affect all picture labels identically, and hence have the same effect on response probabilities as if the category constraint simply failed to apply. To address this concern, we included a second manipulation aimed at distinguishing these possibilities by 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 positive effect of context in this second manipulation.

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").

In the first manipulation, when we compare response probabilities for the same competitor picture in the two contexts, we are comparing a context in which the constraint should have applied and a context in which it should not have applied. The inhibitory constraint causes activation changes to all four noun-only pictures in the verb context but not the noun context. The facilitatory constraint causes activation changes to those noun-only pictures in the noun context but not the verb context; however, because in the noun context the same change occurs for all four pictures, there is no observable response probability difference, and therefore no difference from the context in which there is no constraint applying and so no response probability difference. While the inhibitory constraint does cause a change for all four pictures, that change is something along the lines of multiplication by zero (or, alternatively, return to baseline), and this means that the proportional activation change for the competitor versus distractors is different, because the distractors were already at baseline. This competitor/distractor difference is what creates a visible change in response probability, relative to the context where no changes occurred.

In this second manipulation, because the competitor is noun/verb ambiguous, it should be affected in both contexts, under either constraint. In all cases, the change in activation for the noun/verb ambiguous competitor will be different from the change in activation for the noun-only distractors, such that if we were comparing to a context in which no constraint applied, we would see the change in response probability. We are not comparing to such a context, but the extent of

competitor/distractor difference will vary between the two contexts that we are comparing, which is what leads us to expect a difference in response probability.

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 will increase in the noun context and not the verb context, which 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 context. Therefore, if our assumptions about the representation of category-ambiguous words hold, this second manipulation leads us to expect a context effect under either constraint, and so could provide positive evidence that participants are applying some form of category constraint. This is useful in the case that the first manipulation yields no difference 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.

This second manipulation also provides timing information about the category constraint that would be impossible to glean from the first manipulation, if the category constraint is implemented via facilitation. In the first manipulation, because all four picture labels have the same category status, the application of the category constraint, whether immediate or delayed, should have the same effect on all four pictures, assuming that the effect is proportional. Because our dependent measure is a proportion of fixations, we should not see a change. However, in the second manipulation, a category constraint should affect the noun-verb ambiguous item differentially from the noun-only distractors. If fixations to the soap differ during "sofa" and "socialize," the timing of this difference can indicate when the category constraint is applied.

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 manipulation, 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 manipulations 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 (Duñabeitia et al., 2018) and SVLO (Rossion & Pourtois, 2004) repositories, as well as a small number of supplemental clip-art drawings. 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 in SUBTLEX-US (Brysbaert & New, 2009) and listed as usable only as nouns or only as nouns or verbs. We removed pictures whose names had fewer than 3 phonemes or nasals as the third phoneme following a vowel as the second phoneme. Finally, we took the subset of the remaining pictures that could form a triple with a noun-only and verb-only onset competitor listed in SUBTLEX-US (Brysbaert & New, 2009).

The sentence frames we used (e.g. “She wanted to/the…”, “They chose to/the…”) were adapted from Fox and Blumstein (2016). See **Appendix A** for a full list of stimuli (also available for download on OSF). 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, in which two phonemes are 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. This should not occur systematically so as to introduce a confound for our contrast of interest. 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 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, 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 manipulation in which we used noun-verb ambiguous competitors, we expected competition effects to be larger because of the higher frequency of the picture names. But we also did not expect the interaction between contexts for this manipulation 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).

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. 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 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 courses of the competitor advantage for each participant were 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. 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 is 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 7**) 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 7**), 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.

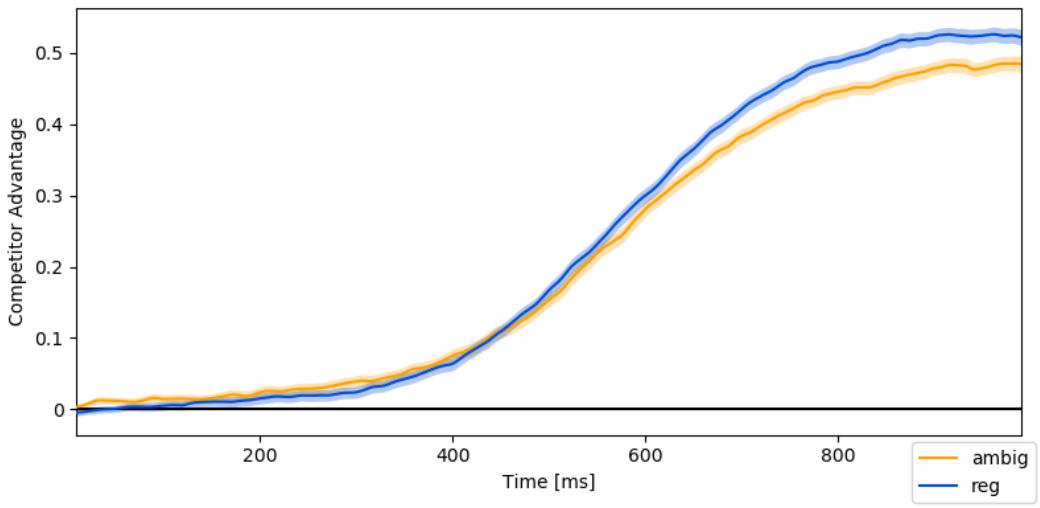


Figure 7. *For filler trials, the smoothed time course of competitor advantage relative to baseline following auditory target onset, for noun-only ('reg') and noun-verb ambiguous ('ambig') competitors, which are fully consistent with the auditory target. Shading indicates one standard error.*

For our first critical manipulation, 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 8**). 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, if slightly larger than, what can be observed in the plots from Huettig and McQueen (2007) and Strand et al. (2018), though neither computes or 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.

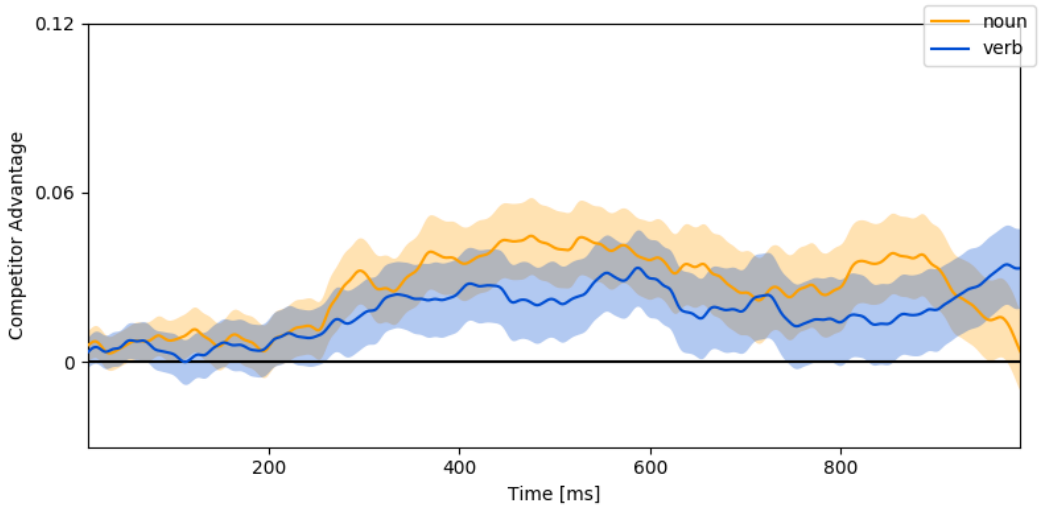


Figure 8. *For the noun-only competitor in critical trials, the smoothed time course of competitor advantage relative to baseline following auditory target onset, in noun and verb contexts. Shading indicates one standard error.*

For our manipulation of noun-verb ambiguous competitors (**Figure 9**), 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 manipulation, 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 ($p < .001$).

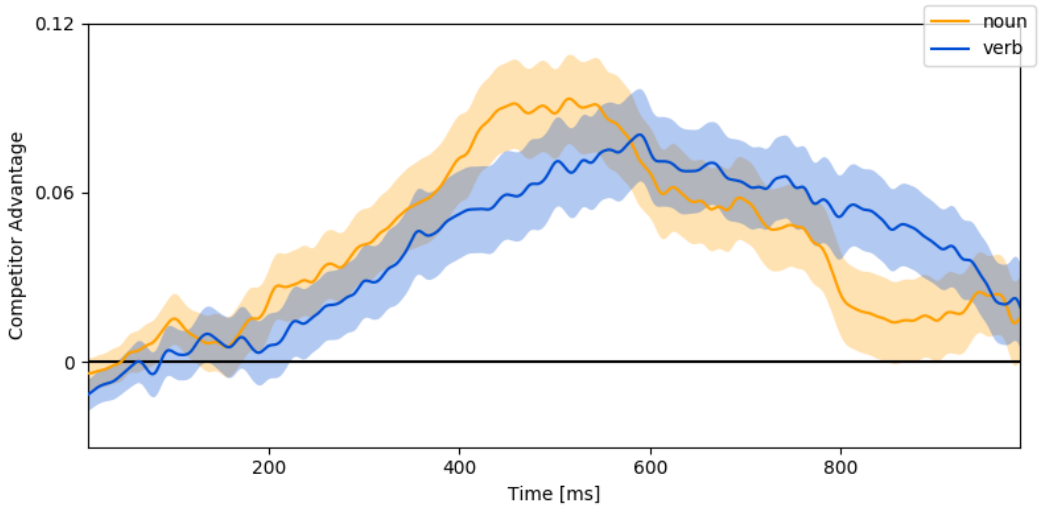


Figure 9. *For the noun-verb ambiguous competitor in critical trials, the smoothed time course of competitor advantage relative to baseline following auditory target onset, in noun and verb contexts. Shading indicates one standard error.*

3.7 Discussion

In this study, we used the visual world paradigm to ask whether and how syntactic context constrains lexical activation. Our design was intended to distinguish between facilitatory and inhibitory mechanisms for the syntactic context constraint. 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? Participants heard auditory targets that they could expect to be either nouns or verbs on the basis of the syntactic context. We measured fixations to pictures of cohort competitors that could only be nouns or were strongly noun-biased. The study was designed to maximize the chance that the activation of wrong-category competitors would be detectable if it were occurring. This was done by not including any other items in the display that could cause changes in response probability and by

making sure that the pictures could not be classified as noun- or verb-associated based on visual properties alone. These decisions were partially informed by our simulations in jTRACE (**Chapter 2**), which demonstrated the importance of the response candidate set in interpreting response probability.

Given these design parameters, we did find immediate cohort competition from wordforms that were incompatible with their syntactic context, with no evidence that it differed from competition in the correct context. This contrasts with previous work in the visual world (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 predictions of an inhibitory account that we would 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. This finding is consistent with both a facilitatory account and an account in which syntactic context did not constrain lexical activation at all.

Our manipulation of 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 that 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. Another possibility is that the designs made it difficult for changes in activation to be detected in response probability.

For Magnuson et al. (2008), the issue is whether the experimental context impacted the way in which images were implicitly labeled by participants. As discussed in **Chapter 1**, 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 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. This 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.

Put another way, if the task requires paying attention only to a shape’s noun descriptor or only to a shape’s adjective descriptor, it would be reasonable for comprehenders to label the referents only with respect to their adjective or noun descriptors. If this is the case, then it is unsurprising that onset overlap between the auditory target and the other, irrelevant descriptor word for the shape would not drive increased fixations, even if wrong-category cohort competitors were activated by the auditory input, because the irrelevant descriptor word for the shape would not be available for a match. Though it is possible that comprehenders do implicitly name both the adjective and the noun even though the task discourages it, the explanation we are proposing cannot be ruled out.

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 one. 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 **Chapter 1**, 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, 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 manipulation 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 would have been 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 that appears in a highly unexpected syntactic context or if the context was misheard; 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 explain the appearance of such an effect in previous work. We have described how this could have occurred in the visual world paradigm. In cross-modal priming and gating, if

there is no constraint we should not observe any modulation of competition due to context. This is fairly simple to reconcile with the case of cross-modal priming, for which early probe points showed no context effects (though a high-powered study would be necessary to more confidently rule them out). The later constraints observed in cross-modal priming are harder to explain under an account where syntactic context has no impact at all, though these timepoints are later than the time window generally analyzed in visual world data. The modulation in competition 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. Nevertheless, we can revisit Tanenhaus and Lucas (1987)'s point that a syntactic category constraint on competition might not actually be useful in easing or speeding word recognition. It could also be that syntactic category is not represented on the lexical item in such a way that it is possible for it to be used in a cue-based search, or that in the style of TRACE it is not represented as a layer whose activation can feed down to influence wordform units.

3.7.4 Caveats

A final explanation for our findings that we must acknowledge is that particular aspects of our own study 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 somewhat 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. If participants could visually identify potential noun and verb referents at the beginning of the trial, and then used that information once they heard "to" or "the" to restrict looks to referents of the right category, the immediacy of their context effect indicates that the necessary constraint from "to" or "the" was available by the onset of the critical word. There is no obvious reason that the same should not be true in our study, since we used similar sentence contexts. 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 manipulation, 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 Integrating with simulations from Chapter 2

We note that both the simulations that we report in **Chapter 2** and the visual world results we report here point to the fact that the presence of early wrong-category competition does not necessarily mean there is no constraint in place. Our simulations suggested that one way to explain cross-method conflicts in the timing of syntactic constraints on cohort competition could be that the immediate constraint apparent in previous visual world studies is correct, and the same underlying activation changes simply take longer to manifest in response probability for tasks with larger response candidate sets. If our visual world study had yielded evidence for inhibition, this would have been consistent with the possibility that the inhibitory constraint apparent in the visual world is simply delayed in other methods. We did not see evidence for inhibition, and a facilitatory account presents an alternative solution for the conflict. However, we leave open the possibility that both

explanations have some truth, or that they interact. The lessons from the simulations apply equally well regardless of the mechanism for the constraint, and could still contribute to our understanding of timing differences across methods even if the constraint is facilitatory.

3.7.6 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 very 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 manipulation, 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.

3.8 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. We showed with simulations in jTRACE that the speed with which bottom-up cohort competition effects can be expected to manifest in a behavioral measure depends on the size and 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 and is important for study design and comparison. In a visual world eye-tracking study building on this understanding, we employed a new design to distinguish whether the potential category constraint operates by boosting 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.

Chapter 4: Studying the cohort via neural effects of phoneme surprisal and cohort entropy[4]

4.1 Introduction

4.1.1 Overview

The core problem, in this dissertation, is how the activation of lexical candidates in response to auditory input is influenced by syntactic context. I am specifically investigating whether syntactic constraints on activation function to facilitate good contextual fits or inhibit lexical candidates whose syntactic category is incompatible with the context. Lexical activation, however, cannot be directly observed. In the previous chapters, we have discussed the relative merits of a variety of behavioral measures that have been used to investigate lexical activation. We have considered cross-modal priming, gating, and in particular, the visual world paradigm, which we argue is the most viable behavioral method for this question. However, the visual world paradigm allows us to probe the activation levels of just a single item in the mental lexicon during an experimental trial, and we must do so indirectly, via fixation proportions. These fixation proportions can easily obscure the activation patterns of a critical item over time, and are subject to other driving forces besides

[4] Experiment 2 was analyzed in collaboration with Christian Brodbeck. Daphne Amir, Fen Ingram, and Stephanie Pomrenke assisted with stimulus selection, and Aura Cruz Heredia assisted with some of the MEG data collection. Unfortunately, data for Experiment 3 could not be collected due to the COVID-19 pandemic.

lexical activation. In Experiment 1, a visual world study reported in the previous chapter, we used a design intended to isolate the link between lexical activation and fixation proportions to the fullest extent possible. That study yielded an effect that is inconsistent with an inhibitory syntactic constraint on lexical activation, but did not yield positive evidence for the presence of a facilitatory syntactic constraint, or information about constraint timing.

In the work reported in this chapter, we advance an experimental approach that holds promise for more effective investigation of the syntactic constraint. Neural measures allow us to record responses to experimental stimuli while participants listen passively rather than, for example, look at pictures or respond to probes. These measures thus require fewer assumptions about the mapping between cognition and behavior. However, we still need hypotheses about how lexical activation should influence measurable neural activity during listening, in order to be able to ask how activation is affected by contextual constraint. A starting point might be the use of simulated lexical activation levels of experimental stimuli to predict neural activity at any given phoneme. One challenge for this approach is that jTRACE does not have a sufficient lexicon or phoneme inventory to do this for a large set of stimuli, though future models might. Another challenge is that even with simulated activation values in hand, we would need to calculate measures that reflect properties of the whole cohort (e.g., size, summed activation, or a measure of variability or dispersion), since our neural measures do not have the resolution to record a response that we could reasonably understand to reflect the activation of just a single lexical representation.

With these difficulties in mind, we pivot to a different conceptual framework in this chapter, moving from auditory word recognition as a process of fluctuating lexical activation levels to auditory word recognition as a process characterized by shifts in probability. In this view, a wordform that is competing as a candidate has increased probability rather than increased activation, and all relevant cognitive computations are expressed in terms of probability. Importantly, this is not a claim that probabilities are simply computed over activations. We follow prior work in hypothesizing that information-theoretic properties of the set of lexical candidates are likely to be reflected in measurable neural activity. Many such properties could be computed over the set of candidates, but in this chapter we focus on the two that have supporting evidence in the neural literature: phoneme surprisal and cohort entropy. Both are based on probabilities, given the phonemes in a word that have been heard. For phoneme surprisal, the relevant probability is that of the current phoneme, while for cohort entropy what matters is the probabilities of the various words that could complete the phoneme sequence. We can easily estimate these probabilities using corpus frequencies, and they satisfy our requirement for summary properties that can be calculated for a single phoneme but are influenced by the dynamics of the entire distribution of candidates.

In earlier chapters, we argued that investigating activation via probability in the visual world paradigm is difficult because the probability of fixation on an item of interest is influenced both by activation of the item of interest and activation of other items in the set under consideration. In this case, we are arguing that the use of probability is specifically beneficial, by essentially the same logic. The crucial

difference is that in the case of the visual world paradigm, the set over which probability is computed is the four items on screen, which is not of any particular theoretical significance. In the cases of cohort entropy and phoneme surprisal, probability is defined with respect to the cohort of lexical candidates or the phonemes that make them up. That cohort is our object of interest, as we try to ask how it is affected by syntactic context. Thus, its influence on probabilities is useful.

There is not strong evidence making either the activation or probability account of word recognition more likely. Because our main concern is the influence of syntactic context, we can be largely agnostic about the correct framework for this problem, at this stage of research. Whether we are investigating activation or probability, what matters is whether it is impacted by a syntactic constraint. With neural measures, we simply need to be able to observe some way in which the cohort of wordforms under consideration influences phoneme-level processing. Though there may be interesting places in which activation and probability make different predictions, we expect them to be highly correlated in many instances. Thus, in this chapter, we switch to a probability framework in large part out of convenience. Of course, in the long term, it will be important to understand which framework better characterizes the cognitive processes underlying auditory word recognition.

4.1.1.1 Cohort entropy and phoneme surprisal

Phoneme surprisal is a variant on the conditional probability of a phoneme given the preceding phonemes in a word, where lower probability phonemes have a higher surprisal. Probability is usually estimated from frequency. Consider a lexicon that has only four equally frequent words, /bri/, /dra/, /blo/ and /blu/. Following a /b/

word onset, /r/ will have a higher surprisal value than /l/, because /l/ occurs more often in that position. However, if the frequency of the word /bri/ were doubled, the surprisal values of /l/ and /r/ after /b/ would become equivalent. Importantly, while phoneme surprisal reflects the conditional probability of just a single phoneme, computing it requires reference to all wordforms in the lexicon and their token frequencies.

Phoneme surprisal can be computed for each phoneme in a word. The surprisal of the first phoneme comes from the probability of that phoneme as a word onset in the language (in the example lexicon above, first phoneme surprisal will be larger for /d/ than /b/). Phoneme surprisal has been seen as a good candidate for a measure that would impact neural activity because of existing evidence from many cognitive domains that predictable stimuli elicit strongly reduced neural responses. Thus, more neural activity is expected in response to higher surprisal, or less predictable, phonemes. Phoneme surprisal at position *i* in a wordform is defined as:

$$surprisal_i = -log_2\, p(k_i \mid k_1, \ldots k_{i-1})$$

where k_i is the phoneme at position *i* and *i* = 1 for the first phoneme in the wordform. $p(k_i \mid k_1, \ldots k_{i-1})$ is therefore the conditional probability of phoneme k_i given the sequence of phonemes that preceded k_i in the wordform. In practice, we use wordform frequencies from a corpus to calculate conditional phoneme probabilities within a word. When C_i is the cohort of possible wordforms *w* consistent with the

sequence of phonemes up to and including position *i*, surprisal at position *i* is calculated as:

$$surprisal_i = -log_2 \frac{\sum_w^{C_i} freq(w)}{\sum_w^{C_{i-1}} freq(w)}$$

Cohort entropy is a very different property of the set of lexical candidates, though we use the same wordform frequencies to calculate it. Phoneme surprisal is determined by the probability of the current phoneme following a specific phoneme sequence. Cohort entropy instead reflects how much uncertainty there is about what the most likely word completion is, given the current phoneme sequence.

Therefore, one property of the set of candidates that entropy will reflect is how many options there are. With the example lexicon (/bri/, /dra/, /blo/ and /blu/), when the listener hears /d/ there is no uncertainty about what word is being heard, as the only option is /dra/. Entropy at /d/ is thus zero. When the listener hears /b/, entropy is higher because there are three options.

However, entropy also reflects probability, as estimated from frequency. When the frequency of /bri/ in the example lexicon was doubled, /r/ and /l/ became equally likely after /b/. So, the entropy of the distribution of options, when hearing /b/, was very high, as there was maximal uncertainty about what would follow. In contrast, when all four items in the lexicon had the same frequency, /l/ was more likely than /r/ following the /b/ onset, because there were two /bl/ options. Thus, there was more certainty about what would follow /b/, and less entropy at /b/. Cohort entropy at position *i* in a wordform is defined as:

$$entropy_i = -\sum_{w}^{C_i} p(w \mid k_1, \ldots k_i) \times log_2\, p(w \mid k_1, \ldots k_i)$$

where w is each wordform in the cohort C_i of wordforms consistent with the sequence of phonemes $k_1, \ldots k_i$. Cohort entropy therefore differs crucially from phoneme surprisal in that it is based on, given a sequence of phonemes, the conditional probabilities of wordforms that might complete the sequence rather than the conditional probability of the phoneme that is occurring in the current position. In practice, we again calculate cohort entropy using corpus frequencies, such that cohort entropy is given by:

$$entropy_i = -\sum_{w}^{C_i} \frac{freq(w)}{\sum_{w}^{C_i} freq(w)} \times log_2 \frac{freq(w)}{\sum_{w}^{C_i} freq(w)}$$

when C_i is the cohort of possible wordforms w consistent with the sequence of phonemes up to and including position i.

Entropy, like surprisal, can be computed at each phoneme in a wordform. However, once a wordform is the only option given the input, entropy becomes zero. One way to link entropy to neural activity is to hypothesize that more uncertainty drives more activity. This is intuitive for the sense in which more competitors will often lead to higher entropy. However, as we have described, the relative distribution of probability among the candidates also matters for entropy, and it is not obvious

why a set of candidates with equal probabilities should necessarily drive more neural activity than a set of candidates with unequal probabilities. Therefore, a second linking hypothesis between entropy and neural activity is that low entropy, and higher certainty about what the word being heard will turn out to be, is a precondition for other processes to be engaged. Ettinger, Linzen, and Marantz (2014) discuss a version of this hypothesis, in which predictions for upcoming input are only made when entropy is low.

Both cohort entropy and phoneme surprisal have been shown in several instances to correlate significantly with neural activity in auditory cortex at a phoneme-by-phoneme level, as we will review in **Section 4.1.2**. However, we cannot know whether entropy and surprisal are the correct, direct linking hypotheses between the cognitive and neural processes underlying word recognition and the neural response as measured by (in most cases) MEG. Correlation with these variables could arise because the variables themselves are correlated with other properties or processes that are the true modulators of the neural response.

It is also unclear whether effects of cohort entropy and phoneme surprisal are linked to different levels of representation involved in the word recognition process. We discuss, in the literature review below and in discussion of Experiment 2, how the profile of entropy and surprisal effects across studies differs, and how this could be informative about their different underlying causes. Phoneme surprisal reflects phoneme probability within a word (thus assuming correctly identified word boundaries), and it is closely related to overall measures of phonotactics in a language. Cohort entropy, on the other hand, concerns the probabilities of lexical

candidates for the word being heard. Effects of cohort entropy seem to require involvement of lexical or wordform representations. This could also be true of effects of phoneme surprisal if they reflect phoneme-level probabilities arising from the set of wordforms under consideration, in real time. However, phoneme surprisal effects could equally well arise from phoneme-level probabilities that are tracked or stored independent of wordform-level probabilities, even if wordform-level probabilities are ultimately still their source. In that scenario, surprisal effects alone are not evidence for a phoneme-by-phoneme cohort updating process. Nevertheless, many studies using these measures have tended to assume that both cohort entropy and phoneme surprisal effects reflect the dynamics of the wordforms under consideration at each phoneme. Thus, distinctions drawn between entropy and surprisal effects are usually with respect to the notions of uncertainty vs. prediction error, rather than wordform vs. phoneme probability.

4.1.1.2 Experiments 2 & 3

Calculating entropy or surprisal requires some knowledge of what wordforms should be competing, given the input (i.e., what the cohort of possible wordforms is), and what the frequencies or probabilities of those wordforms are. The default way to do this is to take all wordforms consistent with the input phonemes, and the overall frequencies of those wordforms in the language. The premise of Experiment 3 in this chapter is that, given a contextual constraint that we want to investigate, we can hypothesize how the cohort might be affected, both in terms of which wordforms compete, and what their relative frequencies are in the context. Given this hypothesis, we can calculate entropy and surprisal using cohorts that are not simply all words

consistent with the input, or whose frequencies are not simply the overall frequency of the wordform. We can then ask whether these entropy and surprisal values do a better job of predicting neural activity than those calculated from default cohorts and frequency values. Specifically, we construct cohorts under the influence of a hypothetical, facilitatory syntactic constraint, and contrast them with cohorts under the influence of an inhibitory constraint, or no constraint at all. The design of Experiment 3 builds on prior work (Gaston & Marantz, 2018) with some refinement of both the design and the analysis method, as well as a new understanding, due to the previous chapter, that a syntactic constraint need not be inhibitory.

Before conducting Experiment 3, however, we believed it necessary to run a simple study on the recognition of single words that would establish baseline entropy and surprisal effects. We had two motivations for this. The first was that even entropy and surprisal effects computed in the default way present with significant variability in the current literature, as we discuss in the next section. They are also usually tested for in service of some other primary question. However, fine-grained conclusions about the restricted cohort require that effects of the unrestricted cohort be both well understood and predictable. The presence of simply any entropy or surprisal effect is not sufficient. Therefore, Experiment 2 in this chapter was intended to demonstrate effects of basic, unconstrained entropy and surprisal in single words, with no other manipulations. We consider Experiment 2 a necessary benchmark in refining the use of entropy and surprisal as reliable measures for probing the composition of the cohort in neural data.

Our second motivation for Experiment 2 was that a new analysis method had been introduced for analyzing neural responses to continuous speech in source-localized MEG data. This method, modeling temporal response functions, is better equipped than previous analyses to deal with acoustic and other confounding variables and the nature of overlapping phoneme responses. We intended to apply this analysis method to Experiment 3, but needed to first demonstrate its efficacy for single words and controlled contexts rather than continuous speech. Experiment 2 allows us to do this.

4.1.2 Literature review

In the following sub-sections, we first review reports of cohort entropy and phoneme surprisal effects in neural data when the stimuli are single words or controlled contexts, and then when continuous speech is used. We discuss several additional studies that we consider related but not strictly relevant because of differing dependent measures or formulations of entropy and surprisal. Finally, we summarize the literature and patterns that we observe in the manifestation of entropy and surprisal effects, which may have relevance for understanding the processes driving these effects.

We do note, to start, that several studies using behavioral measures of auditory word recognition (lexical decision reaction time or production latency after listening) have demonstrated effects of cohort entropy or phoneme surprisal (Baayen et al., 2007; Balling & Baayen, 2012; Bien et al., 2011; Kemps et al., 2005; Wurm et al., 2006). However, as each phoneme in a word has its own surprisal or entropy value, but only one data point per word can be collected with these behavioral measures,

experiments must compute cumulative or composite versions of entropy or surprisal, or ask about the influence of entropy or surprisal at just a single time point within the word on reaction time at the end of the word. These studies therefore have limited informativity for the role of entropy or surprisal in incremental auditory processing.

4.1.2.1 Single words and controlled contexts

The first evidence for cohort entropy or phoneme surprisal effects in neural data comes from Gagnepain et al. (2012), who, using MEG, attempt to explicitly distinguish accounts of spoken word recognition in which the process is characterized by lexical competition (and therefore lexical uncertainty, or entropy) from accounts in which the key computation is segment prediction error (or surprisal). Gagnepain et al. consider these two alternatives to be mutually exclusive. The fundamental distinction that they draw concerns how lexical candidates that are inconsistent with the input are removed from consideration: via segment prediction error or because of inhibition from other lexical candidates that receive more support from the input. This is therefore a mechanistic hypothesis with different predictions for how competition should manifest in neural data. They do not make clear, however, why it could not be the case that uncertainty over lexical candidates could not still be relevant in a segment prediction model of spoken word recognition.

Gagnepain et al. use a novel word consolidation paradigm in which participants are taught new words on Day 1 which, after sleep consolidation and entry into the lexicon, should extend the uniqueness points of existing words. For example, learning the word “formubo” would extend the uniqueness point of the word “formula.” Upon hearing “formula” on Day 2, then, entropy just prior to the new

uniqueness point should increase (relative to words that do not have a new competitor) because there is now lexical uncertainty when previously there was not. Prediction error at that segment should not change. In contrast, just after the new uniqueness point, prediction error should increase (again, relative to words that do not have a new competitor) because there was an alternative at that segment when previously there was not. Entropy at that segment should not change. Rather than examining variation in surprisal or entropy that occurs naturally across words and how well this variation predicts neural activity (the approach of nearly all other studies described in this section) they are instead considering the effects of deliberate manipulations of these variables.

The first dependent measure that Gagnepain et al. employ is averaged global field power, or the root mean square of the sensor data across all gradiometers. In the time window prior to the new uniqueness point, they find no difference between words that did or did not have a new competitor added. In the time window after the new uniqueness point, they do find a difference, with an increase in averaged global field power in the condition for which increased prediction error is expected. This pattern is supported by source-localized data in superior temporal gyrus (STG), and is consistent with the hypothesized outcome of the prediction error account rather than the lexical competition account. This manipulation does not provide timepoint-by-timepoint information about the degree to which neural activity correlates with entropy or surprisal, but it does provide evidence for the predictive value of surprisal, and the segment prediction error account, in which, they propose, STG is coding the difference between predictions and input, and prediction error is fed upwards to

update the lexical representations under consideration. The lack of a neural effect of entropy is not, of course, evidence against the lexical competition account.

Subsequent MEG studies have tested for correlations between neural activity and entropy or surprisal throughout the time course of a word. Ettinger, Linzen, and Marantz (2014) ask whether surprisal effects are heightened in bimorphemic relative to monomorphemic words, perhaps because morphological structure can strengthen the phoneme predictions that may lead to surprisal effects. They also examine effects of entropy and allow for the possibility that both surprisal and entropy might matter. In their continuous variable analysis, they use temporal cluster tests to assess correlation between entropy or surprisal and neural activity in transverse temporal gyrus (TTG), superior temporal gyrus (STG) and middle temporal gyrus (MTG) ROIs during a single-word lexical decision task. Because many of the following studies use the same method, we explain this approach in greater detail here.

Within each ROI, neural activity is averaged over all sources, yielding (for each trial) a single activity value at each timepoint in the epoch for analysis. Phoneme surprisal and entropy values are calculated for each timepoint (so, for example, if the second phoneme in the word starts at 100 ms, each timepoint from 0 to 100 ms will be assigned the phoneme surprisal and cohort entropy value for the first phoneme). At each timepoint a mixed effects model is evaluated with neural activity as the dependent value and entropy or surprisal from 200 ms prior as a fixed effect. The 200 ms lag is a data-driven choice made in the absence of prior literature on the latency of surprisal effects. Testing correlation between surprisal values and neural activity 100, 150, or 200 ms later, they find the strongest effects with the 200 ms lag. After *t*-

values for the correlation coefficients at each time point are computed, a cluster-based permutation test serves as correction for multiple comparisons. The details of permutation tests can vary. In this case, the independent variable (i.e., surprisal or entropy) is randomly permuted 1000 times, and *t*-values for the correlation coefficients at each time point are computed on each permutation. For any clusters of significant *t*-values, *t*-values are summed within the cluster. Significance is evaluated by comparing the cluster with the largest *t*-value sum found with the true independent variable to the distribution of the largest sums generated on each permutation.

Ettinger et al. found significant effects of phoneme surprisal at the end of the word in all three ROIs, in ~200 ms clusters starting roughly 750 ms after word onset, showing more neural activity with higher surprisal. Given the built-in lag, this indicates effects of surprisal for the phonemes occurring in the window ~550-750 ms after word onset. They also found a facilitatory effect of cohort entropy (less neural activity with increasing entropy) in TTG for phonemes occurring in the window 135-177 ms after word onset. For the purposes of this chapter, we disregard their manipulation of morphological complexity. We note that entropy and surprisal were not evaluated in the same model, so it is not possible to say whether either effect persists when the other variable is taken into account. However, the entropy effect does appear to have a different presentation here, both temporally and spatially. Ettinger et al. propose that the dual effects of entropy and surprisal they observe could be because prediction is delayed under conditions of high entropy.

Lewis and Poeppel (2014), examining effects of imageability during spoken word recognition, also consider effects of a variety of cohort-related variables,

including biphone frequency, “cohort competition” (100x the ratio of word frequency to cohort size, defined as “the summed frequency of all items beginning with the same two phonemes”), cohort entropy, and phonological neighborhood density. Like Ettinger et al. (2014), the task was single-word lexical decision, but the stimuli were restricted to monosyllabic and monomorphemic nouns. In a similar temporal cluster test analysis on ROI averages, they find significant effects of biphone frequency in STG (160-191 ms, though it is unclear what lag, if any, is employed), and cohort competition in STS (255-276 ms). In this case, each evaluated variable is residualized with respect to all remaining variables. It appears that effects of neighborhood density in pMTG (327-347 ms) and cohort entropy in STS (250-280 ms) did not survive correction for multiple comparisons via the cluster tests.

Gwilliams and Marantz (2015) study single-word responses in STG and TTG during lexical decision in Arabic. For the final consonant in the root, which is the second to last phoneme in the word, they compute both a typical linear surprisal and a morphological surprisal conditioned only on the interspersed phonemes of the root. In time windows 100-200 ms and 150-350 ms after the onset of the root-final phoneme in their two ROIs, they conduct temporal cluster tests on the correlation coefficients for each variable.

In this study, correlation coefficients came from a single mixed effects model that included all independent variables, and the two surprisal variables were decorrelated in stimulus selection. These are both important methodological considerations given the high degree of correlation that can occur between entropy and surprisal and between different versions of either variable, making it difficult to

partition the variance explained by any single predictor unless steps like these are taken.

Gwilliams and Marantz report significant clusters for morphological surprisal in STG from 130-156 ms and 289-342 ms post-phoneme onset, as well as in TTG from 294-338 ms, showing more neural activity with increasing surprisal in all cases. Note that this approach is different from that of Ettinger et al. (2014) because a 200 ms lag between variable and response is not assumed; instead, surprisal for only one phoneme is evaluated, and the time window in which it is a significant predictor of the neural data is then discovered, as time-locked from that phoneme's onset.

The final MEG study using this general approach is Gaston and Marantz (2018). In this study, we computed three different versions of entropy and of surprisal in order to ask which better predicted neural activity in STG and TTG during the processing of words presented in minimal syntactic contexts. In the calculation of both entropy and surprisal, the set of words assumed to be competing for recognition must be extracted from a corpus, along with their frequencies. We evaluated hypotheses about restriction of the cohort due to the syntactic context by restricting the set of words that we used to calculate entropy and surprisal, the logic being that if, for example, only nouns compete for recognition in noun contexts, phoneme surprisal values calculated from the set of nouns should better predict neural activity than phoneme surprisal values calculated from the set of all words. Put another way, we calculated entropy and surprisal using probabilities that incorporated the syntactic context.

We computed unconstrained entropy and surprisal, to match previous studies, as well as two constrained versions. The first (the "form-conditional" constraint) simply removed from the set of possibilities any word that could not possibly occur in the syntactic context in which the target word was being presented. The second ("usage-conditional") constrained version of entropy and surprisal took the original set of words and altered the frequencies associated with them to reflect frequency of occurrence in the syntactic context. This meant that words that could not appear in the specified context had a frequency of zero (with the same result as the first constraint) but words that could appear in the context were adjusted so that their probability reflected their distribution of possible syntactic categories. For example, the word "clash" can be a noun or verb, and has an overall frequency of 1.314 per million words. Its frequency per million words as a noun is 0.902, so this is the frequency value we would use to calculate entropy or surprisal in noun context. In calculating our constrained variables this way, we made an implicit assumption that a syntactic category constraint would operate by inhibiting items with incompatible syntactic categories. At the time, the possibility of a facilitatory constraint was not yet apparent.

Targets were category-ambiguous noun/verb homonyms (e.g., ache, clash) that were presented in a context meant to trigger an expectation for a noun ("the clash persisted"), an expectation for a verb ("to gleam brightly") or in a context where the pre-target cue was a nonword and therefore did not lead to a category expectation ("juh ache prone"). Participants were asked to judge the acceptability of the three-word phrases, and the third word in the phrase was always selected such that it would

only form an acceptable phrase with the first two words if the context word had been correctly comprehended. For example, “the frown darkly” is an unacceptable phrase if the context word is correctly interpreted as “the” rather than “to.”

Rather than assume a fixed lag between variable and response, our analysis approach was more similar to that of Gwilliams and Marantz (2015), in that we computed correlation between our predictors and neural activity at each source and time point in epochs time-locked to phoneme onsets. The distributed nature of acoustic information makes phoneme boundaries difficult to identify definitively, and it is even more difficult to identify the point in the speech signal at which a phoneme should become identifiable to a listener, as this may be influenced by predictability. For these purposes, we used boundaries as determined by a forced aligner. Because we then use spatiotemporal cluster tests on the correlation coefficients, discovering from the data what the apparent lag is between onset and response, our analysis is less impacted by the uncertainty of the boundary than an analysis that assumes a fixed lag.

The spatiotemporal cluster tests in this study were conducted within a merged STG/TTG ROI. Using a model at each phoneme that incorporated all six entropy and surprisal values, for targets in the noun and verb contexts we found an effect of form-conditional phoneme surprisal 340-450 ms after the onset of the second phoneme, effects of form-conditional and unconstrained phoneme surprisal 150-450 and 320-450 ms, respectively, after the onset of the third phoneme, and an effect of usage-conditional phoneme surprisal 190-370 ms after the onset of the final phoneme. Most but not all effects had the same directionality of previously reported surprisal effects: more neural activity with higher surprisal. The effects at the second and third

phonemes were interactions, such that the correlation coefficients differed between noun and verb contexts. Generally, it appeared that correlations were stronger in the noun contexts.

We took these findings to indicate that there is some form of syntactic constraint operating on the cohort, perhaps in a step-wise fashion, and that some sensitivity to the unconstrained cohort is also maintained. In light of the evidence from the visual world paradigm reported in **Chapter 3**, we now see that simultaneous effects of constrained and unconstrained variables, when the constraint is computed to be inhibitory, may indicate that a facilitatory constraint better captures the state of the cohort (because category-incompatible items have non-zero probabilities). The interactions with context may indicate that our assumptions in applying the syntactic constraint to the cohort were more accurate for noun than verb contexts. In other words, the assumption that "the" leads to an expectation for nouns may be more accurate than the assumption that "to" leads to an expectation for verbs.

We also note that the constrained and unconstrained cohort variables are highly correlated with each other, and entropy and surprisal also show a high degree of correlation at some points in the word. One consequence of this is that an effect of entropy found in an initial analysis evaluating entropy and surprisal variables in separate models was not replicated in the unified model mentioned above. Even when correlated variables are evaluated in the same model, caution is warranted in distinguishing their effects and interpreting the significance of individual estimates, as is always true when multicollinearity is present in a linear model. Replication across datasets is therefore especially important.

A final issue is that the response to each phoneme likely overlaps significantly with the response to previous phonemes, and this is unaccounted for in the analysis. It is difficult to know exactly what impact this might have on analysis outcomes. Overlapping neural responses mean that the neural activity values we are evaluating for correlation with an associated cohort variable are noisy. This is complicated by correlation between cohort variables for successive phonemes.

4.1.2.2 Continuous speech

We next consider two studies that use naturalistic, narrative speech as the stimulus, rather than single words, and that employ an analysis method that accounts for acoustic variables and the overlapping nature of successive phoneme responses. In both cases, though words appear in context, cohort variables do not reflect the context and are calculated in the same way that they are in single-word studies.

Brodbeck, Hong, and Simon (2018) analyze MEG data by estimating the response time course at each source dipole as the sum of the estimated responses for each of a series of acoustic and lexical predictors. Each estimated response is the linear convolution of a response function and a time series for the predictor; overlapping phoneme responses are therefore explicitly modeled. Correlation can then be assessed between the estimated and observed responses for different models that do or do not include any given predictor, at each source point, and cluster tests used to determine at which source points the correlation differences are significant. We describe the TRF analysis method in great detail in the **Analysis** section for Experiment 2 in this chapter. Brodbeck, Hong, and Simon (2018) evaluate the following potential predictors of neural activity: acoustic envelope, acoustic onsets,

phoneme onsets, word onsets, and both word-initial and non-initial cohort entropy, phoneme surprisal, cohort size, and cohort reduction. The cohort variables are calculated without regard for context, and assuming that word boundaries are accurately recognized.

We reproduce their results in **Figure 10** below. They find that of the lexical predictors, only non-initial phoneme surprisal and cohort entropy significantly improve correlation with the neural response, in superior and middle temporal areas, peaking at 114 and 125 ms after phoneme onset, respectively. The lack of first-phoneme effects is consistent with the previously reported studies. Acoustic envelope, acoustic onsets, word onsets, and phoneme onsets, none of which are modeled in previous studies, are all significant predictors and should therefore be accounted for in future studies in case of confounds with variables of interest. Brodbeck, Hong, and Simon also report data for a cocktail-party paradigm in which participants attend to one talker in a two-talker mix. Acoustic effects for the speech of both talkers still occur. However, Brodbeck, Hong, and Simon find no effects of the lexical predictors corresponding to the unattended speech, and for the attended speech they observe effects only of word onsets and of (non-initial) cohort entropy. This accords with cohort entropy reflecting probability over wordforms and phoneme surprisal reflecting probability over phonemes; if surprisal is a phoneme prediction error signal, it makes sense that it is no longer relevant when predictions cease to match the form of the input, which is mixed with the second speaker.

Di Liberto et al. (2019) use a combination of TRF and canonical correlation analysis for EEG data collected during naturalistic listening. While they also include

acoustic variables, their primary question relates to a variable reflecting phonotactic probability as computed by the BLICK model (Hayes & Wilson, 2008), for which they find significant coupling with EEG signal. They report that this phonotactic measure performs better than cohort entropy or surprisal, but they do not report whether they observe significant correlation for those variables.

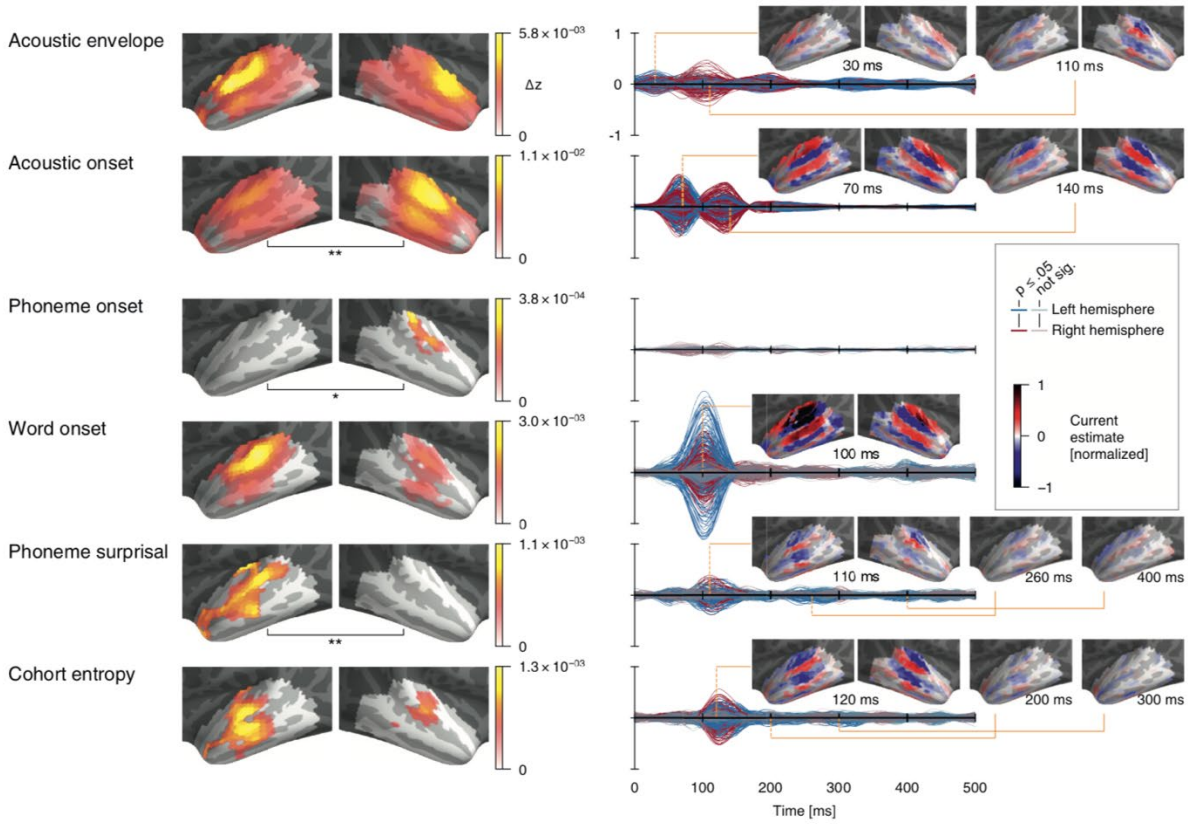


Figure 2. Brain Responses to Single Speaker

Left column: significant predictive power ($p \leq 0.05$, corrected). Colors reflect the difference in *z*-transformed correlation between the full and the appropriately shuffled model. Color-maps are normalized for each predictor to maximize visibility of internal structure, as appropriate for evaluating source localization results: due to spatial dispersion of minimum norm source estimates, effect peaks are relatively accurate estimates, but strong effects can cause spurious spread whose amplitude decreases with distance from the peak. See also Table S2. Right column: Temporal response functions (TRFs) estimated for the reduced model. Each line reflects the TRF at one virtual current dipole, with color coding its location by hemisphere, and saturation coding significance ($p \leq 0.05$, corrected). Anatomical plots display TRFs at certain time points of interest (only significant values are shown), with color coding current direction relative to the cortical surface. Acoustic TRFs were averaged across frequency band for display as visual inspection revealed no major differences apart from amplitude differences between frequency bands. See also Figure S2 and Table S3.

Figure 10. *Figure 2 from Brodbeck, Hong, and Simon (2018), showing significant predictors in TRF analysis of continuous speech. Reprinted from Current Biology, Vol. 28, Issue 24, Brodbeck, Hong, and Simon, Rapid transformation from auditory to linguistic representations of continuous speech, pages 3976-3983.e5, Copyright (2018) Elsevier Ltd, with permission from Elsevier.*

4.1.2.3 Alternative approaches

Finally, four additional MEG studies have relevance for the overall feasibility of using information-theoretic variables with neural data, but are not directly comparable to the studies discussed so far because they employ a different dependent measure or a different means of calculating entropy or surprisal.

Brennan et al. (2014), in a semantic priming paradigm with lexical decision on the target word, test for but fail to find a correlation between cohort entropy at the first phoneme of the target and 10-20 Hz power in the window of a priming effect starting at ~270 ms. They note that it may be difficult to observe entropy effects for monosyllabic words as they exhibit relatively little variation in entropy.

Kocagoncu et al. (2017) compute cohort entropy not from corpus frequencies but from confidence scores provided by participants for each lexical candidate proposed in a gating task with five increments of 25 ms. Description of this gating task in the Methods section of the paper is ambiguous, so it is unclear whether the first fragment ends 50 or 125 ms before the word's corpus-defined uniqueness point. Kocagoncu et al. compute the change in entropy from the first gate to the gating-defined uniqueness point of the word, which is the point at which the word is correctly identified by 80% of participants, with at least 80% confidence. This point is on average 69 ms later than the corpus-defined uniqueness point. They then use this measure as a predictor of neural activity (during a nonword detection task) prior to and following the gating-defined uniqueness point. They find significant effects in STG and supramarginal gyrus from -400 to -376 ms, in MTG from -224 to -180 ms, and in IFG from -244 to -172 ms.

Gwilliams et al. (2020) present naturalistic speech during MEG recording and train decoding models to predict stimulus features from sensor activity. The logic behind this type of approach is that decoding accuracy should vary to the extent that these stimulus features are relevant at a given time point. Gwilliams et al. evaluate decoding accuracy as a function of entropy and surprisal (among other things). For non-first phonemes, they find that decoding accuracy is higher for low surprisal phonemes in the window 120-132 ms after phoneme onset, concluding that processing can start earlier for predictable phonemes. They also (again for non-first phonemes) find higher decoding accuracy for phonemes with higher cohort entropy from 304-328 ms after phoneme onset, interpreting this to mean that maintenance of phonetic information is prolonged during lexical uncertainty.

Finally, Donhauser and Baillet (2020) train an artificial neural network to predict the next phoneme and the next word in naturalistic auditory input given the context, which is information about the previous 35 phonemes (phoneme identities, durations, pauses, and the roughly 10 words that the phonemes make up). They use these probabilistic phoneme predictions to compute a surprisal measure ("surprise") which reflects the conditional probability of a target phoneme given the context, and an entropy measure ("uncertainty") that reflects the uncertainty of the predictive distribution of the next phoneme, given that target phoneme. Their surprisal measure is in line with the calculation of surprisal used in the previously reviewed studies, but uses the probability distribution generated by the neural network rather than corpus frequencies. Their entropy measure, in contrast, reflects entropy over upcoming phonemes (as predicted by the neural network), while the entropy measures in the

previously reviewed studies reflect entropy over possible wordforms that match the input. Using a TRF approach for MEG analysis along the lines of Brodbeck, Hong, and Simon (2018) and Di Liberto et al. (2019), Donhauser and Baillet find effects of both uncertainty and surprise, and that these effects persist when biphone and triphone probabilities are taken into account. Donhauser and Baillet thus argue that their measures do not simply reflect phonotactics.

The problem with this approach, however, is that we cannot know what information in the previous 35 phonemes is contributing to the neural response. A context of 35 phonemes consists of the preceding phonemes in the word being heard as well as several previous words, which then allows for the contribution of syntactic and semantic information from the sentence in which the word appears. The more typically used surprisal measure is conditioned just on the target word, and to what extent the context prior to the target word influences surprisal (as explored by Gaston and Marantz (2018)) is an important question. By showing that uncertainty and surprise are accounting for variance above and beyond biphone and triphone probabilities, Donhauser and Baillet show that their 35-phoneme context is contributing something beyond very local context, and this is useful. However, we cannot know to what extent this 35-phoneme context out-performs even simple word-based surprisal, let alone whether syntactic and semantic information are also important.

4.1.2.4 Summarizing

To summarize the literature on neural data, we start with the seven studies presenting single words or small, controlled contexts (Brennan et al., 2014; Ettinger

et al., 2014; Gagnepain et al., 2012; Gaston & Marantz, 2018; Gwilliams & Marantz, 2015; Kocagoncu et al., 2017; Lewis & Poeppel, 2014). Within this group, phoneme surprisal effects are found in all four studies that test for them (Ettinger et al., 2014; Gagnepain et al., 2012; Gaston & Marantz, 2018; Gwilliams & Marantz, 2015). All of these effects localize to STG, TTG, or MTG (though in almost all cases they are not looked for outside of these ROIs). Their timing, however, is variable. The effects reported by Ettinger et al. (2014) and Gagnepain et al. (2012) are not time-locked to phoneme onsets, but manifest toward the end of the word. Gwilliams and Marantz (2015) test only (and time-lock from) the second to last phoneme, and find effects in both the 100-150 ms range and the 250-350 ms range. Gaston and Marantz (2018) observe effects time-locked from the second, third, and final phonemes, in the 350-450 ms range as well as more sustained effects starting in the 150-200 ms range.

We suspect that this variability is largely due to the variation in analysis approaches. The Gagnepain et al. effect reflects a manipulation specific to that time range, and so it would not have been possible to see surprisal effects elsewhere in that design. The Ettinger et al. analysis assumes a fixed 200 ms lag between stimulus time point and neural response, which severely restricts their ability to detect any correlation that doesn't conform to this assumption. The Gwilliams and Marantz (2015) and Gaston and Marantz (2018) results are more comparable because they both use (spatio-) temporal cluster tests, but the effects reported by Gaston and Marantz are ~50-100 ms later. We acknowledge, though, that the precise onset and offset estimates from cluster tests should not be overly interpreted (see Sassenhagen and Draschkow (2019)). The studies are also conducted in different languages, and

examine different variants on surprisal. We would need to see surprisal effects from two studies with similar stimuli and the same analysis method to know more definitively whether the observed variation is cause for concern.

Entropy effects occur in three of the six studies that tested for them. However, for Gaston and Marantz (2018) the reported entropy effect does not then persist in a model that includes surprisal (so we cease considering it), and Ettinger et al. (2014) do not test a combined model. The effect reported by Kocagoncu et al. (2017) is for an entropy measure that uses gating responses rather than corpus frequencies and is not computed on a phoneme-by-phoneme basis. We conclude then that the evidence for phoneme-level cohort entropy effects in non-naturalistic designs is weak, while phoneme surprisal effects are temporally variable but reliably present.

Among the studies with naturalistic stimuli (Brodbeck, Hong, and Simon, 2018; Di Liberto et al., 2019; Donhauser & Baillet, 2020; Gwilliams et al., 2020), it is more difficult to form generalizations because of the variation in methods, dependent measures, and means of calculating the variables. Di Liberto et al. (2019) do not report whether entropy and surprisal are significant predictors in their dataset (only that a measure of phonotactics performs better). Gwilliams et al. (2020) are not testing for the presence of entropy and surprisal effects, but report higher decoding accuracy for lower surprisal phonemes 120-132 ms after phoneme onset, and higher decoding accuracy for higher entropy phonemes 304-328 ms after phoneme onset.

Brodbeck, Hong, and Simon (2018) report surprisal effects with a peak latency (at 114 ms) slightly earlier than any effects reported by Gwilliams and Marantz (2015) or Gaston and Marantz (2018), the two most comparable single-word

studies. Brodbeck, Hong, and Simon also report a closely following entropy effect with peak latency at 125 ms. This is just a single data point, but taken at face value, it contrasts with the weakness of evidence for entropy effects in the non-naturalistic studies. A potential task or stimulus distinction for cohort entropy effects is unlike the robust appearance of surprisal effects across different types of studies. One explanation could be that the effect size for cohort entropy is smaller and less likely to be detected in (likely) underpowered studies. Another possibility is that there are more (and more varied) processes likely to correlate with surprisal than with entropy. However, we did note earlier in this chapter that while cohort entropy effects seem to require the involvement of wordform representations, there is ambiguity as to whether this is true for phoneme surprisal effects, which could arise because wordform probabilities lead to phoneme probabilities, or because phoneme probabilities are independently tracked. A dissociation in the types of studies that yield cohort entropy and phoneme surprisal effects could support the possibility that only cohort entropy effects arise from wordform probabilities, if wordform probabilities are not invoked by some tasks or stimuli. Stronger support for this dissociation (and its interpretation) will require deliberate task or stimulus manipulations with the same analysis method.

4.1.3 The current studies

We now describe two studies, one completed and one proposed, with the goal of advancing the use of entropy and surprisal as a means to study context effects on the cohort. In both, we have used or plan to use the TRF analysis method that allows us to incorporate acoustic variables and deal with phoneme response overlap.

Experiment 2 is a simple, single-word design that allows us to establish baseline expectations for entropy and surprisal effects in single words, using TRFs, in a task that is not lexical decision and should encourage processing above the form level. We find robust surprisal but not entropy effects, and for non-initial phonemes only.

Given the results of Experiment 2, we describe the proposed design for Experiment 3, which will present simple sentences that provide a syntactic context for auditory word recognition and will allow us to compare facilitatory and inhibitory versions of surprisal and entropy for stimuli in which the variables are de-correlated. This is an advance over Gaston and Marantz (2018), who do not examine the possibility of facilitation, use more highly correlated constrained and unconstrained variables, and can't account for overlapping phoneme responses or acoustic variables. While Brodbeck, Hong, and Simon (2018) deal with some of these methodological issues, and their stimuli occur (naturalistically) in syntactic contexts, they do not evaluate context-constrained variables. In full narratives, it is extremely difficult to quantify the many different influences that various aspects of the context might have on the cohort. Fully accounting for these influences is a longer-term goal, but our starting point is isolating the contributions of a relatively simple contextual constraint (syntactic category), whose expected influence on the cohort is more easily quantified.

4.2 Experiment 2

In this study, we presented single, monomorphemic words with randomly occurring semantic relatedness probes, while recording MEG data. We analyze the

neural response to auditory word recognition using temporal response functions. This is the first time this analysis technique has been applied to single words rather than continuous speech. We believe the benefits of this new application to be largely with respect to our understanding of cohort effects in single-word recognition when acoustic variables and phoneme response overlap are accounted for. However, the analysis of single words with the TRF approach is also useful because there should be far less overlap with the neural response from the previous word than occurs in continuous speech. This is a relatively new method for continuous speech in MEG, and establishing consistency between different stimulus types would be a helpful validation of the approach. Any consistent divergence, of course, would also be informative.

Our primary concern is how basic entropy and surprisal effects manifest in a design with no other manipulations, and how they compare to previous effects obtained from (1) continuous speech analyzed with the same method and (2) single words analyzed with different methods. We are also interested in whether the lack of first phoneme cohort effects noted by two previous studies replicates in our data. This experiment is intended to lay the groundwork for the use of TRF analysis for auditory word recognition in simple syntactic contexts.

4.2.1 Materials & Methods

4.2.1.1 Participants

We collected data from 24 people (a subset of those who participated in the study reported in **Chapter 5**). All participants were right-handed, native speakers of English, and seven were also native speakers of additional languages. None reported

history of neurological or linguistic impairment, brain injury, or hearing loss. All reported normal or corrected-to-normal vision. The procedure was approved by the University of Maryland Institutional Review Board and all participants provided written informed consent. Participants were compensated with their choice of $15 or 1 course credit per hour of participation. The full session (including Experiment 4 and the localizer reported in **Chapter 5**) lasted 2 hours.

One dataset was excluded without looking at the data because the participant was very tired and an earbud fell out during the experiment. After this exclusion, we computed accuracy on the relatedness task and excluded any participant with accuracy lower than a cutoff 1 standard deviation below the mean. This excluded three of 23 participants. After preprocessing, two additional datasets were excluded due to extreme noise. 18 datasets are therefore included in our analysis.

4.2.1.2 Stimuli

The Massive Auditory Lexical Decision (MALD) database makes lexical decision data and recordings of 26,793 words (and 9592 pseudowords) freely available (Tucker et al., 2019). The timing of phoneme boundaries in each recording is also provided. Rather than make new recordings of our chosen stimuli, we opted to use MALD recordings because we would then have lexical decision data, phoneme boundaries, and a variety of other lexical variables readily available for all of our stimuli. This would also allow us to consider our data in the context of other published experiments and analyses using the MALD dataset. We also hope to make our MEG data freely available to add to the richness of this open dataset.

To create our stimuli, we started with the full list of real words included in MALD. We then removed any items that were not monomorphemic as indicated by MALD, and then from this set, anything still tagged as multimorphemic according to CELEX (Baayen et al., 1995). We removed any items with missing information in the MALD dataset of item-level variables. We then removed any items whose list of parts of speech included: Preposition, Interjection, Name, Unclassified, Conjunction, Pronoun, Determiner, Letter, Not, Ex, Article, To. Finally, we sorted the items according to frequency and removed the lowest 10%. We were left with 4144 items.

From this list, we removed by hand any additional multimorphemic items that had not been caught by the MALD or CELEX tagging. Because the recordings were made by speakers of Canadian English, we also removed any item for which the pronunciation in the recording was noticeably divergent from American English. Finally, we removed inappropriate and particularly evocative words. This left us with a total of 2676 items. From a random sample of 1500 of these items, we removed any homophones, and then of those remaining we used a random sample of 1000 in the experiment. 1000 was our target number because the length of the experiment had to be kept to a maximum of 20 minutes in order to fit into the recording session along with the studies reported in **Chapter 5**.

The full lists of stimuli and probes (see below), are reproduced in **Appendix A**. These items, as well as associated stimulus variables from MALD, are also available for viewing and download on OSF.

4.2.1.3 Task

To ensure attention, we pseudo-randomly presented a semantic relatedness probe after some words. Probes were single words, presented visually, and participants were instructed to answer as to whether the probe was related at all to the word that they had just heard. There was no advance warning for probe trials, so attention was required on each auditory trial in case it was to be followed by a probe trial.

We selected this task so that it would apply equally well to all types of words, and because we did not want button presses to occur on critical trials (as would happen in e.g. lexical decision). The probe trials for which we expected participants to answer "No" were selected randomly from the list of eligible words that we did not end up using for auditory trials. Probe trials for which we expected participants to answer "Yes" were synonyms taken from the WordNet (https://wordnet.princeton.edu) page of the preceding auditory item, and were also monomorphemic so as not to be trivially distinguishable from "No" trials. There was no overlap between probe words and words used in auditory trials. Which auditory trials would be followed with a probe were randomly selected. "Yes" and "No" probes were equally distributed.

4.2.1.4 Procedure

This study was always at the end of the experimental session, following Experiment 4 and the localizer from **Chapter 5**. Participants wore foam earbuds and volume was adjusted to the comfort level of the participant. Participants lay supine inside the magnetically shielded room and looked at a screen overhead, while holding

a button box in each hand. They were instructed that they would hear a long series of random words, and that they should simply listen to the words. They were instructed to keep their eyes open because a probe word would occasionally appear on the screen with a question mark, and they were to answer (with left hand for No, right hand for Yes) whether the word on the screen was related in any way to the word they had heard just before it.

We used Presentation (Neurobehavioral Systems, Inc., www.neurobs.com) to present the experiment. Our parameter and scenario files are available for download from OSF. There were 1000 auditory trials interspersed with 97 probe trials. The amount of time between trials was 267 ms. A visual fixation cross was on screen continuously during auditory trials and during the inter-trial interval. Each auditory trial simply consisted of presentation of the auditory stimulus, and lasted the length of the auditory stimulus. Probe trials were pseudo-randomly distributed throughout the experiment with a maximum interlude of 20 trials between probes. During probe trials, a probe word would be presented visually, with a question mark (e.g. "podium?"). The probe stayed on the screen until the participant pressed a button to indicate whether the probe was related in any way to the word that had played immediately before it.

The experiment lasted roughly 17 minutes. There was no built-in break, but participants were instructed that if they wished to take a break, they should simply delay their button press on a probe trial.

4.2.1.5 MEG data collection

Before recording, we used a Polhemus 3SPACE FASTRAK to digitize participant head shapes as well as the positions of five affixed marker coils. These marker coils were used to record head position relative to the MEG sensors before and after each study in the session.

We recorded continuous MEG data, inside a magnetically shielded room, with a 160-channel axial gradiometer whole-head system (Kanazawa Institute of Technology, Kanazawa, Japan). Our sampling rate was 1000 Hz, and we used an online 60 Hz notch filter and 200 Hz low-pass filter.

4.2.1.6 MEG pre-processing

We processed the data using mne-python version 0.19.2 (Gramfort et al., 2013, 2014) and eelbrain version 0.31.1 (Brodbeck, Proloy Das, et al., 2019). The TRF analysis was conducted with mne-python version 0.20.5 and eelbrain version 0.32.dev0.

During file conversion with mne-python's kit2fiff GUI, we excluded any faulty marker measurements. We co-registered each digitized head shape with the Freesurfer (Fischl, 2012) "fsaverage" brain, using mne-python's co-registration GUI. We first used rotation and translation to align the digitized head shape and average MRI by the three fiducial points. We then used rotation, translation, and 3-axis scaling to minimize the distance between digitized head shape and average MRI points using the iterative closest point (ICP) algorithm. Convergence was always achieved within 40 iterations. For one participant, outlying points on the digitized head shape were removed between fitting to the fiducials and applying ICP.

Flat channels were automatically removed, and we used temporal signal space separation (Taulu & Simola, 2006) for removal of extraneous artifacts, with a buffer duration of 10 seconds.

We then used ICA (independent components analysis, with extended infomax method) for removal of ocular, cardiac, and other extraneous artifacts. Decomposition was performed using all of the data, with a 1-40 Hz filter applied, and then we visualized the components with a 1500 ms epoch from each word's onset. After selecting ICA components for removal, we proceeded with a 1-20 Hz filter, and down-sampled the data to 100 Hz for analysis.

To compute the noise covariance matrix, we used 2 minutes of empty room data recorded before or after each session. We defined the source space on the white matter surface with a four-fold icosahedral subdivision, with 2562 sources per hemisphere. Orientation of the source dipoles was fixed perpendicular to the white matter surface. For minimum norm current estimation, we used the MNE noise normalization method with SNR of 1 and did not use depth weighting.

The anatomical labels we used to create search areas for the spatiotemporal cluster tests came from the Freesurfer 'aparc' parcellation.

4.2.1.7 Analysis

Behavioral data

Mean accuracy was computed after the exclusion of one participant a priori. The mean number of correct probe responses was 73.6 (out of 97) with a standard deviation of 18.4. The number of correct probe responses was lower than one standard deviation below the mean for three participants, so they were excluded from

further analysis. One participant answered 13 of 97 probes correctly. We kept this participant in the dataset because this was so far below chance that we assumed they had reversed which hand they were supposed to use to make Yes and No responses.

Neural data

Our TRF analysis largely followed Brodbeck, Hong, and Simon (2018), with some differences noted below. Brodbeck, Presacco, and Simon (2018) describe the method in detail; we re-describe it here in a manner intended to be more accessible for readers not overly familiar with linear kernel estimation.

For each stimulus variable of interest, a time series was created indicating the value of the predictor at each time point in the stimulus. We used the stimulus variables that had been found significant in the Brodbeck, Hong, and Simon (2018) data: acoustic envelope, acoustic onset, word onset, phoneme onset, cohort entropy, and phoneme surprisal. For acoustic predictors (acoustic envelope and acoustic onset), the value of the predictor can vary continuously at each time point. See Brodbeck, Hong, and Simon (2018) for description of how these variables are calculated. For lexical predictors, values are non-zero only at time points labeled as phoneme onsets (i.e., lexical predictors consist of impulses at phoneme onsets). Of these lexical predictors, the phoneme onset and word onset predictors each consist of binary impulses, while the predictors for cohort entropy and phoneme surprisal consist of impulses that are scaled continuously according to the variable value at that phoneme. In **Figure 11** below, we reproduce Figure 1 from Brodbeck, Hong, and Simon, which illustrates how these stimulus variables are modeled. Entropy and surprisal are calculated using frequency information from SUBTLEX (Brysbaert &

New, 2009) and phoneme sequence information from the CMU pronouncing dictionary (Weide, 1994).

Our study did not actually present a continuous stimulus (rather, individual words with short intervening pauses), but a single time series reflecting predictor values (or pauses) throughout the entire experiment could still be created. Probe trials were modeled simply as silence. The timing of phoneme onsets was taken from the forced aligner information made available with the MALD recordings.

For each participant, at each source point, each stimulus variable is then convolved with a -100 to 500 ms kernel (or "temporal response function") to create an estimated response for that predictor for the full duration of the stimulus. The sum of these linear convolutions across all predictors is the estimated response for the source point, for which correlation with the actual neural response can be evaluated. The kernel (or TRF) for a predictor can be thought of as an estimated evoked response occurring in response to each time point in which that predictor is non-zero, and it scales with the predictor value. We are estimating the evoked responses for all predictors we think might be relevant, summing them when they overlap in time, and then summing over all predictors to create the estimated neural response.

The first step in this process is, for each source point, to jointly estimate an optimal kernel for each predictor, using a coordinate descent algorithm. The kernel has a window of -100 to 500 ms around each event. To model the neural response at a given time point, we take, for a given predictor, the predictor value at time t in the stimulus multiplied by the kernel value at 0 ms in that predictor's kernel, the predictor value at time t-10 ms in the stimulus multiplied by the kernel value at 10ms, the

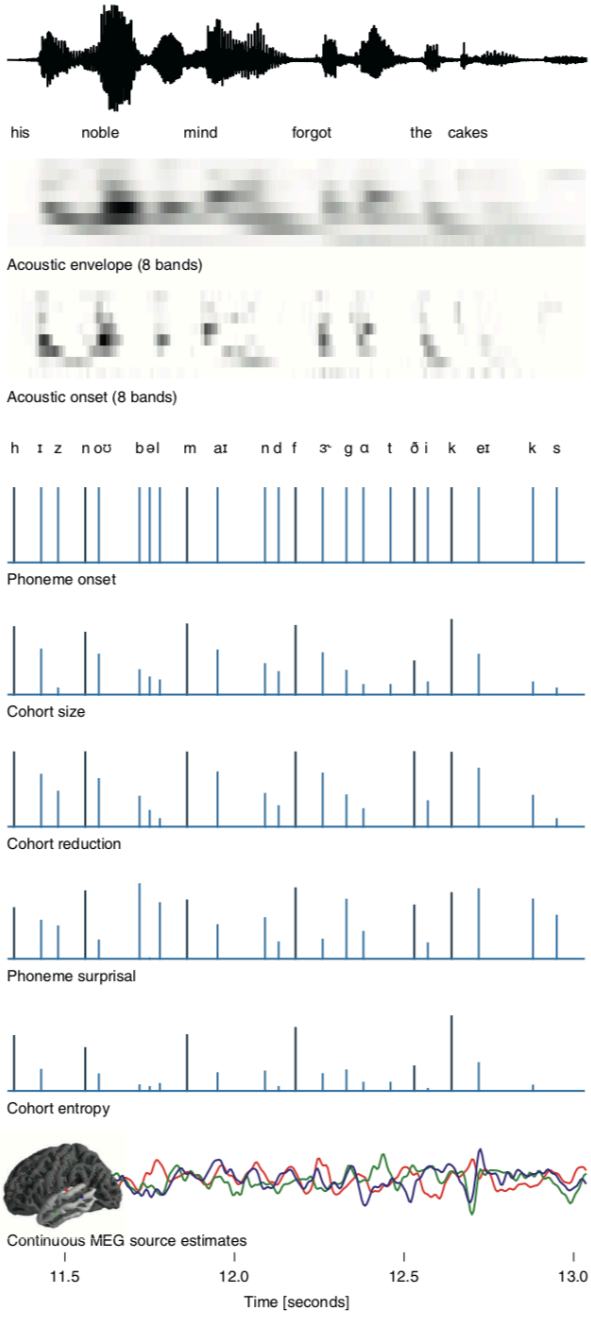


Figure 1. Analysis Framework, Illustrated with an Excerpt from One of the Stimuli

The acoustic waveform (top row) is shown for reference only. Subsequent rows show the predictor variables used to model responses to a single speaker. Acoustic predictors were based on an auditory spectrogram aggregated into 8 frequency bands. For the phoneme-based predictor variables, the initial phoneme of each word is drawn in black, whereas all subsequent phonemes are drawn in blue. The last row contains estimated brain responses from three virtual current dipoles, representative of the modeled signal. The anatomical plot of the cortex is shaded to indicate the temporal lobe, the anatomical region of interest (only the left hemisphere is shown, but both hemispheres were analyzed). See Table S1 for correlations between different predictor variables and Figure S1 for corresponding scatter-plots (of the phoneme-based predictor variables).

Figure 11. *Figure 1 from Brodbeck, Hong, and Simon (2018), illustrating how different types of stimulus variables are represented in their analysis. Reprinted from Current Biology, Vol. 28, Issue 24, Brodbeck, Hong, and Simon, Rapid transformation from auditory to linguistic representations of continuous speech, pages 3976-3983.e5, Copyright (2018) Elsevier Ltd, with permission from Elsevier.*

predictor value at t-20 ms multiplied by the kernel value at 20 ms, etc, repeated up to t-490 ms in the stimulus and 490 ms in the kernel, and we sum over each of these timepoints. This is because we are modeling the overall response as being the sum of the response contributed at that timepoint by each kernel that has had its onset in the last 500 ms. The 10 ms time steps are because the data are down-sampled to 100 Hz. We also repeat this procedure for 100 ms in the other direction, and add this value to the sum. This step helps account for any advance cues in the auditory signal. We do this for each predictor and sum over all predictors. Finally, we add an error term.

For a given predictor at a given time point, for each overlapping kernel contributing to the response, if the predictor value for that kernel or the kernel value for that time point is zero, there is no contribution to the response from that predictor for that kernel point. The better the predictor, and the higher its value at that time point, the higher its kernel value should be.

How, in practice, can we estimate the optimal kernel for many predictors at once in a continuous time series? The algorithm to estimate the optimal kernel starts with a value of zero for each time point in each predictor's kernel. At each time point in the duration of the kernels, on each iteration of the algorithm, the predictor for which an increase in its kernel value at that time point leads to the largest reduction in error predicting the summed response in the training data is identified, and this change is evaluated for its error reduction with held-out test data. Iterations stop when error can no longer be decreased in the training data or increases in the test data. To maximize the amount of training data available, we use four-fold cross validation so that each partition of the data serves once as the test data when the other three

partitions are training data. The optimal kernels estimated on each of the four repetitions are then averaged together.

Once the averaged, optimal TRF for a predictor is identified, it is convolved with the stimulus variable for a fifth held-out partition of the data, and the convolutions for each predictor are summed to create the estimated response for that partition. Once an estimated response has been created for each partition, from a kernel estimated via the other four, the five partitions are recombined, and correlation between the estimated and actual response over the full course of the stimulus at this source point can be computed. Thus, the four-fold cross-validation described above for estimating the kernel is actually nested within five-fold cross-validation; on each repetition, four partitions are used to estimate the kernel, and that estimated kernel is tested using the fifth partition. The rationale for this fifth "fold" in the cross-validation process is that we do not want to use the actual neural response to help create an estimated response, and then ask about the correlation between them; we instead use some of the data to create an estimated response, and then see how well this estimated response correlates with the remaining data.

To evaluate whether a given predictor significantly improves the estimated neural response (i.e., increases the correlation between the estimated and observed neural response), we go through the entire described process for a model that contains the predictor, and a model that doesn't[5]. At each source point, then, we have two

[5] Note that for Brodbeck, Hong, and Simon (2018), cross-validation was not used for model comparison. Instead, cross-validation was used to estimate the optimal kernel, and then this optimal kernel was convolved with the complete stimulus variable, for

correlation values for each participant, and we can compute a *t*-value for the difference between the two correlation values. We use a permutation cluster test over the left and right temporal lobes to ask if there are clusters of sources in which the *t*-values are elevated, indicating that correlation between the estimated and actual response was significantly higher when the variable of interest was included in the model. If it was not, we remove that variable from the model. This approach can also be used simply to ask whether the correlation values for a specific model are significantly different from zero.

Having established that each predictor in a model is contributing significantly to the model fit with the observed data, we re-estimate a final set of TRFs using the full dataset rather than holding out the fifth partition. We can then examine the individual predictors' TRFs in order to ask at what source and time points they are significantly different from zero. A specific predictor's TRF has a value at each source and time point, for each participant. We can compute, at each source and time point, a *t*-value for the difference between the TRF value and zero, in this set of participants, and evaluate significance with a permutation cluster test over both sources and time points.

Our intention was to use model comparison to evaluate whether there were entropy and surprisal effects in our data, and to follow up on the lack of such effects

---

an estimated response of the entire dataset. This was done using the true predictor, and then again using a shuffled version of that predictor which should have had no predictive value. They then tested for a difference between the correlation values for the model with the true and the shuffled predictor. We now use cross-validation for this step, following Brodbeck et al. (2019), because it better handles correlation between predictors.

at the first phoneme noted by Brodbeck, Hong, and Simon (2018) and Gaston and Marantz (2018). We then planned to examine the TRFs of significant predictors for timing and location information.

4.2.2 Results

Our starting model included acoustic envelope, acoustic onset, word onset, non-initial phoneme onset (i.e., phoneme onsets that are not the first phoneme onset, which is co-extensive with word onset), phoneme surprisal, and cohort entropy at each phoneme. Our first question was whether first phonemes should be excluded from the phoneme surprisal and cohort entropy estimates, following the prior evidence discussed above on this point. To answer this question we compared the starting model to a model in which surprisal and entropy at the first phoneme are modeled as separate predictors from surprisal and entropy at non-initial phonemes. The more complex model, in which they are modeled separately, was significantly better (t_{max} = 4.02, p = .010). There was no significant difference upon removal of both first phoneme surprisal and first phoneme entropy simultaneously (t_{max} = 2.61, p = .499), removal of first phoneme surprisal alone (t_{max} = 3.30, p = .375), or removal of first phoneme entropy alone (t_{max} = 2.11, p = .862). Thus, we proceeded without either predictor. The lack of first phoneme cohort effects is consistent with Brodbeck, Hong, and Simon (2018) and Gaston and Marantz (2018).

Having removed entropy and surprisal for the first phoneme as predictors, we asked whether correlation improves when entropy and surprisal of the second phoneme are tested as separate predictors in the model from entropy and surprisal of subsequent phonemes. The model in which they are not separate predictors is

significantly better (t_{max} = 7.69, p < .001), indicating that (other than word onsets) the ordinal position of the segment is not relevant to modeling entropy and surprisal. Similarly, although word onset was already being modeled separately from non-initial phoneme onsets, to confirm that this reflects a qualitative difference about word onsets we also checked whether there was any significant improvement when the second phoneme onset was modeled separately from subsequent phoneme onsets. There was not (t_{max} = 2.67, p = .572).

Having resolved these preliminary questions about how best to formulate the model, we tested our primary research questions: do phoneme surprisal and cohort entropy improve the estimated neural response in a single-word design? We found that indeed, a model with (non-initial) phoneme surprisal was significantly better than a model without it (t_{max} = 7.64, p < 0.001). However, removing (non-initial) cohort entropy led to no significant difference (t_{max} = 2.93, p = .400). Even when surprisal is absent from the model, removal of entropy still does not yield a significant difference (t_{max} = 3.18, p = .127). This contrasts with Brodbeck, Hong, and Simon (2018), for whom entropy and surprisal were both significant.

Finally, we tested the remaining variables, and as expected, word onset was a significant contributor (t_{max} = 5.59, p < .001), as were the acoustic predictors (t_{max} = 8.20, p < .001).

Having established that model fit with the observed neural data is improved by the acoustic predictors, word onset, non-initial phoneme onset, and non-initial phoneme surprisal, we then examined the estimated response for phoneme surprisal, which for the left hemisphere shows a first peak at 70 ms and a second peak at 290

ms, in posterior superior and middle temporal areas. The right hemisphere response appears somewhat more temporally diffuse. **Figure 12** shows the TRF for each source point (**Figure 12A**), the difference in correlation between the estimated and actual response at each source point when the model does or doesn't include surprisal **(Figure 12B**), and the current estimate at each source point for the two peaks in the TRF (**Figure 12C**). Note that while model comparison was carried out with a -100 to 500 ms kernel, for visualization purposes only we extend the TRF to 600 ms.

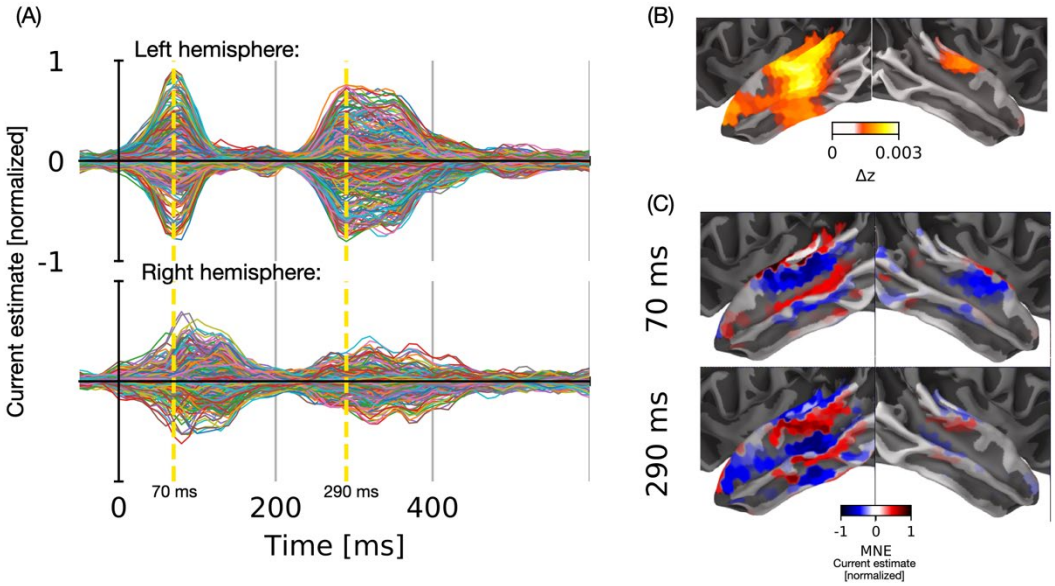


Figure 12. *TRF results for phoneme surprisal.* ***(A)*** *TRF for each source point for the left and right hemisphere. Latency of left hemisphere peaks is marked with yellow dashed lines.* ***(B)*** *Difference in correlation between estimated and actual response at each source point when the model does or doesn't include surprisal, for left and right hemisphere.* ***(C)*** *Current estimate at each source point for the early and late peak, for left and right hemisphere.*

4.2.3 Discussion

4.2.3.1 Summary

Experiment 2 was intended to establish baseline expectations for entropy and surprisal effects in single words, using the TRF analysis method. As have many previous studies, we found that phoneme surprisal is a significant predictor of neural activity. The spatial distribution of the effect, along posterior superior and middle temporal gyrus, appears largely the same as that observed by Brodbeck, Hong, and Simon (2018) (see **Figure 10** for comparison) as well as previous single-word studies that did not use the TRF method.

We see that the TRF for surprisal appears to peak twice, first at 70 ms after phoneme onset and then again at 290 ms. This pattern differs from Brodbeck, Hong, and Simon, who observed only an early peak for the surprisal response (at 114 ms) in continuous speech, but is more in line with Gaston and Marantz (2018) and Gwilliams and Marantz (2015), both of which report early and late surprisal effects. Nevertheless, the effect is still earlier than in any of the previous studies. Even in comparing the early peaks from the two TRF results (this study and Brodbeck, Hong, and Simon (2018)), we cannot say whether the timing difference (70 vs. 114 ms) is due to the use of single words versus continuous speech, or for example due to specific properties of the speaker, biases of the forced aligners, or the presence or absence of co-articulation from the previous word. Modulation of the timing of surprisal effects will need to be explored in future work.

Like both Brodbeck, Hong, and Simon (2018) and Gaston and Marantz (2018), the surprisal effect we found was only for phonemes that are not the first

phoneme in the word. Because surprisal is driven by the conditional probability of a phoneme given the preceding input, a surprisal effect at the first phoneme would have reflected probability conditioned on silence, or prediction of a specific phoneme as the first phoneme in an upcoming word that is expected to occur.

Replicating prior surprisal findings is extremely useful for us in both proceeding to Experiment 3 and establishing this measure as a reliable index of lexical processing. It appears that surprisal effects ~100 ms post-phoneme-onset are particularly robust. The later-stage surprisal effect observed in the single-word studies deserves more investigation in future work.

Unlike Brodbeck, Hong, and Simon (2018), we did not find effects of cohort entropy. In the Introduction of this chapter, we pointed out that all of the studies in which entropy effects are weak or non-existent use single words or phrases rather than naturalistic speech as stimuli. Our null effect, with single-word stimuli, fits this pattern. Prior to this result, one explanation for the entropy pattern could have been that the TRF analysis method is more sensitive to entropy effects because it accounts for more variation in the speech signal. However, we use nearly the same TRF analysis method as Brodbeck, Hong, and Simon, so this explanation now seems unlikely.

4.2.3.2 Task or stimulus effects on involvement of the cohort

Another explanation we suggested was that entropy effects are driven by probabilities tracked at the level of wordform representations, while phoneme surprisal effects are driven by probabilities tracked only at the phoneme level. If this is the case, the data could be indicating that in single-word paradigms, word

recognition does not involve incremental, phoneme-by-phoneme modulation of wordform probabilities. Specifically, we mean designs in which single words are recorded separately and then presented sequentially with intervening pauses. This definition therefore includes the Gaston and Marantz (2018) study in which short phrases are constructed out of separate audio files. Recognition of single words presented in this way could in theory occur without initial involvement of a cohort of possible candidates. Instead, listeners could wait until some or all of the word has been perceived before retrieving lexical information that matches the wordform. We believe the tasks employed in the single world paradigms we have reviewed could all be completed with such a strategy.

Gagnepain et al. (2012) use a pause-detection task with an inter-stimulus interval of 1850-2150 ms between words. It is unclear whether lexical access would be necessary at all to complete this task. In the studies using lexical decision or nonword detection tasks, the non-words were pronounceable (Brennan et al., 2014), phonotactically legal (Kocagoncu et al., 2017), contained only legal bigrams (Lewis & Poeppel, 2014), or were actually very low frequency real words (Ettinger et al., 2014). In all cases, then, verification of the wordform would be necessary on every trial to perform the task, as the presence of legal phonotactics alone would not be sufficient. However, it does seem that this could be done once the entire wordform has been perceived. Gaston and Marantz (2018) use a phrase acceptability task, which would require lexical access, but again this could occur once the target wordform is known. Finally, our Experiment 2 used random semantic relatedness probes for the experimental task, and we expected that this would encourage lexical access.

However, as in all of the previously described studies, this task could be completed with lexical access delayed until the full wordform has been perceived.

By contrast, the speed of naturalistic speech and the imperative to recognize words quickly for the sake of sentence-level interpretation could be what drives the cohort process (and therefore leads to entropy effects) in continuous speech paradigms. Under this explanation for the lack of entropy effects in single-word paradigms, we would have to postulate that sensitivity to phoneme probabilities is preserved even when wordform access is postponed. We might expect that entropy effects would be observed for single words if a task were designed such that earlier identification of the word is encouraged and the cohort process becomes more advantageous.

This explanation extends in an interesting way to another generalization in the literature: that the presence of multimorphemic words in a study also seems to be relevant for whether or not entropy effects occur. The words appearing in the Brodbeck, Hong, and Simon (2018) stimuli are a mix of multi- and mono-morphemic, as occurs naturally. We have previously dismissed the few entropy effects that have been reported for single-word paradigms as constituting weak evidence, for various reasons. However, the two that stand (Ettinger et al., 2014; Kocagoncu et al., 2017) both included multimorphemic words, while all of the studies that did not find entropy effects (Brennan et al., 2014; Gagnepain et al., 2012; Gaston & Marantz, 2018; Lewis & Poeppel, 2014) used only monomorphemic words. Our study joins this list, using monomorphemic words and failing to find an entropy effect.

We suggest that the relevant difference between monomorphemic and multimorphemic words is that multimorphemic words are a form of continuous speech. We have suggested that the reason for the lack of entropy effects in single-word studies could be that the pauses between words make it unnecessary to engage in phoneme-by-phoneme access to candidate wordform representations, and participants instead wait until the word has ended to make contact with the mental lexicon. Multimorphemic words are two separate lexical units without an intervening pause, and so even in single-word designs this could motivate incremental updating of the cohort so that the first morpheme can be recognized in time to begin processing the second.

Note that this explanation does not apply to the processing of mono- versus multimorphemic words in general, or even to the processing of single words in general, but instead to studies in which *only* monomorphemic words are presented. What we are describing is a strategy in which participants delay lexical access until there are fewer or there is just a single lexical candidate remaining, because it is very clear in the experiment that there is always time to do so. This could be tested with manipulations of the inter-stimulus interval in single-word paradigms. If this is a pervasive strategy in single-word studies, it would have important implications for our understanding of existing neural and behavioral data, and would motivate increased use of more naturalistic designs (or, at least, multimorphemic words). We acknowledge that, given pervasive assumptions about auditory word recognition, it would be surprising if cohort competition turned out to be task-modulable rather than automatic. However, it is a possibility that we cannot rule out.

4.2.3.3 Task or stimulus effects on wordform versus lexical involvement

An alternative explanation to that described in the previous section for an entropy/surprisal distinction would be that competition between candidate wordforms does occur phoneme by phoneme even in single-word recognition, but that the tasks employed do not require lexical access (i.e., involvement of the level of the abstract lexical representation rather than just the wordform) or they allow it to be delayed until only a single or very few candidates remain. If entropy effects are driven by abstract lexical representations for all cohort competitors, this would explain why they do not occur in the single-word paradigms. Again, the idea is that continuous speech (and multimorphemic words) do require lexical access to occur for cohort competitors. This explanation allows for the possibility that surprisal arises via competition purely at the phoneme level, but it also allows for surprisal reflecting competition at the wordform level. Surprisal reflecting competition at the wordform level would seem to fit better with the syntactically conditioned phoneme surprisal effects reported by Gaston and Marantz (2018), unless strictly phoneme-level predictions can also be rapidly influenced by context.

If both of these explanations are incorrect and surprisal effects do reflect lexical-level access for cohort competitors, we would have to postulate some other process, engaged by continuous speech and not otherwise, on which cohort entropy effects are contingent. For example, if correlations between neural activity and cohort entropy are not driven by entropy per se but by a process that is sensitive to entropy, an effect of single words versus continuous speech on that process would make it

appear that entropy effects were also modulated. Within-study manipulations of task and stimulus type will likely be necessary for further clarity.

4.2.3.3 Moving forward

Our goal in this study was to establish a better understanding of basic, unconstrained cohort effects when the stimulus is modeled as thoroughly as current methods allow. Our replication of previous surprisal effects allows us to proceed in using this variable to study syntactic constraints on the cohort in Experiment 3. Our failure to find an effect of cohort entropy is in line with previous failures to do so in single-word paradigms.

In the previous section, we have discussed what the asymmetry in surprisal and entropy effects might reflect about the levels of representation that are driving these effects. We raised the possibility that surprisal effects do not reflect wordform-level processing. This would make it less likely that surprisal could then show the effects of a syntactic constraint, though syntactically conditioned phoneme probability effects are not impossible.

Though we did not find entropy effects in Experiment 2, we will still test for them in Experiment 3. The most convincing evidence for entropy effects comes from a study using continuous speech rather than single words. Experiment 3, presenting sentences rather than single words, is a form of continuous speech. We cannot be sure that we will observe entropy effects, as it is of course possible that the previously observed entropy effect was a false positive, or that disconnected sentences are not similar enough to continuous speech in whatever property it is that is relevant for entropy effects. Testing for entropy effects in Experiment 3, even if not informative

for the question of the syntactic constraint, will aid in advancing our understanding of what drives entropy effects.

Experiment 2 has also helped validate the TRF approach as a promising option for future work in single-word paradigms. We have reported the results of a very constrained analysis here, but this dataset does have the potential to inform many other questions. For example, we are beginning to explore effects of uniqueness point and whether the effects of common lexical predictors like word frequency are in fact better understood as being time-locked from the uniqueness point rather than word onset. Incorporating effects of syllable structure in our models would be another useful direction for development.

4.3 Experiment 3

Having established baseline expectations for entropy and surprisal effects in a single-word paradigm with the TRF analysis method, we now propose a study intended to follow up on both Gaston and Marantz (2018) and on Experiment 1 in **Chapter 3**. Data for this new study could not be collected due to the pandemic, so we report only the design. **Table 5** provides the concept for the stimuli.

Gaston and Marantz (2018) calculated two syntactically constrained versions of entropy and surprisal in which (1) wordforms that could not appear in the syntactic context were removed from the hypothesized cohort or (2) all wordforms in the cohort had their frequencies updated to reflect only their frequency of occurrence in the syntactic context. For example, a wordform with a frequency of 10, occurring 60% of the time as a noun and 40% of the time as a verb, would compete with a frequency of 4 in a verb-constraining context. Both of these implementations of the

constraint assumed that it was inhibitory, such that incompatible wordforms could not compete, or competed less. Gaston and Marantz found effects of both constrained and unconstrainted surprisal, interpreted to mean that perhaps sensitivity to the original cohort was somehow maintained despite the constraint.

In Experiment 1 of this dissertation, I reported data from the visual world paradigm that is consistent only with a facilitatory constraint or no constraint at all. If there is a facilitatory syntactic constraint, this could explain why Gaston and Marantz (2018) observed correlation with both unconstrained and (inhibitorily) constrained surprisal: each is capturing a different aspect of the facilitated distribution, since it is true both that contextually inconsistent wordforms are competing, and that contextually consistent wordforms are competing more than contextually inconsistent ones. Therefore, in this proposed study, we explicitly compare cohort measures calculated under the assumption of no syntactic constraint, a facilitatory constraint, and an inhibitory constraint, attempting to resolve questions raised by Experiment 1 as well as by Gaston and Marantz (2018), and provide the necessary evidence for or against a facilitatory constraint.

Table 5*. Concept for sentence stimuli for Experiment 3. For nouns and verbs, at each phoneme, we will compare surprisal and entropy values that either do not reflect the syntactic context or reflect a facilitatory or inhibitory syntactic constraint.*

Example sentences
The birds planned the candles.
The remarks hid the bystanders.
The cakes described the kangaroos.

This study incorporates a variety of improvements over previous designs. One serious issue for Gaston and Marantz (2018) was the high degree of correlation between constrained and unconstrained variables. Experiment 3 will therefore use a set of target stimuli selected specifically such that constrained and unconstrained variables are as de-correlated as possible. Stimuli will be simple sentences with a repetitive structure (The Noun Verb the Noun) to maximize both syntactic expectedness and analyzable data per trial, and to allow analysis of both nouns and verbs. Rather than natural sentences, we use random pairings of noun-verb-noun so that lexical predictability is absent in the experiment and we avoid the influence of lexical prediction on the cohort to the extent possible. However, we will still record these sentences as continuous speech, so the results of this study should aid in our understanding of the variation in entropy effects discussed in the previous sections. Finally, we use a TRF analysis to appropriately model overlapping phoneme responses and acoustic variables.

4.3.1 Design features

4.3.1.1 Structurally predictable sentences

Gaston and Marantz (2018) studied syntactic context by presenting minimal phrases: “the clash persisted,” “to gleam brightly.” In this study, we scale up to sentences rather than phrases, with a fixed template: Determiner Noun Verb Determiner Noun. This makes our stimuli and the syntactic contexts somewhat more naturalistic, and allows us to present and analyze both nouns and verbs in the same trial type. In the Gaston and Marantz design, target words were noun/verb homonyms, and interpretation of the phrase rested on correct comprehension of the

context word as either “to” or “the.” Similarly, in Experiment 1 of this dissertation, the syntactic category of the target word was cued in advance only by the immediately preceding “to”/ “the”. In this design, the syntactic category of each word is fixed and predictable on every trial. This intentionally maximizes the potential strength of the constraint, allowing for more certainty about each word’s syntactic category than would typically be possible. Any context effects that we observe should of course be followed up on in a design whose sentences do not have a fixed structure, to better understand in what circumstances the constraint applies. Our study, however, is focused on how the constraint is implemented when it does apply. By not using category cues that occur only immediately before the target word, we also address the concern from Experiment 1 that there was not sufficient time for the identity of the context cue word to be processed and for the category expectation to be generated before the onset of the target word.

The deterministic syntactic category expectations allowed by this design also justify the simplified manner in which we construct our hypothesized cohorts for the calculation of constrained entropy and surprisal. For nouns, we boost the frequencies of items that can be nouns or inhibit those that can’t. For verbs we boost the frequencies of items that can be verbs, or inhibit those that can’t. We do so because we do not yet know to what extent category expectations in normal, un-predictable contexts are non-deterministic (i.e., to what extent both nouns and adjectives are expected after “the”), or how this might be implemented. A design in which the occurrence of nouns after “the” is perfectly predictable makes these simplifications more justified, and should allow, subsequently, for incremental exploration of the full

hypothesis space for the constrained cohort, ideally using a parsing model that provides a distribution over syntactic categories at each position in the sentence.

Using predictable syntactic structures will also allow us to address the unexpected interactions reported by Gaston and Marantz (2018), in which surprisal effects were generally stronger for items presented in noun context than verb context. We can then ask whether nouns and verbs in our design are similarly differentiated despite more comparable category cues. We also have the opportunity to examine nouns in subject vs. object position. The subject noun is particularly interesting because there is no possible influence of the lexical content of previous words in the sentence.

Finally, the use of simple sentences in which all content words are analyzable targets allows us to maximize the proportion of MEG recording time that contributes to our dataset. Adding sentence frames for our target words solely to provide context, in such a way that they cannot also be analyzed for cohort effects, would be a waste of statistical power.

We will use a memory probe task to ensure attention.

4.3.1.2 Avoidance of lexical predictability

Our design will avoid lexical predictability by constructing sentences with random noun-verb-noun combinations, such that the identity of each word does not constrain which words are likely to occur next in any way other than syntactic

category[6]. We do this because in natural sentences both lexical association and compositional lexical predictions are likely to change the probabilities of words that are competing for recognition, but not in ways that we can straightforwardly model. As we try to isolate the contributions of syntactic category, any unaccounted-for changes to the probability distribution of competitors will add noise to our measures and make it more difficult to identify category effects. Eventually, we hope that fully specified models of lexical prediction can be incorporated into analyses of more naturalistic stimuli.

4.3.1.3 De-correlation of stimulus variables

A major feature of the design for this study is the selection of target words such that constrained and unconstrained versions of our cohort variables are far less correlated than typically occurs.

To do this, we took all words that occur both in SUBTLEX (Brysbaert & New, 2009) with a part of speech tag (Brysbaert et al., 2012b) and in the CMU

---

[6] We note that it is not guaranteed that presenting stimuli that are not lexically predictable leads comprehenders to stop engaging in lexical prediction, and that different aspects of prediction might be susceptible to this manipulation in different ways. For example, compositional predictions for the object based on the subject and verb may no longer be possible when the subject and verb are a very unlikely combination, whereas word-to-word lexical association effects could continue. However, we believe the largest risk is simply that the noise we were seeking to avoid is in fact still present, which would make this design no worse off than if we did not try to eliminate lexical prediction. In **Chapter 5**, we discuss this issue extensively for a similar but visually presented design, including the possibility that lexical and syntactic prediction are interrelated in such a way that the syntactic prediction cannot occur without lexical prediction. If syntactic effects on the cohort are indeed modulated by semantic coherence or predictability, this would be unexpected but open up many interesting questions.

pronunciation dictionary (Weide, 1994). SUBTLEX provides overall lexical frequencies as well as frequency in each part of speech. The CMU pronunciation dictionary provides phoneme parses. This information would ordinarily constitute the lexicon which would be the basis for calculation of phoneme and word probabilities and therefore entropy or surprisal. Such calculations require only the set of words whose phoneme sequence is consistent with the input, and their frequencies. For this study, we created four alternative lexicons in which frequency and therefore probability values were altered to reflect the hypothesized effects of a syntactic constraint: one in which the context has led to an expectation for a noun and the constraint operates via facilitation, one in which the context has led to an expectation for a verb and the constraint operates via facilitation, one in which the context has led to an expectation for a noun and the constraint operates via inhibition, and one in which the context has led to an expectation for a verb and the constraint operates via inhibition.

To approximate probability distributions that have been affected by such constraints, we take the following steps. For the noun/facilitation lexicon, items that do not have a noun tag retain their normal lexical frequency. Items that do have a noun tag have their noun-specific frequency doubled, i.e., their noun-specific frequency is added to their normal lexical frequency. We follow the same procedure for the verb/facilitation lexicon. For the noun/inhibition lexicon, items that do not have a noun tag have their frequency set to zero. Items that do have a noun tag have their normal lexical frequency replaced with their noun-specific frequency. We follow the same procedure for the verb/inhibition lexicon. Any item with a zero

frequency in the relevant lexicon is removed from that lexicon. We then use these lexicons to compute entropy and surprisal at each phoneme. We compute unconstrained entropy and surprisal using the default frequencies from SUBTLEX, and compute entropy and surprisal under inhibition and under facilitation for nouns and for verbs, using the four different constrained lexicons. Our implementation of facilitation, in particular, is arbitrary. There is no prior data indicating the magnitude of probability increase that could be expected. This parameter will have to be fine-tuned in future work if the facilitation-constrained variables are supported in our dataset.

Having calculated these values, in order to select our stimulus set, we restricted the set of possible words to those with five or more phonemes. We did this so that we can analyze up through the fifth phoneme for all words and not have a loss of power at later phonemes in some words. We do not make any restrictions on the number of morphemes. Following Gaston and Marantz (2018), Brodbeck, Hong, and Simon (2018), and Experiment 2 in this chapter, we do not expect effects of entropy and surprisal at the first phoneme. Therefore, we do not attempt to reduce correlation between our constrained and unconstrained variables at the first phoneme. In **Table 6** below, we report the correlations between our constrained and unconstrained surprisal variables for nouns. We use this case for illustration purposes, but the situation is very similar for verbs and for entropy. The unconstrained and facilitation-constrained variables exhibit the most problematic degree of correlation. Our goal is to reduce correlation to 0.7 or below. 0.7 is an arbitrary, generally accepted standard, but it has

also been shown to perform reasonably well as a threshold for avoiding severe distortion due to collinearity (Dormann et al., 2013).

Table 6. *Correlations between unconstrained, facilitation-constrained, and inhibition-constrained surprisal variables for nouns at each phoneme position*

	Unconstrained & facilitation-constrained surprisal	Unconstrained & inhibition-constrained surprisal	Facilitation- & inhibition-constrained surprisal
Phoneme 2	0.951	0.666	0.746
Phoneme 3	0.977	0.773	0.833
Phoneme 4	0.990	0.828	0.874
Phoneme 5	0.992	0.841	0.885

This is a difficult problem and there is no obvious best solution. One perspective is that de-correlation (i.e., selecting subsets of words that show lower correlations) will lead to a strange, un-representative set of words. Because of this, any positive result from this study will ideally be paired with a similar finding in a randomly sampled stimulus set with the default correlation levels. The problem is that with the randomly sampled set alone we cannot with any certainty distinguish variables with correlation as high as 0.99.

Any approach to de-correlation will have flaws. Here, we describe three possibilities, still using as our example the correlation (for nouns, at the second phoneme) between unconstrained and facilitation-constrained surprisal because this is the most severe case. In **Figure 13**, we plot the two variables at this phoneme before any attempted de-correlation.

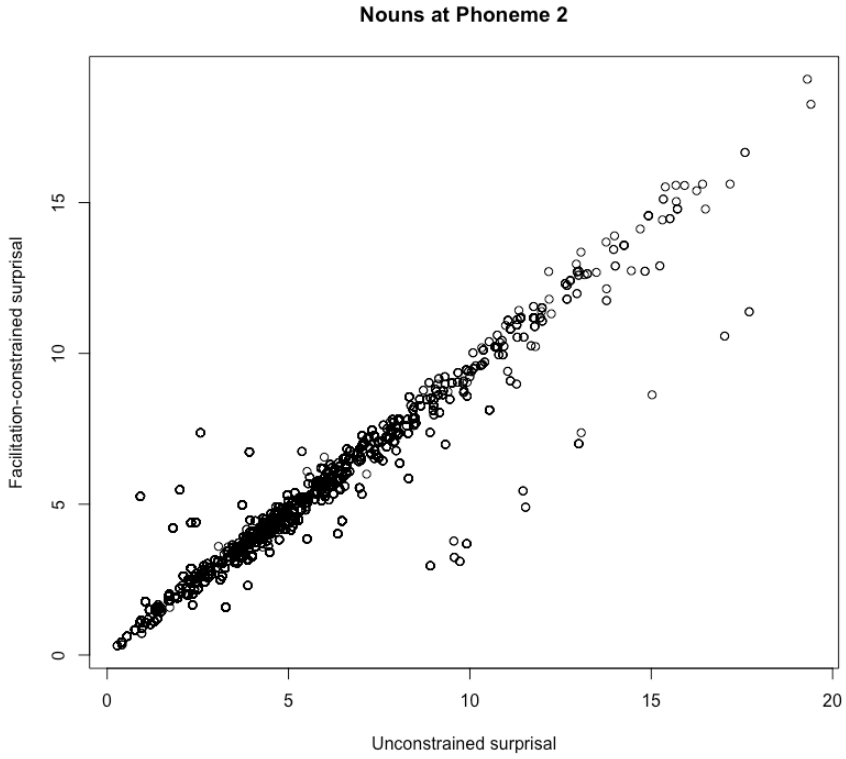


Figure 13. *Scatter plot for unconstrained and facilitation-constrained surprisal at phoneme 2, for nouns, before attempted de-correlation.*

One way to obtain a de-correlated set would be random sampling within a severely restricted range for both variables (in effect, "zooming in" on the correlation line). This is not feasible in our case both because there would not be enough target words within such a restricted range, and because this restricted range would make it more difficult to observe correlation with the neural data.

Another approach is to calculate the difference between the unconstrained and facilitation-constrained surprisal for each item, and restrict to items with some arbitrarily sufficient distance. For example, in the set of nouns, if we restrict to items with a difference greater than 0.8 between unconstrained and facilitation-constrained surprisal at the second phoneme, we reduce correlation to 0.710. By increasing the

cutoff to 1.0, we can further reduce the correlation to 0.570. In **Figure 14** below, we plot correlation for the set with the 1.0 cutoff.

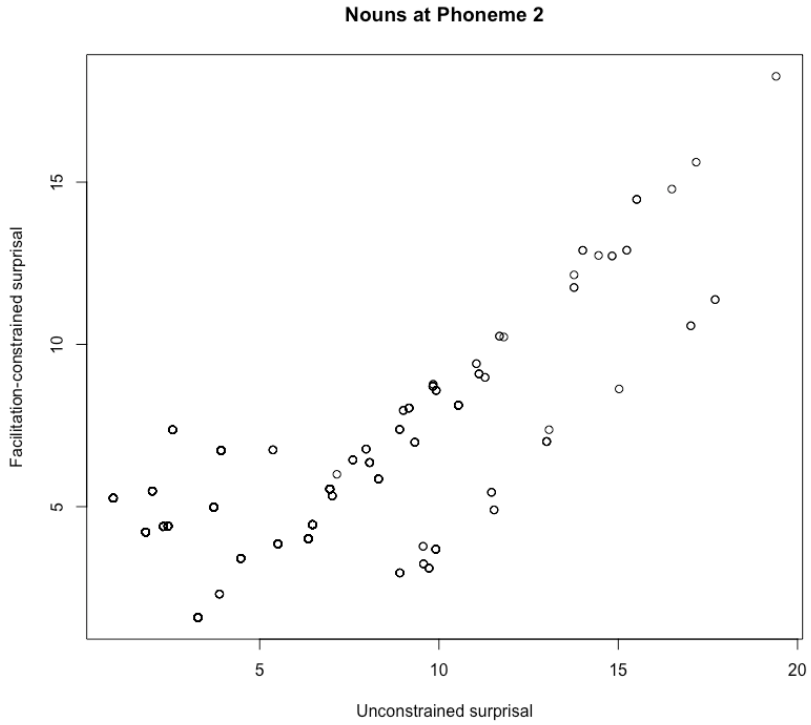


Figure 14. *Scatter plot for unconstrained and facilitation-constrained surprisal at phoneme 2, for nouns, after attempted de-correlation via a cut-off for how close the two values can be for any given item.*

This approach, however, produces too small a set of target candidates. The de-correlation step has to be applied iteratively for each phoneme and for both entropy and surprisal and still yield 2000 nouns or 1000 verbs, as we intend to present 1000 sentences (the maximum accommodated in a MEG session with reasonable length). It does have the advantage, though, of preserving the full range for each variable.

For our final attempted approach to de-correlation, we look for an arbitrary cutoff value for surprisal such that this value is to the right of the peak of the distribution of unconstrained surprisal values, and to the left of the peak of the

distribution of facilitation-constrained surprisal values. We then restrict to the set of words for which it is true both that their unconstrained surprisal is below the arbitrarily selected upper bound and that their facilitation-constrained surprisal is above the arbitrarily selected lower bound. We then reverse this (choosing a lower bound for unconstrained surprisal and an upper bound for facilitation-constrained surprisal) and additionally take the set of words for which these reversed constraints are true. This leaves us with a somewhat, but not severely restricted range for both variables, retaining most of the original outliers from the correlation line as well as one middle section. It also selects a large enough stimulus set that further iterations for subsequent phonemes and for entropy are still possible. With this process, for nouns at the second phoneme, we reduce the unconstrained/facilitation-constrained correlation from 0.951 to 0.599, the unconstrained/inhibition-constrained correlation from 0.666 to 0.170, and the facilitation-constrained/inhibition-constrained correlation from 0.746 to 0.551. In **Figure 15** below, we plot correlation for the set of words constrained in this way.

4.3.2 TRF analysis method

This dataset will be analyzed with the TRF method described for Experiment 2. Starting with a simple model including only acoustic variables, we will test the addition of unconstrained surprisal and entropy, facilitation-constrained surprisal and entropy, and inhibition-constrained surprisal and entropy, collapsing over phonemes 2-5. For any significant improvement, we will test whether separately modeling entropy and surprisal, the different phoneme positions, syntactic categories, or word positions improves model fit. If multiple versions of a specific variable improve

model fit (e.g., both unconstrained and facilitation-constrained surprisal), we can use spatiotemporal tests to ask whether the strength of the correlation with the observed data differs between the two variables.

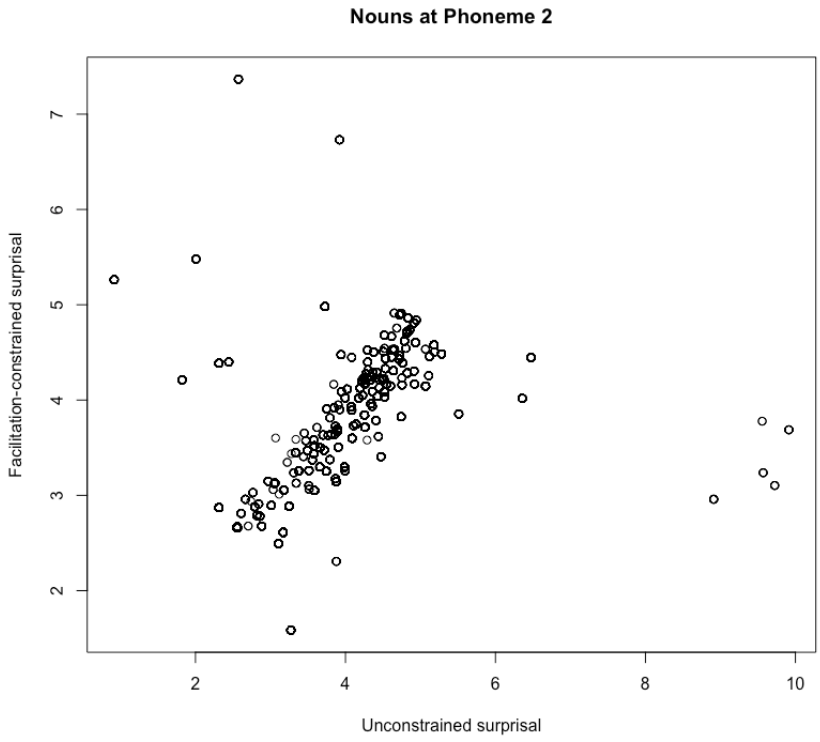


Figure 15. *Scatter plot for unconstrained and facilitation-constrained surprisal at phoneme 2, for nouns, after attempted de-correlation by restricting the ranges of both variables.*

4.3.3 Discussion

4.3.3.1 Contributions of Experiment 3

Our primary goal for the design that we will employ in Experiment 3 is that we gain more definitive evidence for the nature of the syntactic constraint on word recognition. Experiment 1 (in **Chapter 3**) was unable to distinguish a facilitatory

constraint and a lack of constraint. Our design for Experiment 3 is also intended to provide timing information. If, e.g., a constrained version of surprisal better predicts neural activity than an unconstrained version of surprisal, is this the case immediately and throughout the word, or is there a delay for the constraint?

Beyond the question of the syntactic constraint, Experiments 2 and 3 together should constitute considerable progress in refining the use of entropy and surprisal to probe word recognition and lexical access. In prior literature there has been variation in the manifestation of these effects across designs and analysis methods. With the addition of Experiment 3, we will have data for entropy and surprisal effects in continuous speech (Brodbeck, Hong, et al., 2018), single words (Experiment 2), and lexically unpredictable sentences (Experiment 3), all analyzed with the same method, which will allow more serious comparison between the stimulus types. Entropy effects are reported for continuous speech but not for single words, so the sentences in Experiment 3 are an important intermediate as we try to further our understanding of what aspects of the stimulus modulate the appearance of entropy effects.

4.3.3.2 Directions for future work

If Experiment 3 yields definitive evidence for one of the mechanisms for the syntactic constraint, we intend to move toward more complex hypotheses about both the constraint mechanism and the syntactic parses feeding the constraint. Evidence for the faciliatory constraint, for example, would invite fine-tuning of the parameters for the increase in probability. We would also need to consider how a constraint could be implemented when, for example, the context allows both adjectives and nouns.

One issue to keep in mind is to what extent the degree of correlation between different variants on entropy or surprisal will hold us back from pursuing finer-grained distinctions. This is very likely to be a problem, and we suspect that a combination of more deliberate manipulations (as in Experiment 3) and naturalistic stimuli will be necessary.

Chapter 5: Syntactic prediction in posterior temporal lobe[7]

5.1. Introduction

5.1.1 Overview

In the previous chapters, I have been largely concerned with the impact on auditory word recognition when the syntactic category of the word being recognized is known. In studying this issue, we ignore the full complexity of how it is that the syntactic category of an incoming word could be known or expected in advance, when that knowledge becomes available, and what exactly it consists of. We don't know if syntactic category expectations arise from hierarchical syntactic structure-building, and if so, via what parsing algorithm, or if they arise from simpler bigram co-occurrence probabilities. We don't know whether they are deterministic or not. We also don't know if syntactic predictions occur automatically for each upcoming word, or only in specific circumstances. We don't know to what extent they can be modulated by the reliability (at whatever level) of the input. Thus, when we assume in designing an experiment that comprehenders expect a noun after hearing "the," and then try to study the impact of expecting a noun, there are myriad ways in which we might be misrepresenting the process whose impact we are studying. A full and accurate accounting of the interaction between top-down and bottom-up information

[7] Macie McKitrick, Fen Ingram, and Aura Cruz Heredia assisted with some of the MEG data collection for Experiment 4 and its localizer. Macie McKitrick also assisted with creating stimuli for Experiment 4.

in word recognition will require a far more detailed understanding of the top-down contribution than is currently available. The work described in this chapter is an attempt in that direction, building on an extensive neuroimaging literature that investigates where (in the brain) and when (during sentence processing) effects of syntactic structure can be isolated in neural data.

Our goal was to identify neural effects that reflect the prediction of syntactic structure. A neural signature of syntactic prediction, regardless of modality, would be instrumental for investigating top-down effects of syntactic category. We used MEG to examine the response to a determiner phrase when readers expected that it would be the subject of a sentence ("the toasty tractors entered the scenic cathedrals") or the first item in a list ("the toasty tractors the scenic cathedrals"). A promising recent report of a syntactic prediction effect in MEG (Matchin et al., 2019) used natural sentences in which both syntactic and lexical prediction are possible. Using nonsense stimuli to discourage lexical prediction, we did not find evidence for a neural effect of syntactic prediction.

As we will discuss in the following sections, there are a variety of possible explanations for our null result, but one very clear new understanding we have gained is that "subtracting" lexical prediction from syntactic prediction is a far more complicated prospect than we had assumed, and that this issue may affect a large number of findings that are purportedly syntactic in nature. We speculate that the occurrence of syntactic prediction may actually depend on the occurrence of lexical prediction. The idea that lexical and syntactic processing could be more entwined

than we had previously imagined also opens up new and interesting questions surrounding top-down/bottom-up interactions during word recognition.

5.1.2 Background: Matchin et al. (2019)

Matchin et al. (2017, 2019) report fMRI and MEG data from a classic manipulation of structure (sentence/phrase/list) and content (natural/jabberwocky). **Figure 16** illustrates their design.

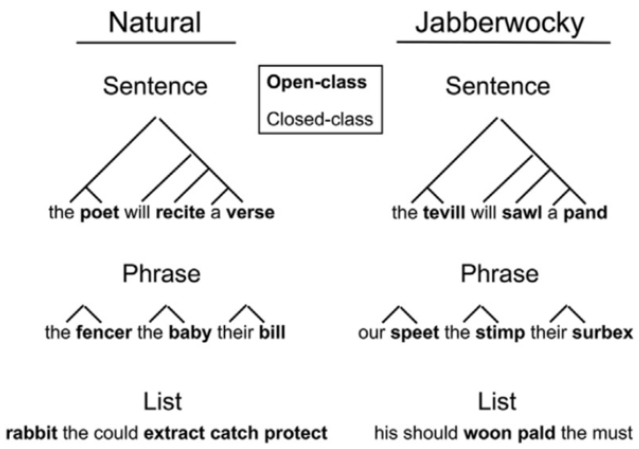


Fig. 1 – Schematic of stimulus design. Tree diagrams represent the constituent structure in each condition.

Figure 16. *Matchin et al. (2017)'s Figure 1, illustrating the stimuli for their structure by content manipulation. Reprinted from Cortex, Vol. 88, Matchin, Hammerly, and Lau, The role of the IFG and pSTS in syntactic prediction: Evidence from a parametric study of hierarchical structure in fMRI, pages 106-123, Copyright (2016) Elsevier Ltd, with permission from Elsevier.*

The motivation for Matchin et al. (2019) was that most paradigms comparing structured and unstructured stimuli (often, sentences and lists) in order to ask about combinatory processing use fMRI. These studies reliably report effects of structure in a language network (see **Figure 17**) including angular gyrus (AG, also referred to as

the temporo-parietal junction (TPJ)), anterior temporal lobe (ATL), posterior temporal lobe (PTL), and inferior frontal gyrus (IFG). However, they do not provide information about the timing of effects of structure. This makes it difficult to understand the specific contribution that each of these regions is making to online sentence processing, because the implication of an effect of structure varies according to how much bottom-up input has already been processed, and at what word positions in the sentence the effect occurs. Matchin et al. (2019) instead use MEG time course data, which provides novel evidence for the latencies of different effects of structure.

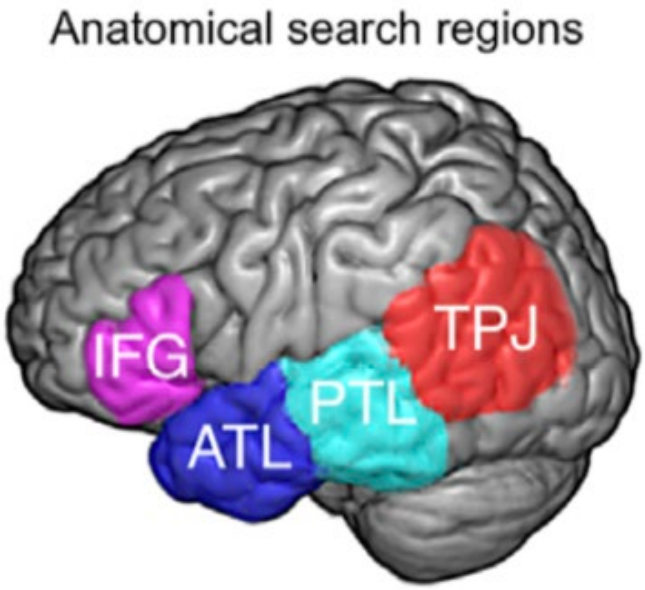


Figure 17. *Reproduction of Matchin et al. (2019)'s Figure 3, showing the anatomical search regions used in their study for typical language network areas showing structure effects: inferior frontal gyrus (IFG), anterior temporal lobe (ATL, posterior temporal lobe (PTL), and temporo-parietal junction (TPJ, also referred to as angular gyrus (AG)). Reprinted from Human Brain Mapping, Vol. 40, Issue 2, Matchin, Brodbeck, Hammerly, and Lau, The temporal dynamics of structure and content in sentence comprehension: Evidence from fMRI-constrained MEG, pages 663-678, Copyright (2018) Wiley Periodicals, Inc., with permission from John Wiley and Sons.*

When structure manipulations are crossed with content manipulations that compare natural and jabberwocky stimuli, it has often been suggested that combinatoric effects can be diagnosed as syntactic or semantic in nature (e.g., Pallier et al. (2011)). Matchin et al. (2017), using fMRI, replicated Pallier et al. (2011) in demonstrating an increased response for natural sentences relative to phrases and lists in inferior frontal gyrus (IFG), posterior temporal lobe (PTL), anterior temporal lobe (ATL), and angular gyrus (AG), and for jabberwocky sentences, in contrast, effects of structure in IFG and PTL only. This was taken by both groups to indicate that the role of AG and ATL is confined to semantic/conceptual or thematic processing. IFG and PTL are generally associated with lexical retrieval and storage as well as syntax, and because some degree of syntactic structure is assumed to be stored with lexical items, Matchin et al. propose that IFG and PTL are involved in lexical-syntactic processing.

Crucially, in this fMRI data, IFG and PTL did not show an increased response for phrases relative to lists, despite the fact that phrases involve more structure than lists. This suggests that the sentence > phrase response in those areas does not reflect basic syntactic structure building, but might instead reflect sentence-level syntactic prediction of the sort that for unconnected phrases would not be necessary or would not have to be maintained over any distance.

Matchin et al. (2019), using the MEG data for this same paradigm, report further evidence for a predictive structure-building role for posterior temporal lobe in syntactic processing. Experiment 4, reported in this chapter, is in direct response to

this finding. Therefore, we now describe Matchin et al. (2019)'s design and findings in greater detail.

5.1.2.1 Design

Matchin et al. (2019) fully cross structure (sentence, unconnected phrase, and list conditions) and content ("jabberwocky" and natural stimulus conditions), as in the fMRI version of the study. In the MEG data, however, they focus on the sentence vs. phrase comparison rather than the sentence vs. list comparison. They do this because open-class and closed-class words appear in the same positions in the sentence and phrase conditions. The contrast between open and closed-class words is known to elicit large electrophysiological differences, which might dominate any structure effects in the sentence vs. list comparison. This is a problem only for the MEG analysis because the MEG analysis examines each position in the sentence separately.

Sentences always had the same simple structure (Determiner-Noun-Modal-Verb-Determiner-Noun). Jabberwocky conditions substituted pronounceable nonwords for the content words, scrambled across sentences with function words fixed in place. Phrase trials consisted of a sequence of three Determiner-Noun or Modal-Verb phrases taken from the sentences. Repetition of lexical items was counter-balanced across subjects, so that the same word was not seen multiple times by the same person.

A critical element of the design was that stimuli were presented in blocks of the same condition, with a warning before each block telling participants what condition they were about to be presented with. The instructed task was to respond to a memory probe word, which was presented pseudo-randomly after 2 of every 6

trials. Participants had to indicate whether or not they had seen the probe word in the preceding sequence.

5.1.2.2 Results

Functional regions of interest within each of the four anatomical areas (inferior frontal gyrus (IFG), posterior temporal lobe (PTL), anterior temporal lobe (ATL), angular gyrus (AG)) for each participant were chosen from the fMRI data for use in the MEG analysis. Average time courses of activity within each ROI for each participant were then extracted for use in temporal cluster tests. Because the focus of this chapter is syntactic prediction, we will only discuss their reported MEG effects of structure in PTL and IFG.

Prediction effect

Matchin et al. (2019) find an effect of structure in PTL 272-484 ms after the onset of the second word (the subject noun). We reproduce their findings in **Figure 18**. PTL, in this study, is defined as "the superior temporal sulcus (STS) or MTG, posterior to primary auditory cortex and anterior to the end of the sylvian fissure." This finding seems to support the proposal that PTL supports syntactic prediction, because it occurs during the first phrase of the sentence. At this point, the bottom-up input in the sentence and phrase conditions is identical, and the only difference is that participants know what block type they are in, and therefore that sentence-level structure is upcoming in the sentence condition. This is consistent with what would be expected under a left-corner parsing model, in which input consisting of a subject determiner or noun phrase would lead to the projection of structure for a verb.

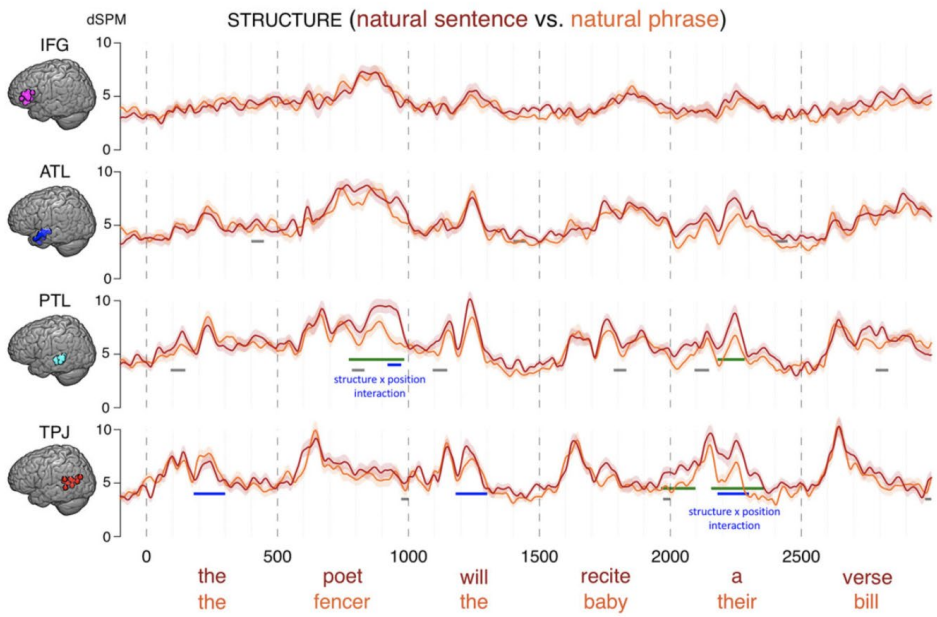


FIGURE 4 Analyses of STRUCTURE within each ROI (dSPM). Red: NATURAL SENTENCE, orange: NATURAL PHRASE. Gray lines indicate significant main effects of STRUCTURE in the word-level analyses, blue lines indicate significant interactions of STRUCTURE and POSITION in the word-level analyses, and green lines represent significant effects of STRUCTURE in the sentence-level analyses. X axis is time in milliseconds relative to onset of the first word in each six-word trial. Shading indicates the within-subject standard error (Loftus & Masson, 1994). The blue text "STRUCTURE × POSITION interaction" and the corresponding blue underline reflects significant time periods for the STRUCTURE × POSITION interaction in the word-level analysis. In the PTL, only the time period at word 2 survived a post-hoc pairwise comparison, while in the TPJ, the time periods at words 1, 3, and 5 all survived a post-hoc pairwise comparisons [Color figure can be viewed at wileyonlinelibrary.com]

Figure 18. *Matchin et al. (2019)'s Figure 4, showing the results of their structure manipulation. Reprinted from Human Brain Mapping, Vol. 40, Issue 2, Matchin, Brodbeck, Hammerly, and Lau, The temporal dynamics of structure and content in sentence comprehension: Evidence from fMRI-constrained MEG, pages 663-678, Copyright (2018) Wiley Periodicals, Inc., with permission from John Wiley and Sons.*

Non-predictive structure effects

Matchin et al. also observe a main effect of structure for all open-class words at 284-332 ms and all closed-class words at 92-148 ms. This is proposed to reflect increased costs of attention and maintenance for the syntactic structure associated with each lexical item, which would presumably increase in a sentence context where that syntactic structure requires integration. Matchin et al. suggest post-hoc that the latency difference between open and closed-class words is because closed-class words can be recognized more quickly.

In this dataset, PTL also shows content effects from the contrast between natural and jabberwocky sentences. This supports the idea that activity in this area reflects both lexical and syntactic processes. Interestingly, for open-class words the structure and content effects occur at roughly the same time. Content effects for open-class words should reflect the difference between real words and nonwords, as the latter cannot be successfully retrieved. In contrast, for closed-class words the content effect lags the structure effect by roughly 100 ms. These closed-class words are the same in jabberwocky and natural sentences, so this cannot reflect an issue in lexical retrieval. Instead, one difference between conditions is that in natural sentences these closed-class/function words are followed by real words, which can be predicted or constrained by selectional restrictions, while in jabberwocky sentences this type of prediction would not be possible.

Finally, Matchin et al. observe an effect of structure in PTL 180-284 ms after the fifth word (the second determiner), which also occurs in AG and is likely related to the processing of event semantics due to the verb in the sentence condition.

Lack of IFG effects

IFG, surprisingly, showed no effects of structure in the MEG data despite robust structure effects in the fMRI data. Matchin et al. offer several potential explanations. One is the possibility of reduced signal in frontal regions, but this is difficult to reconcile with the finding of content effects in IFG, unless the signal generated by content versus structure effects is qualitatively different. Another is the possibility that structure effects in IFG are occurring at different points in different trials, which would still sum to an effect in an fMRI analysis but would not in MEG,

where we analyze timepoint-by-timepoint. However, variation in effect latencies would be difficult to explain given that every sentence had an identical structure. Finally, they raise the possibility that the IFG effects observed in fMRI are sentence wrap up effects, which are not observed in MEG because the window of analysis ends 500 ms after the last word. In pursuit of this explanation, follow-up work would ideally include a later analysis window to look for wrap up effects.

Summary & questions

To summarize, Matchin et al. (2019) found that in a contrast between blocks whose trials consist of simple sentences and blocks whose trials consist of unconnected two-word phrases, there is an increased response to the first noun in the trial in sentence blocks relative to phrase blocks. This occurs in left posterior temporal lobe, 272-484 ms after the onset of the first noun. In the sentence blocks, this first noun is the subject of the sentence. The primary claim in this study is that the MEG data provides novel support for a critical role of PTL in syntactic structure-building, and in particular for predictive syntactic structure building. In following up on this claim, there are two uncertainties that we will address.

The first is that this design does not distinguish whether the projection of sentence-level structure is triggered by the head noun of the subject or by the first content word in the sentence. This is relevant in order for this data to help characterize top-down syntactic parsing. Second, it is possible that this result actually reflects lexical rather than syntactic prediction. In this design, presentation of the subject noun allows specific lexical items to be predicted, in addition to the syntactic structure necessary for an upcoming verb phrase to be integrated. Prediction of the

upcoming word is not possible in phrase blocks because there is no semantic relationship between the first and subsequent determiner phrases in a trial, whereas there necessarily is such a relationship between the subject and verb/object of a natural sentence. For example, “recite” and “verse” are far more related to and predictable from the subject “poet” than “baby” and “bill” are from “fencer.” We address both of these issues in Experiment 4.

5.1.3 Experiment 4

In the following sections, we present a study that builds on Matchin et al. (2019)’s evidence for predictive structure building in simple sentences. A straightforward interpretation of their effect is that comprehenders project sentence-level syntactic structure when they encounter the subject noun in the sentence blocks, whereas this structure is not necessary and therefore not generated in phrase blocks. This is just a single data point, but would seem to be evidence for a predictive structure-building function for posterior temporal lobe (PTL), in line with previous evidence for effects of syntactic structure specifically in posterior superior temporal sulcus (pSTS).

5.1.3.1 Primary questions addressed

Our follow-up study addresses several important questions raised by this finding, in pursuit of a better understanding of the role of PTL in syntactic prediction and the nature of the observed prediction effect. First, does the prediction effect reflect structural prediction or lexical prediction? Second, why was the prediction effect observed in response to the subject noun specifically? Matchin et al. (2019) used a complex manipulation of both structure (sentence, phrase, and list blocks) and

content (natural and Jabberwocky). Addressing our questions required only the sub-part of the design that is crucial for the prediction effect (natural sentences and phrases).

Our first priority was addressing the possibility that the prediction effect observed in Matchin et al. (2019) is actually a *lexical* rather than a *structural* prediction effect. Therefore, instead of using natural sentences in which the verb is at least somewhat predictable given the subject, we use sentences in which the subject determiner phrase and verb are randomly paired, so that they have the same (lack of) relationship as the first and second phrase in phrase trials. For example, for Matchin et al. (2019), since all sentences are natural sentences, participants reading "the poet" can expect that the upcoming lexical items will be semantically related to "poet." The set of possible verbs that are likely to follow is thus predictable, and indeed the sentence is completed as "the poet will recite a verse." The sentences in our experiment are instead along the lines of "the toasty tractors entered the scenic cathedrals." With sentences like this, reading "the toasty tractors" does not allow participants to expect any specific set of potential verbs or objects. This allows us to ensure that any apparent prediction effects we observe in response to the first determiner phrase are due to structural and not lexical prediction.

To discourage lexical prediction to the full extent possible while still using real words, we also use random pairing of adjectives and nouns within the determiner phrases (removing only the most severe violations), and use adjectives that are not in general predictive of any single noun. We also randomly pair the verbs and object determiner phrases. Assuming that neither lexical nor syntactic prediction are

occurring in the phrase condition for Matchin et al. (2019), our intention was that in our study this would again be true in the phrase condition, but in the sentence condition lexical prediction would not be possible. If lexical prediction had made no contribution or was only partially responsible for the prediction effect, that effect should still occur in our study. Of course, it is possible that lexical prediction still occurs automatically even in unpredictable sentences, and so our removal of lexical predictability does not guarantee that a prediction effect occurring in our design is purely syntactic.

Second, we address whether the prediction effect in Matchin et al. (2019) is in response to the subject noun per se, or is triggered by some other property of that stimulus item in the design (e.g., that it is the first content/open-class word, or second word in the sentence). All determiner phrases in Matchin et al. (2019) consisted of Determiner + Noun, but we add an intervening adjective. This allows us to ask whether the prediction effect moves from the second to the third word, indicating that it is indeed associated with the subject noun, or continues to occur in response to the second word, indicating one of the other possibilities mentioned above (first content/open-class word, or second word in the sentence).

5.1.3.2 Other changes from Matchin et al. (2019)

We made several other modifications to the Matchin et al. design.

For Matchin et al. (2019), the subject and object of the sentence were always determiner phrases, but the phrase trials were sometimes made up of verb phrases. We altered the set of stimuli such that (between subjects) we are comparing the response to the *same* phrases in a sentence trial (e.g., “the verbose wizards trampled

the opulent baths") or a phrase trial (e.g., "the verbose wizards the opulent baths"). This way, the block context is truly the only difference between conditions.

We also included a condition that was more distantly related to the syntactic prediction question. Following Lau and Liao (2018), we added coordination blocks in which trials consist of two coordinated determiner phrases (e.g., "the verbose wizards and the opulent baths"). This allows us to try to replicate the sustained effect of coordination that Lau and Liao observe during their second noun phrase, as well as to explore the difference between effects of sentence-level structure and structure for coordination (both predictive and otherwise).

For Matchin et al. (2019) analysis of MEG data for each specific participant could be spatially constrained to regions that had been activated (for that same participant) in fMRI data from the same paradigm. We did not collect fMRI data to complement the MEG data for Experiment 4. Instead, we attempted to narrow down our region of interest for syntactic structure effects by running a localizer task (in MEG) with the same participants, and in the same session, as Experiment 4. The localizer compared responses to natural sentences, scrambled natural sentences, and consonant strings. This was intended to restrict our region for analysis within PTL and potentially IFG.

Finally, we also collected EEG data for the same paradigm to allow for cross-method comparison, but the EEG analyses are not reported in this dissertation.

5.2 Materials & Methods

Our stimuli, informally pre-registered analysis plan, deviations from the analysis plan, and analysis scripts for both Experiment 4 and the localizer are available for viewing and download on OSF.

5.2.1 Participants

35 participants (16 women) completed this study (mean age: 21.2, range: 18-30). All participants were right-handed, native speakers of English, and 12 were also native speakers of additional languages. None reported history of neurological or linguistic impairment, brain injury, or hearing loss. All reported normal or corrected-to-normal vision. The procedure was approved by the University of Maryland Institutional Review Board and all participants provided written informed consent. Participants were compensated with their choice of $15 or 1 course credit per hour of participation. The full session (this experiment, the localizer, and Experiment 2 reported in **Chapter 4**) lasted 2 hours.

For both this experiment and the localizer, three of the 35 participants' datasets were excluded from analysis, because of (1) the participant expressing lack of attention due to claustrophobia, (2) the participant repeatedly falling asleep, and (3) extreme noise observed in the data. For this experiment, we made one additional exclusion because of excessive alpha activity.

5.2.2 Stimuli

All stimuli are reproduced in **Appendix A**. Our main objective in creating stimuli was to ensure that prediction of upcoming lexical items from the first

determiner phrase was equally not possible in all conditions, such that any apparent prediction effects could not be due to prediction of specific lexical items, and could more likely be attributed to syntactic structural prediction.

Our stimuli were sequences of words in which lexical predictability between words was minimized and our conditions varied in syntactic structure: sentences ("the toasty tractors entered the scenic cathedrals") or lists of phrases ("the violent novels [blank] the crinkly cheeses"). Trials were presented one word at a time, in blocks in which the participant knew they were only being presented with sentences or only being presented with phrases. We used the same structure for all stimuli. Sentences had the following pattern: the *adjective noun verb* the *adjective noun*. Phrase list trials consisted of: the *adjective noun* [blank] the *adjective noun*.

As mentioned above, we included a third, parallel condition of coordinated phrases ("the fleshy soldiers and the pale bikes"). Coordinated phrase trials always consisted of: the *adjective noun* and the *adjective noun*. As with the sentence and phrase conditions, participants were instructed at the beginning of the block that in that block they were only being presented with coordinated phrases.

In total, each participant was presented with 70 items from each of the three conditions, but only 66 from each condition were analyzed.

Nouns were always plural and verbs were always in the past tense. The set of nouns and verbs came from the nouns and verbs used by Lau and Liao (2018) and Matchin et al. (2017, 2019). Adjectives were selected from a list of all adjectives appearing in the Corpus of Contemporary American English (Davies, M., 2008) for which there are no single nouns with greater than 15% probability of following that

adjective. We did this to ensure that our adjectives did not lead to strong expectation of any specific noun.

We removed all low frequency adjectives, prefixed adjectives, adjectival participles, nouns used as adjectives, hyphenated adjectives, and adjectives that would need to be capitalized. We then randomly paired our adjectives and nouns, and removed any resulting phrases that were especially emotionally evocative or that violated selectional restrictions between the adjective and noun. This is likely to have increased predictability within the determiner phrase (from adjective to noun), but we considered this unproblematic because we were comparing the response to the same determiner phrases between conditions.

We then assembled random pairings of our determiner phrases to create three sets of 70 determiner phrase pairs. Any given participant would see one set of these determiner phrase pairs in the sentence condition (with a verb in between each of the two determiner phrases), the second set in the coordination condition (with "and" in between each of the two determiner phrases), and the third set in the phrase condition (with just a blank screen in between each of the two determiner phrases).

For the sentence condition, we randomly chose 70 verbs to insert between the determiner phrases. We removed any resulting combinations in which the combination of the verb and second determiner phrase resulted in a selectional restriction. We crucially did not alter the random pairings between the first determiner phrase and the verb except for a handful of highly emotionally evocative cases. This meant that upon encountering the first determiner phrase in the trial in any condition, upcoming lexical material, whether verb or second determiner phrase, was

not any more predictable in the sentence condition than the coordination or phrase conditions.

Impressionistically, sentences constructed from truly random word and phrase pairings made processing extremely difficult, and we suspected that it might lead participants to stop processing meaning altogether. Therefore, in cases where we used non-random combinations, our goal was to make the sentences as easy as possible to parse while still maintaining low lexical predictability.

Because this design was counter-balanced, the condition in which each set of 70 determiner phrases appeared was rotated across participants. Therefore, the same set of 70 verbs appeared with each of the three different sets of determiner phrases, and we applied the above criteria in each case.

There was no repetition of any single word across the set of stimuli presented to a single participant. Across participants, the same pair of determiner phrases occurred in the sentence condition, the coordination condition, and the phrase condition, so that we could compare responses to identical determiner phrase stimuli when expectation for the syntactic structure of what was upcoming was the only thing that varied.

We used three basic lists, where each list contained all three conditions, and each set was in a different condition in each list. With three conditions, presented in blocks, each list had 6 possible block orders. We therefore used 18 different stimulus lists. However, within each list, instead of presenting 70 trials from the same condition in a single block, we broke each condition into two blocks of 35 trials, and repeated the fundamental block order twice (i.e., if the fundamental order for List 1a

was *sentence*, *phrase*, *coordination*, we presented six 35-trial blocks in the following order: *sentence*, *phrase*, *coordination, sentence*, *phrase*, *coordination).* Though participants did not know this, the first two trials in each block of 35 were triggered as practice items and not analyzed.

10 additional trials from each condition (no overlap with experimental items) were used for a short, three-block practice run of the experiment completed before the MEG recording, outside of the magnetically shielded room.

5.2.3 Task

We included single-word memory probes at the end of each block to encourage attention. Participants were instructed that they would be reading sequences of words, in six short blocks with breaks in between. They were warned that at the end of each block, there would be test trials in which they would be presented with single probe words and asked whether they had seen the word or not. They were given two button boxes and instructed to press the button in their left hand for "No" and the button in their right hand for "Yes." Before each block, a prompt appeared indicating whether that block would present sentences, pairs of items, or lists of items. Each block contained 35 trials, with an option to rest after the 10^{th}, 20^{th}, and 30^{th} trials. Then, 12 probe words were presented, one at a time, with a question mark. Half of the probes had appeared in the block and were supposed to elicit a "yes" answer, and half had not appeared in that block or previously in the experiment. "Yes" probes were chosen semi-randomly from words that had appeared in the block, with the restrictions that they were split evenly in coming from the first determiner phrase or the second determiner phrase, split evenly in being nouns or adjectives, and

were roughly evenly distributed in where in the block they had occurred. “No” probes were chosen from the list of candidate nouns and adjectives originally used to create the stimuli, and had the same noun/adjective split. Participants were not told about any of these restrictions on the probes.

“Yes” and “No” probes were presented in a random order. There was no time limit on the response to the probes, but participants were instructed to answer according to their first reaction. After they pressed either button with their answer, the next probe word appeared. Breaks within and between blocks were also un-timed, and participants could press a button to continue whenever they were ready. The entire task lasted roughly 25 minutes, depending on how much rest time the participant chose to take.

This task was very similar to the task used by Matchin et al. (2019), except that in that study probes were distributed randomly during the block, rather than all being presented at the end of the block. We made this change to discourage participants from trying to explicitly remember each specific stimulus in preparation for an upcoming probe, which could lead to phonological rehearsal.

5.2.4 Procedure

Before beginning the experiment, participants completed a short practice run of three blocks. This mimicked the real experiment in structure but had only 10 trials and three probes per block, and answers to the memory probes were provided by keyboard rather than button box.

Once set up for recording inside the magnetically shielded room, participants lay supine and viewed the stimuli projected on a screen above their heads, holding

one button box in each hand. Stimuli were presented in white font on a black background. The study began with an instruction slide, and participants could press a button to begin when they were ready. At the beginning of each of the six blocks, there was a prompt screen indicating which type of block it would be and reiterating the instructions. Then, participants could press either button when they were ready to begin. Within each block, there were 35 trials. After the 10th, 20th, and 30th trials, a REST! prompt appeared, and then instructions to press either button to move on when ready. After the third block, participants were alerted that they were halfway done with the task.

Each trial began with a 1500 ms fixation cross, a 300 ms blank, and then each of the seven words appeared on the screen, one at a time, for 300 ms, with a 200 ms blank between words. A 300 ms blank followed the final word, before the fixation cross for the next trial began. On phrase trials, the fourth word position was instead a blank screen. This trial timing was identical to the localizer except that there were seven rather than nine words per trial.

After the 35th trial, 12 probes were presented. Each presentation of a probe began with a 1500 ms fixation cross, a 300 ms blank, and then the word "TEST" appeared for 300 ms. After a 300 ms blank, the probe word appeared on screen with a question mark, and disappeared only when a button was pressed to answer.

5.2.5 MEG data collection & preprocessing

All MEG data collection and preprocessing details were the same as described in **Chapter 4**, except for the following details.

We computed and selected ICA components for rejection using a high-pass filter of 1 Hz and a low-pass filter of 40 Hz. For ICA component selection, we visualized the entire 1500 ms epoch (for the localizer) or the 3500 ms epoch (for Experiment 4). We then proceeded with the 40 Hz low-pass filter, down-sampling the data to 500 Hz. To select epochs for rejection, we applied an absolute threshold of 2 pT, and removed any additional extraneous artifacts identified by visual inspection. Baseline correction is specified for each specific analysis in the upcoming sections. For source estimation, we used a depth weighting parameter of 0.8 and dSPM noise normalization with a signal to noise ratio of 3.

5.2.6 Sentence localizer

We included a localizer so that we could generate functional regions of interest in addition to the region of interest indicated by Matchin et al. (2019). Following a very common localizer approach for sentence processing studies (e.g. Fedorenko et al. (2010) and others) we compared natural sentences, scrambled sentences, and lists of nonwords. The sentence/scrambled contrast was intended to isolate syntactic structure-building effects, while comparison with the nonword lists would elicit effects reflecting both syntactic and lexical processes. We intended to use any clusters arising in IFG or PTL for the sentence/scrambled contrast as regions of interest in Experiment 4, and if no such clusters arose, we intended to use clusters from the sentence/consonant contrast. The sentence/consonant contrast would otherwise be used purely for exploratory analysis. While our analyses identify candidate regions in both time and space, we intended to use only the spatial information in constructing group-level ROIs for Experiment 4. The timing

information, however, is helpful in interpreting the nature of the effects. Including both a syntactic and a lexical contrast also unexpectedly allowed us to make some interesting (speculative) inferences about the nature of Matchin et al. (2019)'s findings.

To maximize statistical power with a short recording time, we collapsed over all word positions in the sentence. This means that any observable structure effects are those that occur in response to each or most words of the sentence at roughly the same time. A different approach would be needed to look for structure effects specific to any specific position in the sentence.

We conducted spatiotemporal cluster tests on the localizer data in the left temporal lobe and left IFG. This analysis was conducted before any data from Experiment 4 had been visualized or analyzed in any way, so as not to bias our decisions. Within the left temporal lobe, we were looking for clusters in PTL, but we allowed a larger search space because MEG does not have extremely high spatial precision and we did not want to cut off a cluster extending outside of PTL.

We note that because we used a fixed orientation of the dipole for source localization and threshold-based cluster tests for our analysis, sources within any given cluster will have the same polarity. Thus, we avoid the issue of activity cancellation that can occur in ROIs that include sources of different polarity. See Gwilliams and Marantz (2016) for further discussion of this problem and the merits of functional ROIs for fixed orientation analyses.

5.2.6.1 Stimuli

The stimuli were borrowed from the localizer used by Lau and Namyst (2019). The localizer consisted of three conditions: natural sentences, lists of real words (scrambled sentences), and lists of nonwords. Each sentence or list was a sequence of nine words or nonwords. The sentences were themselves adapted from Rogalsky and Hickok's (2009) localizer, with some additions. The scrambled word lists were created by scrambling the set of words across all sentences and recombining in sets of nine. We used two different lists for presenting the stimuli, such that content words that appeared in a natural sentence trial in List A appeared in a scrambled sentence trial in List B, and so a single participant would not see the same content word in more than one condition. The nonwords were unpronounceable consonant strings that were matched with the real words for length.

5.2.6.2 Task

The localizer always occurred between Experiment 4 and Experiment 2, both of which involved very demanding tasks. Participants were instructed that this was a brief interlude and they should simply read what appeared on the screen.

5.2.6.3 Procedure

Stimuli were presented in white text at the center of a black screen, one word at a time. Each trial consisted of a 1500 ms fixation cross, a 300 ms blank screen, and then each word was presented for 300 ms, with a 200 ms blank screen between words. The final word was followed by a 300 ms blank screen, before the fixation cross for the next trial appeared.

Each participant was presented with 60 trials (20 from each condition) in a random order. Different from Experiment 4, but consistent with most of the prior fMRI literature using this kind of localizer design, there was no pre-trial cue indicating what the nature of the stimulus would be for that trial. The entire task lasted 6.5 minutes.

5.2.6.4 Analysis

The epoch for analysis for the localizer was -100 to 500 ms from the onset of each word/nonword.

We did not analyze the first word in each trial because it could not be evident to participants whether they were reading a natural sentence or a scrambled sentence until the second word at the earliest. Therefore, condition averages for each participant included all words in all trials, except for the first word in each trial, with a maximum of 160 data points contributing to each condition average. When trials were missing due to artifact rejection, we equalized the number of trials included from each condition.

We did not apply baseline correction because, given the short 500 ms SOA, the typical -100 to 0 ms baseline period would have been the final 100 ms of the previous epoch, when condition-specific differences are still likely.

We conducted spatiotemporal cluster tests (see Nichols and Holmes (2002)) for our two contrasts of interest: sentence vs. scrambled and sentence vs. consonant. These tests were conducted over the entire 500 ms epoch but within a constrained search space of brain regions defined in the Desikan-Killiany cortical atlas in Freesurfer. From the left temporal lobe, we included: the superior, middle, inferior,

and transverse temporal gyri, banks of the superior temporal sulcus, fusiform gyrus, temporal pole. From the left frontal lobe, we included: pars opercularis, pars triangularis, pars orbitalis.

For each contrast, a two-tailed repeated-measures *t*-test was conducted at each source and time point within the search space. For each cluster of adjacent *t*-values exceeding a threshold equivalent to $p < 0.05$ (with a minimum of 10 contiguous sources and 25 ms duration), we summed the *t*-values in the cluster. We then compared this sum to a distribution of the largest such cluster sum values generated in each of 10,000 random permutations of the data (shuffling condition labels within each subject). The *p*-value for each of the clusters found in the real dataset was therefore the proportion of the 10,000 random permutations on which the largest cluster sum value was larger than that of the currently observed cluster.

We expected our sentence/scrambled contrast to yield effects related to the presence of syntactic structure, and we were interested in any such effects localizing to inferior frontal gyrus or the posterior (superior or middle) temporal lobe. For this localizer analysis we report any clusters with $p < .2$, because the goal was to find regions of interest for the analysis of Experiment 4, rather than confirmatory hypothesis testing that would warrant stricter Type I error control.

5.2.6.5 Results

For spatiotemporal cluster test results for both the localizer and Experiment 4, we plot both the spatial extent of the cluster and the time course of activity within the cluster. For the spatial plots, the color at each source point reflects the maximum *t*-value within the time window of the cluster; not all source points are necessarily

significant at every time point. The time course plot shows, at each time point, the averaged neural activity for all sources that are part of the cluster.

Sentence vs. Scrambled

Our spatiotemporal cluster test yielded no clusters for the sentence/scrambled contrast in the left frontal lobe region that we analyzed (pars opercularis, pars triangularis, pars orbitalis). In the left temporal lobe, we found three clusters, which we report in **Table 7.**

Table 7. *Spatiotemporal clusters with $p < .2$ for sentence vs. scrambled contrast.*

Cluster	Time	Location	*p*-value
1	116-242 ms	anterior inferior temporal gyrus	$p = .005$
2	124-232 ms	anterior superior temporal sulcus	$p = .125$
3	212-312 ms	posterior superior temporal sulcus (dorsal bank)	$p = .172$

Clusters 1 and 2, along anterior inferior temporal gyrus and anterior superior temporal sulcus, display time courses of activity almost exactly inverse of each other, with a peak at roughly 130 ms (negative-going for the inferior temporal cluster, positive-going for the superior temporal cluster) displaying a scrambled > sentence effect, followed by a polarity reversal and a more sustained sentence > scrambled effect. Given their almost identical time windows of significance, we strongly suspect that these two clusters reflect the same effect measured on either side of the middle temporal gyrus. The latency (shortly after 100 ms) and directionality (scrambled > sentence) of the first peak in the significance window may reflect the (lack of) predictability of the visual wordform in the scrambled condition. Another possibility, given the early time window, is that this is a "wrap-around" effect owing to the short

SOA, and actually reflects processes occurring ~500-600 ms after the onset of the previous word. In either case, these clusters are not relevant as regions of interest for Experiment 4 because they are in anterior rather than posterior temporal lobe.

Cluster 3 (see **Figure 19**), on the dorsal bank of posterior superior temporal sulcus, displays a clear negative-going peak at 250 ms with a stronger response for sentence > scrambled, and then a positive-going peak at 400 ms displaying no apparent difference between the conditions. This time window and location are very consistent with the effects of syntactic structure found by Matchin et al. (2019), making this cluster suitable as a region of interest for the analysis of Experiment 4.

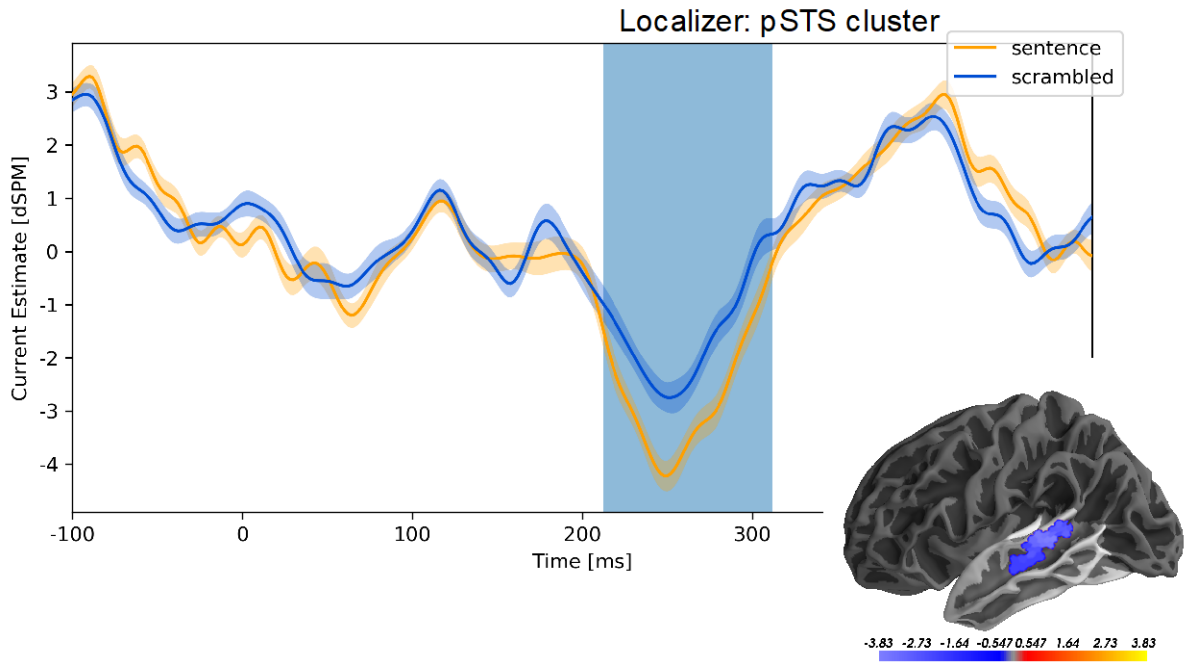


Figure 19. *Sentence vs. scrambled cluster in pSTS for localizer, plotting time course and location of neural activity. Color bar shows maximum t-value at a source point.*

Sentence vs. Consonant strings

The sentence/scrambled contrast yielded a PTL cluster suitable as a region of interest for the analysis of Experiment 4. However, we still engaged in exploratory analysis of the sentence/consonant contrast because it is a frequently employed contrast in the fMRI literature, and time course information is of potential interest. Because this was a secondary contrast for us and it yielded many more clusters, we used a stricter significance threshold of $p < .05$ for serious consideration.

We did not find any clusters in the frontal region of interest that met this threshold. Clusters with $p < .05$ in the temporal region of interest are reported in **Table 8**. Cluster 4, which appears strongest along posterior inferior temporal gyrus, is outside of our primary area of consideration for lexical and/or syntactic effects and probably reflects differences in conceptual access; therefore, we do not discuss it further here.

Table 8. *Spatiotemporal clusters with p < .05 for sentence vs. consonant contrast.*

Cluster	Time	Location	*p*-value
4	188-384 ms	posterior inferior temporal gyrus	$p = .0404$
5	206-334 ms	superior temporal sulcus (dorsal bank and posterior ventral bank)	$p = .0338$
6	302-486 ms	anterior superior temporal gyrus	$p = .0275$
7	326-474 ms	transverse temporal sulcus	$p = .0331$

On visual inspection, we see that Clusters 5-7 (see **Figure 20** and **Figure 21**) are all in the area of superior temporal gyrus/sulcus and transverse temporal gyrus/sulcus, and all have a time course of activity extremely similar to the PTL cluster from the sentence/scrambled contrast (Cluster 3), though sometimes flipped in polarity. This time course is characterized by peaks in activity at 250 and then (in the

opposite direction) at 400 ms. Cluster 3 showed a sentence > scrambled effect in a window containing the 250 ms peak, and these superior temporal clusters (Clusters 5-7) elicited by the sentence/consonant contrast similarly appear to show an increase for the sentence condition at this peak. Cluster 5 (see **Figure 20**) is significant in a nearly identical time window to Cluster 3, but with a somewhat more diffuse localization, extending more anteriorly along the dorsal bank of STS, and, at the posterior end of the cluster, extending to the ventral bank of STS.

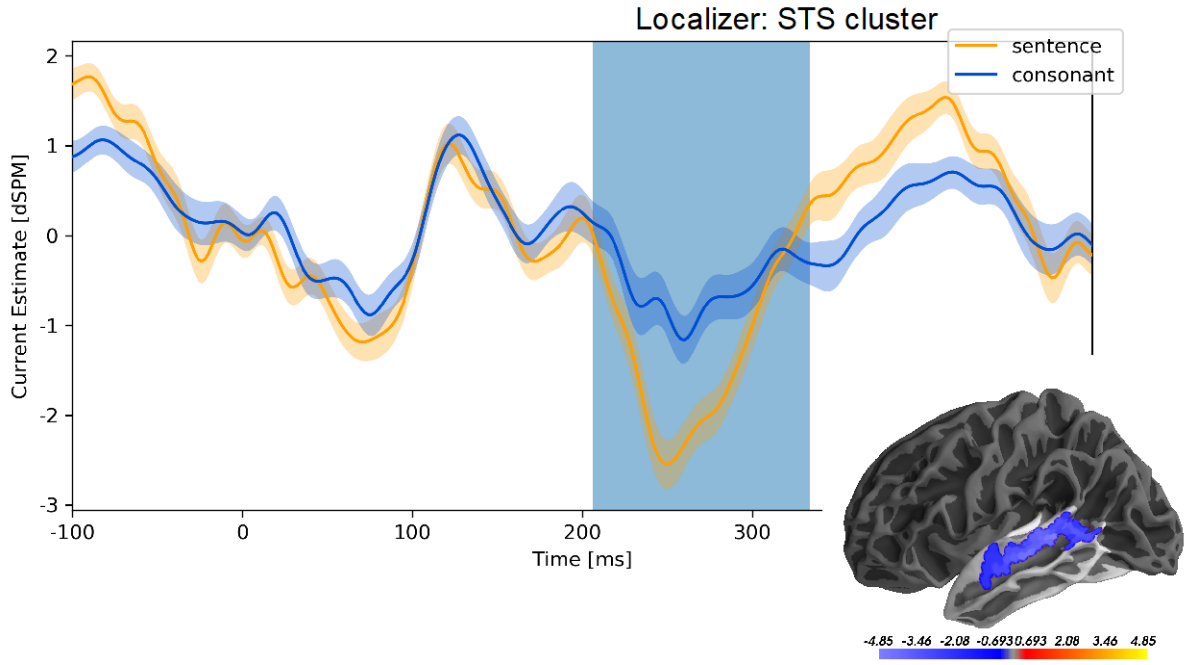


Figure 20. *Sentence vs. consonant cluster along STS for localizer, plotting time course and location of neural activity. Color bar shows maximum t-value at a source point.*

This set of clusters also shows a decreased response to the consonant condition relative to the sentence condition at the 400 ms peak; this did not occur for the scrambled condition in Cluster 3. Clusters 6 (along STG) and 7 (in TTS) are

significant in the window of the 400 ms peak, showing increased activity for sentences relative to consonant strings (see **Figure 21**).

5.2.6.6 Discussion

Our sentence/scrambled contrast yielded an effect of structure (sentence > scrambled) along the dorsal bank of the posterior superior temporal sulcus, in a 100 ms window roughly centered around a negative-going peak at 250 ms. This cluster will serve as a functional ROI for structure effects in Experiment 4.

Since the analysis collapses over all word positions, we expect that any effects we are able to detect are not specific to any single position in the sentence. Matchin et al. (2019) found effects of structure in posterior temporal lobe from 92-148 ms in closed-class words and 284-332 ms in open-class words. Our analysis does not distinguish between open and closed-class words, and so it is possible that our averaged time course is reflecting a mix of exactly this earlier effect of structure for closed-class words and later effect of structure for open-class words.

Our sentence vs. consonant string contrast yielded the same STS effect at 250 ms, but also a later difference along STG and TTS at 400 ms. Though this later effect is elicited only by the contrast between lexical (sentence) and non-lexical (consonant string) stimuli, we cannot be sure that it is not also driven in some way by the difference in structure between sentences and consonant strings.

The biphasic response also suggests the possibility that Matchin et al. (2019)'s syntactic prediction effect could have been driven by a combination of distinct early and late responses. Matchin et al.'s effect, on the second word, extended across these early and late windows (272-484 ms). Crucially, Matchin et al.'s source modeling

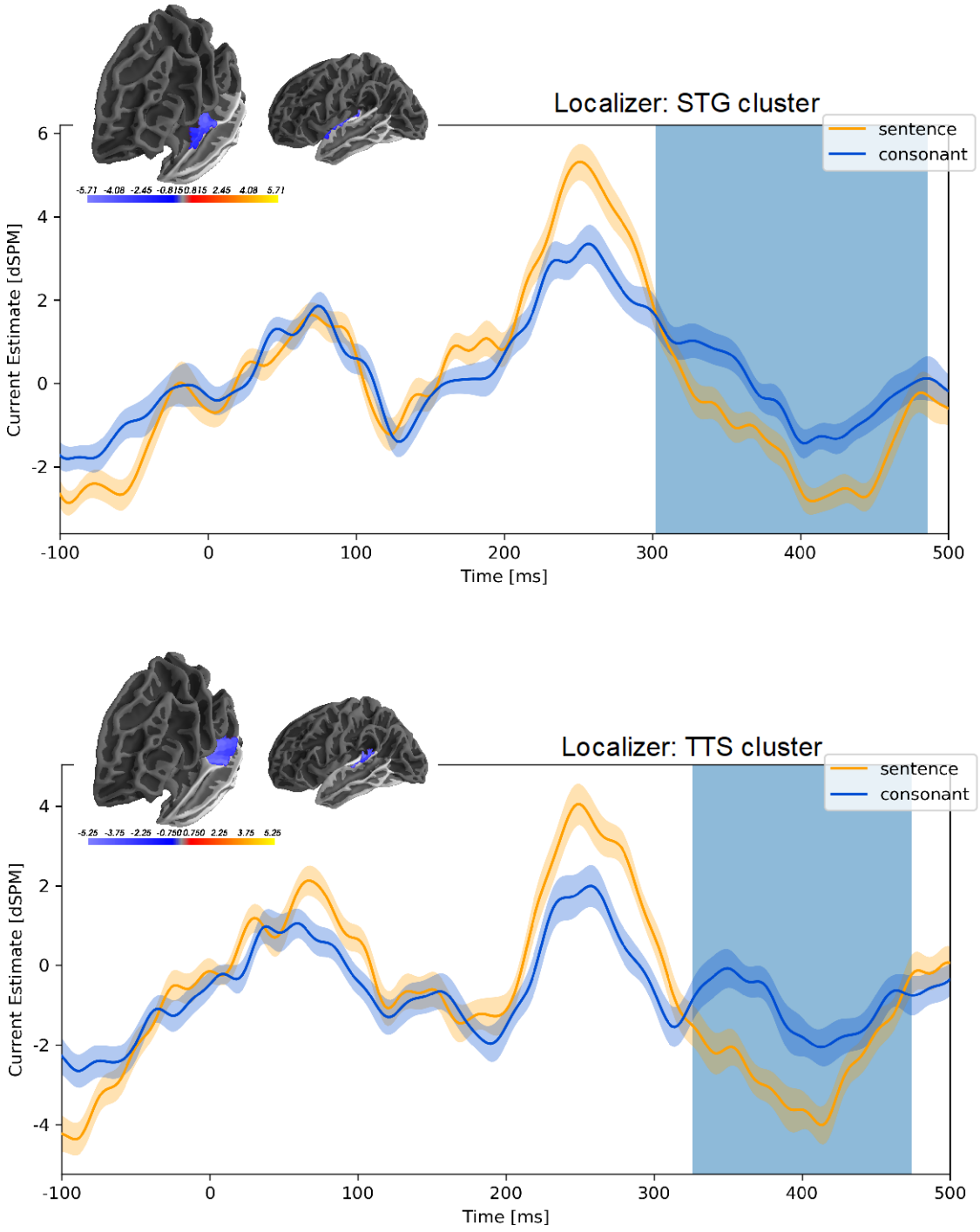


Figure 21. *Sentence vs. consonant clusters along STG (top) and TTS (bottom) for localizer, plotting time course and location of neural activity. Color bar shows maximum t-value at a source point.*

approach assumed loose orientation of the dipoles, such that activity had to manifest as positive, whereas our analysis of the localizer used fixed orientation, distinguishing positive and negative current estimates. Therefore, we speculate that the sustained prediction effect observed by Matchin et al. (2019) could actually have been driven by the two peaks that we observe with opposite polarities at 250 and 400 ms along STG and STS, which would, in a loose orientation analysis, manifest as a single, sustained, positive effect over that time window. Because of this speculative relationship to the prediction effect observed by Matchin et al. (2019), we use the STG and TTS clusters as functional ROIs in Experiment 4, in addition to the pSTS ROI from the sentence/scrambled contrast.

5.2.7 Data analysis

5.2.7.1 Behavioral data

Participants completed a memory probe task at the end of each block of stimuli. In planning the study, we decided we would not use accuracy on this task as a criterion for inclusion in the analysis of the neural data, because we expected participants to find it quite difficult and we expected that accuracy could be close to chance even for those who were paying attention. Therefore, we report mean accuracy in each condition for full transparency, but this information was not used or considered any further.

5.2.7.2 Neural data

For the purposes of data processing (ICA visualization and epoch rejection) we used a single epoch encompassing the entire trial (3500 ms). However, our primary epoch for analysis was 0-1500 ms from the onset of the trial, which

encompassed the presentation of the first determiner phrase (the *adjective noun*). For tests in this first analysis window, we used the 100 ms window prior to the onset of the first word for baseline correction.

Our secondary epoch for analysis was the 1500-3500 ms window beginning with the onset of the intervening item between the two determiner phrases. Depending on the condition, this intervening item was a verb, the coordinator "and", or a blank screen. These 2000 ms encompassed the intervening item as well as the second determiner phrase. For tests in this second analysis window, we used the 100 ms prior to the onset of the intervening item for baseline correction, as all conditions were indistinguishable up to this point.

All tests reported in this section are two-tailed, despite cases in which a one-tailed test would have been justified by the preceding literature, due to the fact that we used a fixed orientation of the dipole current in source localization and would therefore be analyzing both positive and negative neural activity values. We used three different testing approaches to investigate effects of syntactic prediction in the first epoch.

Matchin et al. (2019) ROI

First, in order to ask whether we could replicate (in a more narrow sense) the apparent effect of syntactic prediction found by Matchin et al. (2019), we created a region of interest based on the PTL ROI they had used. Because Matchin et al. had MRI data for each participant, in each anatomical region (IFG, AG, ATL, PTL) they extracted peak coordinates for the averaged structure and content contrasts and constructed each participant's MEG ROI so as to surround that peak. For our ROI, we

took the average (along each dimension) of each their individual participant peak fMRI coordinates for the posterior temporal lobe (PTL), which were along posterior superior temporal sulcus (pSTS) or middle temporal gyrus (pMTG). We then created a region of interest with radius 25 mm around the averaged peak coordinates. Using this spatial ROI (see **Figure 22**), we analyzed two temporal windows of interest as determined by Matchin et al. (2019)'s results. Their syntactic prediction effect was 272-484 ms after presentation of the subject noun, but it is unclear in their design whether the structure effect is in response to the subject noun or the first content word. Therefore, within the spatial ROI we computed a separate average of source-localized activity for each of our participants (for each condition) over the time window 272-484 ms after the first adjective and the time window 272-484 ms after the first noun. For each time window, we ran a repeated measures *t*-test on the participant averages, comparing the sentence and phrase list conditions.

Functional ROIs from localizer

Our second analysis approach was an attempted replication in a broader sense, intended to detect predictive structure effects even if the specific location and timing of Matchin et al. (2019)'s effect was not representative. For this analysis, rather than using an ROI derived from Matchin et al. (2019)'s fMRI data, we used an ROI determined by the results of our localizer task. The localizer yielded a cluster along pSTS for the contrast between sentences and scrambled sentences, from 212-312 ms after word onset (see **Figure 19**). Note that this was a group level cluster, and we did not have an individual ROI for each participant. The spatial coordinates of this cluster were used as the ROI for a temporal cluster test within the time window of the first

determiner phrase (0-1500 ms), again comparing the sentence and phrase list conditions. We did not restrict our test to the time window in which the cluster was significant in the localizer because the tasks and stimuli were quite different and we did not expect exact correspondence between the nature of structure effects in the localizer and Experiment 4. This was also a natural way to deal with the problem that on the basis of Matchin et al. (2019) we could not predict whether the syntactic prediction effect might manifest on the adjective or the noun.

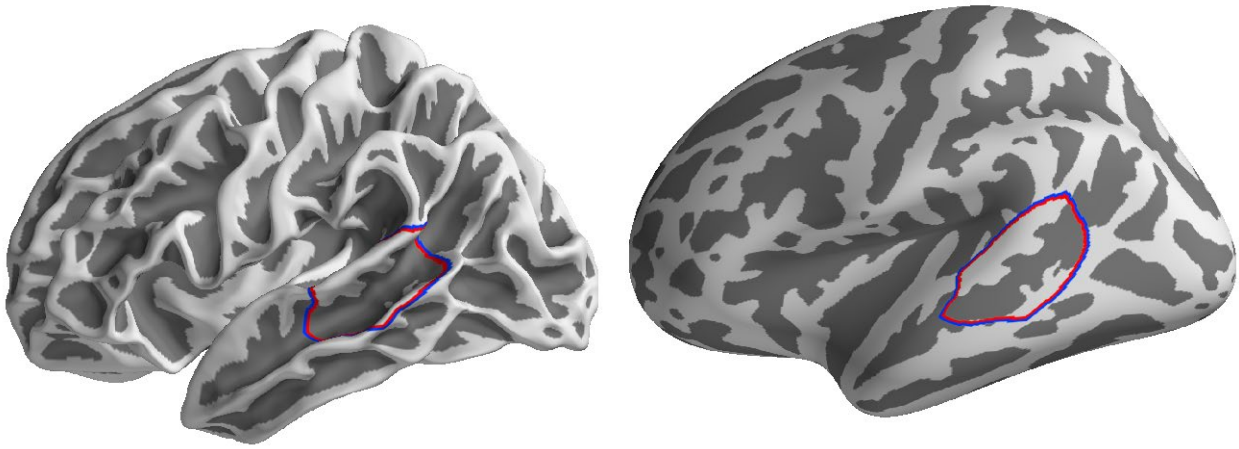


Figure 22. *ROI based on Matchin et al. (2019) PTL ROI, visualized in two ways.*

The temporal cluster test was identical to the spatiotemporal cluster test described for the localizer in **Section 5.2.6.4**, except that we had averaged over space (within the ROI) for each time point, and so were computing *t*-values only for points in time and not for points in space. This ROI is a sub-region of the Matchin et al.

(2019) ROI described above, and its origin as a cluster of sources with unified directionality may make it less susceptible to the cancelling out of current estimates that can occur in fixed orientation analyses when estimates from positive and negative sources are averaged.

Our localizer had generated two other clusters from the sentence/consonant string contrast (**Figure 21**) that we planned to use as ROIs for temporal cluster tests. One reason for investigating these ROIs was that the location and time window of their effects in the localizer (nearly contiguous along anterior superior temporal gyrus (STG) and the transverse temporal sulcus (TTS), roughly 300-500 ms after word onset) were close to Matchin et al. (2019)'s apparent syntactic prediction effect. A concern with Matchin et al. (2019)'s finding is that it reflects lexical rather than (or in addition to) syntactic prediction, which our design is intended to address. Because these clusters were generated by a partially lexical contrast, we considered them potentially useful in understanding the relationship between Matchin et al. (2019)'s findings and our own. However, we consider this analysis purely exploratory.

We followed the same procedure as for the first localizer cluster, again comparing the sentence and phrase list conditions in the time window of the first determiner phrase.

Spatiotemporal cluster tests

All of the above analyses defined spatial ROIs and evaluated their time courses for differences between conditions. As a complement to these ROI tests, in order to ensure that we did not miss any areas demonstrating an effect of our crucial contrast, we also conducted a spatiotemporal cluster test, which does not assume a

fixed spatial ROI in advance but can be limited to broad search areas of interest (here, left temporal lobe and left inferior frontal gyrus). This spatiotemporal cluster test again evaluated the sentence/phrase list contrast, in the window of the first determiner phrase (0-1500 ms from trial onset, using the preceding 100 ms for baseline correction). We also conducted an exploratory spatiotemporal cluster test, again comparing the sentence and phrase list conditions, during the window of the verb and second determiner phrase, when there is bottom-up evidence of the need for structure.

Analyzing the coordination condition

Finally, our analyses of the coordination condition were also conducted with spatiotemporal cluster tests in the same regions. The first was a hypothesis-driven (confirmatory) test investigating whether the effect of coordination observed by Lau and Liao (2018) in EEG could also be observed in MEG. This test compared the response to coordinated phrases relative to phrase lists in the second epoch, spanning the 2000 ms from the onset of the coordinator or blank middle position. This was necessarily a spatiotemporal rather than a temporal test because we could not know from the EEG result where we could expect to observe the effect in MEG.

We also conducted an exploratory spatiotemporal cluster test exploring the possibility of effects of predicting coordination during the first determiner phrase. This test compared the coordination and phrase list conditions in the first epoch.

5.3 Results

Results and discussion for all tests in the second analysis window (1500-3500 ms) are reported in **Appendix B**.

5.3.1 Behavioral data

The mean proportion of accurately answered memory probes was 0.630 (SD 0.0974) in the sentence condition, 0.653 (SD 0.106) in the phrase condition, and 0.647 (SD 0.112) in the coordination condition. There was no significant difference between accuracy in the sentence and phrase conditions ($t(30) = -1.0$, $p = .327$) or in the coordination and phrase conditions ($t(30) = -0.3$, $p = .763$).

5.3.2 Neural data

5.3.2.1 Matchin et al. (2019) ROI

In the region of interest created from the averaged PTL ROIs in Matchin et al. (2019), we averaged source-localized activity over all source points in the time window 272-484 ms after the first adjective and the time window 272-484 ms after the first noun. From Matchin et al. (2019), we expected an effect of sentence > phrase in one of those windows. However, the results of our repeated measures *t*-test on the participant averages, comparing the sentence and phrase list conditions, were not significant in the window following the adjective ($t(30) = 0.53$, $p = .603$) or the window following the noun ($t(30) = 0.10$, $p = .922$). Our data therefore does not allow us to reject the null hypothesis that there was no difference in neural activity for the sentence and phrase conditions in this ROI in either time window. In **Figure 23**, we plot the time course of activity for these two conditions in this ROI, during the 1500 ms window of the first determiner phrase. Though it appears that activity is numerically greater for the sentence condition than the phrase condition at some points in the epoch, this difference does not exceed the variation we expect due to noise.

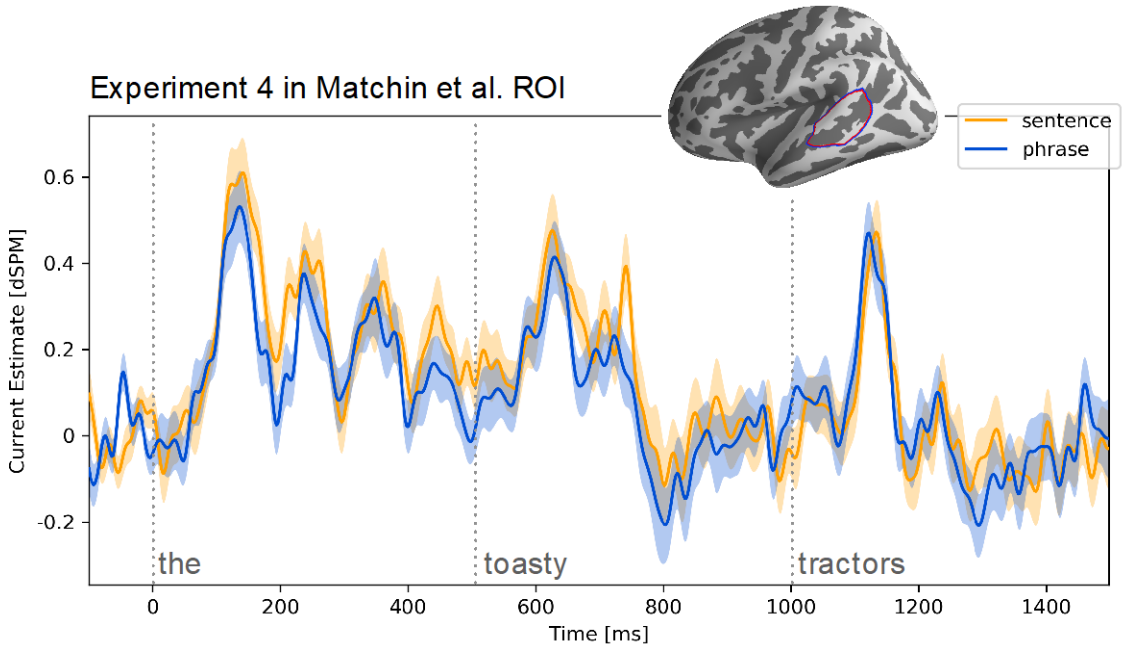


Figure 23. *Time course of neural activity in the PTL ROI based on Matchin et al. (2019), in 0-1500 ms window. Example item shows onset of each word.*

5.3.2.2 Functional ROIs from localizer

Our second analysis for the sentence/phrase contrast used functional ROIs created from the localizer. We conducted temporal cluster tests over the 0-1500 ms window within a pSTS ROI derived from the localizer's sentence/scrambled contrast, and within STG and TTS ROIs derived from the localizer's sentence/consonant contrast. Although the conditions were numerically different at the onset of the epoch, these temporal cluster tests yielded no significant clusters. In **Figure 24**, we plot the time course of activity in the 0-1500 ms window within each of these ROIs. The pSTS and TTS ROIs appear to have inverse time course patterns because they are on either side of the same gyrus. We note that current estimate values in the pSTS ROI

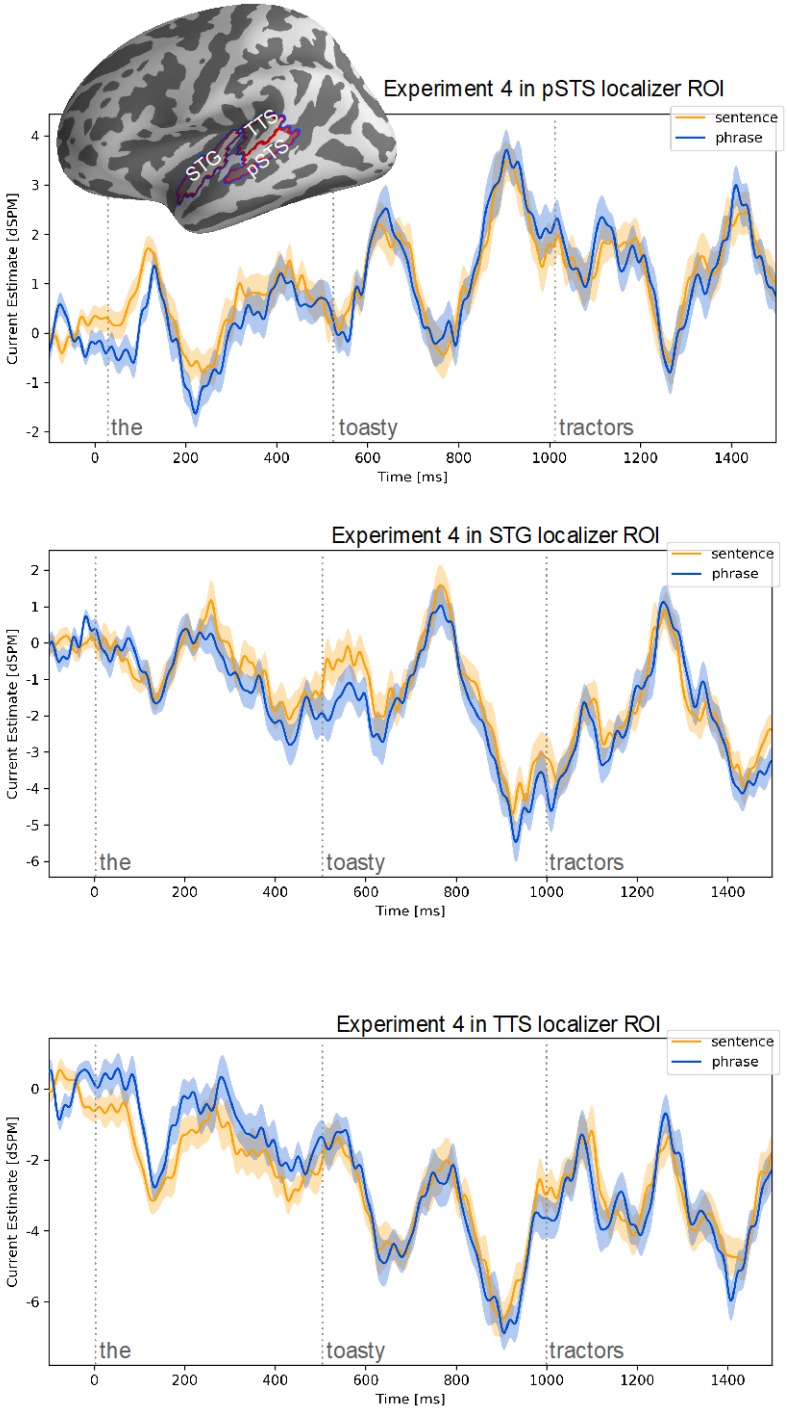


Figure 24. *Time course of neural activity in ROIs from localizer, in 0-1500 ms window. Example item shows onset of each word.*

are far higher than for the Matchin et al. (2019) ROI, suggesting that in that analysis sources with opposing directions may have been averaged together, cancelling out.

5.3.2.3 Spatiotemporal cluster tests

Our final step in testing for the apparent effect of predictive syntactic structure-building in Matchin et al. (2019) was a spatiotemporal cluster test in the 0-1500 ms window of the first determiner phrase, over the left temporal lobe and left inferior frontal gyrus. This was the same search area used for the localizer. We found no significant clusters for this contrast.

Our planned but exploratory spatiotemporal test, examining the contrast of coordination vs. phrase in the 0-1500 ms window of the first determiner phrase, yielded no significant clusters.

5.4 Discussion

5.4.1 Outline

Experiment 4 compared the neural response to a determiner phrase when participants expected that it would be the subject of a sentence or the first item in a list. Our design avoided lexical predictability so as to try to isolate syntactic prediction. In this Discussion section, we first summarize the results of our analyses in the window of the first determiner phrase. We failed to observe the syntactic prediction effect expected on the basis of Matchin et al. (2019)'s result, and it is fully possible that this null effect is simply a false negative. Our lack of significant effect is not evidence that there is no true effect of structural prediction. However, we will discuss implications that would hold if the difference between our dataset and Matchin et al. (2019)'s reflects a true difference between the cognitive computations

elicited by these paradigms, as there are a variety of possible scenarios in which it is simultaneously the case that Matchin et al. (2019) observed a true positive and we observed a true negative. We first walk through the possibility that the apparent effect of syntactic prediction actually reflects lexical prediction, which does not occur in a design lacking lexical predictability. We consider how this explanation relates to N400 effects, jabberwocky effects, and how it can be framed alternatively as a shift in probability across the lexicon rather than specific lexical prediction. We also consider what kind of design would be necessary to truly disentangle effects of structure and lexical predictability, and to what extent they are correlated in previous fMRI designs. We finish with a brief review of a number of other possible explanations for the discrepancy. Note, again, that the results and discussion for analyses in the window after the first determiner phrase are in **Appendix B**.

5.4.2 Summary of results

In this experiment, we first looked for evidence of a structural prediction effect analogous to the one observed by Matchin et al. (2019). Any effects in the window of the first determiner phrase would have to be predictive in nature, because the stimulus is identical between conditions. We created an anatomical ROI based on the ROI used by Matchin et al., and in the same time window as the original effect, during the response to the first determiner phrase's adjective and noun, we tested for a difference between the sentence and phrase conditions. We observed no significant differences.

In addition to the anatomical ROI based on Matchin et al. (2019)'s ROI, we tested for the sentence > phrase syntactic prediction effect in functional ROIs from

the results of a localizer task also completed by our participants. We used a cluster in pSTS from the sentence/scrambled localizer contrast, and two clusters along STG and TTS from the sentence/consonant localizer contrast. In these ROIs, we ran temporal cluster tests rather than restricting our window of analysis to the window in which Matchin et al. (2019) had observed their effect. In that window of the first determiner phrase, we found no clusters for the sentence/phrase contrast, in any of the three ROIs. This is interesting, as it shows that the processes driving the sentence > scrambled structure effect across all words in our localizer sentences are not in play during the first phrase in our sentence > phrase structure manipulation.

Finally, we conducted a spatiotemporal cluster test over the left temporal lobe and IFG, in the window of the first determiner phrase. This test relaxed all assumptions about where and when a predictive structure effect might occur. Again, we found no effects for the sentence/phrase contrast. We similarly found no effects for the coordination/phrase contrast with this test. There is therefore no evidence in this dataset for increased neural activity during the processing of a determiner phrase when sentence-level or coordinating structure can be expected next.

While this was surprising, we do not consider our result a failure to replicate. Our study was by no means an exact replication of Matchin et al. (2019)'s design. We eliminated lexical predictability both within and between phrases, we added an adjective to the determiner phrases, we added a coordination condition, and we compared the response to the same phrases in each of the different conditions. Our question was therefore whether the effect observed by Matchin et al. (2019) was

robust enough to persist despite these changes, and we did not find evidence that it was.

5.4.3 The role of lexical prediction

How should we interpret Matchin et al. (2019)'s evidence for neural activity that appears to reflect anticipation for upcoming sentence-level structure, together with our null effect? We first consider the possibility that the original effect occurred due to the lexical predictability that is present in the sentence condition but not the phrase condition. This would not necessarily mean that the original effect was driven solely by lexical predictability, but that it was not completely syntactic. We discuss this lexical prediction account in great detail relative to the alternative accounts that will follow. In doing so, we do not intend to imply that it is necessarily more plausible. We believe it warrants such extensive discussion simply because it would have many interesting implications if it were true.

Additionally, in proposing a lexical account of the syntactic prediction effect, we are not intending to cast doubts on the existence of neural effects of syntactic structure. All of our discussion is restricted to structure effects in posterior temporal lobe (PTL), which has also been extensively implicated in lexical processing. The deep association between lexical predictability and syntactic structure in natural language must therefore be thoroughly accounted for before we can confidently isolate syntactic structure effects in PTL.

5.4.3.1 Accounting for Matchin et al. (2019)'s prediction effect

Matchin et al. (2019) used natural sentences, such that processing of the subject determiner phrase allowed for prediction of the verb on both a lexical and a

syntactic level, and in general it was the case that content words in the sentence trials were far more related to each other than content words in the phrase trials. Our study was specifically designed to eliminate lexical predictability in the sentence condition, using random combinations of subject, verb, and object so that upcoming words in the sentence were not any more predictable on the basis of previously occurring words than they would be in a random list of words or phrases with no relationship between them. We wanted to minimize the possibility that any 'predictive' effects we observed were due to lexical prediction differences between sentences and phrases, rather than the syntactic prediction differences argued for by Matchin et al. Instead, as Matchin et al. (2019)'s prediction effect no longer occurs in the absence of lexical predictability, it is possible that it instead reflected specific prediction for the verb.

We also note that the null effect for the phrase/list contrast in the Matchin et al. (2017) fMRI data is just as consistent with a lexical prediction account as a syntactic prediction account. If the sentence/phrase contrast is due to prediction of higher-level syntactic structure, then we expect no difference between phrases that require only low-level (i.e., not sentence-level) structure and lists that require no structure. But if the sentence/phrase contrast is due to lexical prediction, it is also the case that neither the phrase condition nor the list condition allows prediction of content words.

Specific prediction of the verb would entail that a specific candidate or family of candidates for the next content word position are facilitated due to their relatedness and likelihood of co-occurring with the subject noun, and the rest of the lexicon is unaffected. These lexical predictions may be partially syntactic in nature, to the extent

that syntactic category information is one of several types of information used in the process of predicting the next word. However, repeated violations of lexical predictions in nonsense stimuli might cause the comprehender to cease generating those lexical predictions altogether. Then it would both be the case that the originally observed effect was (partially) syntactic in nature, and that removing lexical predictability prevented it from occurring. This wouldn't necessarily mean that syntactic predictions had stopped, but that we had been observing them via their impact on another process, which itself had stopped. A stronger version of this would be that syntactic and lexical predictions are so intertwined that the former cannot occur without the latter.

For this to explain the variability in the presence or absence of syntactic prediction effects, of course, generation of these lexical predictions has to be subject to modulation based on their utility in the broader context, rather than being fully automatic. It also has to be very clear in the experimental context that specific verbs or upcoming words are not predictable from the first few words in the trial and therefore shouldn't be predicted. In a 'nonsense' design like ours this is trivial because lexical predictability is never present. For Matchin et al. (2019), the block design makes clear when upcoming trials will include lexical predictability (when the upcoming block is cued as a natural sentence block) or not (all other block types). For this explanation to extend to a design with randomly ordered trial types, a cue would have to be present at the beginning of each trial indicating whether lexical predictability will hold within the trial or not.

Before continuing, we note that in discussion of the lexical prediction possibility so far, we frame the process as if it occurs by default and the presence of violations causes comprehenders to stop making predictions. However, it is also possible to frame this such that predictions are initiated only when there is positive evidence for predictability. This latter view might be more straight-forward in that it does not require hypotheses about how many prediction violations are necessary before predictions stop. Instead, any evidence for predictability could be the trigger to start. This is consistent with evidence from Lau et al. (2013) that in a two-word semantic priming paradigm, facilitation in related pairs occurs even when only 10% of pairs are related, and facilitation appears to be heightened when a higher percentage of pairs are related. We could then explain this as predictions being initiated to the extent that they seem helpful (with a very low initial bar). Of course, this question should be answerable with an order analysis examining when facilitation changes in response to the predictability context, perhaps along the lines of the approach used by Delaney-Busch et al. (2019).

5.4.3.2 Relation to the N400

How does this account relate to the extensive ERP literature on contextual prediction and N400 effects? Reduced amplitude of the N400 in supportive contexts is thought to be a reflection (by some views) of the facilitative *impact* of lexical prediction at the point at which the predicted (or unpredicted) word is encountered (Lau et al., 2008, 2009, 2013). Here, in contrast, we are primarily concerned with effects reflecting the *generation* of the prediction. Dikker and Pylkkänen (2013) report MEG evidence for apparent lexical preactivation processes occurring in left

mid-temporal cortex (among other locations) at the timepoint at which the prediction is generated, when a specific word can be predicted from a preceding picture. However, findings from these two perspectives (the generation of the prediction, and its later impact) would ideally constrain each other. Could the effect observed by Matchin et al. (2019) be consistent with generating specific predictions for the verb in particular? Could these be the same predictions that lead to N400 reductions for predicted words? We will focus here on one salient consideration in assessing the potential correspondence between the Matchin et al. effect and the N400, which is whether N400 evidence is consistent with strong top-down modulation of prediction. This would be necessary under a unified account because the Matchin et al. prediction effect appears to cease in list and nonsense conditions when it isn't warranted.

There is evidence that the degree of facilitation that occurs on the N400 is influenced by the broader context. Lau et al. (2013) show that for words presented with a single preceding word for context, there is a larger reduction of the N400 amplitude for the contextually supported item when word pairs are presented among a high vs. low proportion of other related word pairs. A higher proportion of related pairs appears to encourage prediction, which leads to greater facilitation of the related item. Brothers et al. (2019) also show that in a manipulation in which one ("reliable") speaker produces sentences with predictable endings and a second ("unreliable") speaker produces sentences with endings that are plausible but not predictable, facilitation as measured on the N400 is enhanced in the reliable, more predictable context.

Can the predictions driving N400 reductions can be "turned off" completely, as seems to be occur in our study due to the use of nonsense stimuli? One piece of supporting evidence comes from Van Petten (1993), who manipulates whether associated word pairs (e.g. moon and stars) occur within congruent or anomalous sentences, and finds the N400 reduction on the second word to be mitigated in the anomalous sentences. Another relevant data point comes from Payne et al. (2015), who examine the N400 amplitude at each word position in natural sentences, nonsense sentences, and word lists created by scrambling within the nonsense sentences. For open-class words, they find that facilitation increases with word position only in the natural sentence condition. Both results suggest that the necessary predictions are either not occurring in the nonsense sentences, or that they are occurring but not, of course, facilitating the items that end up being presented.

Stronger evidence that these predictions are actually stopped would be if the N400 reduction for a contextually supported item within a natural sentence ceases to occur when that natural sentence occurs in an experimental context that strongly discourages prediction, i.e., if most of the other sentences are nonsense. This would be a challenging experimental design because the presence of natural target trials in which lexical prediction would not be futile would itself undermine the goal of having the experimental context discourage prediction. The proportion of target trials would therefore have to be extremely low. For Lau et al. (2013), for example, the proportion of related pairs in the low-relatedness condition was 10%, and the typical N400 reduction was still observed in those related target trials.

A final consideration in asking whether the Matchin et al. (2019) effect could be reflecting the kinds of predictions that lead to N400 reduction is that, in that study, no facilitation was seen on the verb in the sentence condition, relative to the noun or verb in that position in the phrase condition. This might seem surprising on an account in which prediction of the verb in the sentence condition is what drives the sentence > phrase effect. However, Matchin et al. did not aim to use strongly predictable verbs, nor did they fully match lexical items across conditions in specific sentence positions, which could have been a separate source of N400 variability. It's also worth noting that according to the lexical prediction account, prediction for the object should be occurring on the verb in the sentence condition, but not in the phrase condition. This could induce a prediction effect in the opposite direction of the facilitation we would be looking for. In fact, something like this kind of 'canceling out' of lexical prediction generation and lexical facilitation must be happening on this account, or else we would expect to see prediction generation effects on every content word of the sentence.

We note that Matchin et al. (2019) did find several additional effects of structure (sentence > phrase) at other points in the sentence besides the first DP, and we did not observe any such effects in our study. Could these have been lexical prediction effects as well? We think this is unlikely, because the timing details of the other reported effects of structure do not obviously follow from the lexical prediction account spelled out above. It therefore seems more likely that they reflect other computations. For more detailed discussion, see **Appendix B.2.3**.

5.4.3.3 Alternative framing: lexical probability shifts

Another way to conceive of a lexical prediction effect is that each new word that is comprehended leads to a probability shift over the entire lexicon, rather than a specific prediction for some small subset of items. In many respects this account is very similar to, or possibly a description at a different level, of the specific prediction account. One important difference, though, is that such a probability update could reflect probability of occurring as the next content word, or general probability of co-occurring in the context, and in either case the observable effect would be modulated by the degree of probability shift. Another difference, and potential advantage, is that we don't necessarily have to consider prediction to be "on" or "off." This account allows for two alternative explanations for the prediction effect occurring only on the subject noun.

First, in experiments like ours, using unconnected sentences, the comprehender has no information before the onset of a trial about what words are likely to occur other than the frequency distribution over all words in the lexicon. A new trial is a new context, and comprehension of the first word should allow for a dramatic shift in what other words are likely to co-occur in the trial. Additional words may not lead to additional appreciable shifts in general co-occurrence probability, unless the first word was a homophone whose meaning is disambiguated by the next content word. To use an example from Matchin et al. (2019), the probability of the word "verse" occurring in a trial with "poet" may not be very different from its probability of occurring in a trial with both "poet" and "recite." This is testable with corpus data.

If the probability shift is for probability of occurring as the next content word, then each new word should lead to a more appreciable shift, just as, in the specific prediction account, each new word should lead to a new prediction. However, the difference is that if the effect is due to the cost of making predictions vs. not making predictions, it should be constant on each word when it can occur. But if the effect reflects the degree of probability shift for the upcoming word, given the word just encountered, the size of the effect may vary and in some cases it may not be observable. The effect from Matchin et al. (2019) occurring only on the subject presents an interesting possibility, that in fact subsequent words in the sentence after the first word do not yield any appreciable probability shift over the lexicon. This would be an implication then that the shift in probability for what can occur next after accessing "the poet" (when previously there was no way to know what could come next) is far larger than the shift in probability for what can occur next after accessing "recite" when already having processed "poet."

For the specific prediction account, we considered whether the effects observed by Matchin et al. (2019) on all open and closed-class words could also constitute effects of lexical prediction rather than syntactic structure. We concluded that this was unlikely, and it remains unlikely under the probability shift account, in part because the processing of each determiner, for example, should not lead to a larger probability shift in the sentence than the phrase condition. However, there is one sense in which the probability shift account is more compatible with effects occurring on every word. Rather than reflecting the probability shift itself, these effects could reflect some property of the probability distribution (e.g., entropy) that

continues to be different on the determiners due to changes caused by the content words in the sentence condition.

Finally, we consider to what extent the lexical probability shift proposed under this account plausibly incorporates syntactic information as well, reflecting an influence of the presence of higher-level syntactic structure. It is easy to see how probability shifts due to syntactic category information (i.e., shifts toward nouns and adjectives after a determiner) could occur in parallel or via shared mechanisms with probability shifts driven by lexical content. In the extreme, syntactic prediction could reduce to a probability update over lexical items, according to their syntactic categories, and would then be inherently a lexical process. If the probability updating process is in fact a single mechanism with multiple informational inputs, it could be that severe expectation violations with respect to just one of the inputs affect the entire mechanism. Syntactic predictions might also stop, or we might simply cease being able to observe them because we had only seen their impacts via the lexical effects. Stopping the lexical probability update in list and nonsense conditions would rely on it becoming clear in the experimental context that co-occurrence probability from word to word or generally within a trial is not in line with normal statistics for the language (besides it being necessary for this probability update to be stop-able in the first place). This is largely the same as for the specific prediction account, and is a cue that will likely be confounded with the presence of syntactic structure in most designs.

5.4.3.4 Relation to jabberwocky effects

We argue in the previous sections that at least some apparent effects of structural prediction may actually reflect differences in whether lexical prediction is occurring. This raises interesting questions about the neuroimaging literature on jabberwocky sentences, which do not contain lexical content (and thus, like our nonsense stimuli, should not allow for lexical prediction) but which do still exhibit effects of structure. Jabberwocky manipulations are in fact predicated on the very idea that it is possible for syntactic processing to occur (whether predictive or not) when lexical content is not only nonsense but absent. In what follows, we argue that the structure effects observed in jabberwocky paradigms are likely to arise from an altered parsing algorithm, and thus should not be expected to correspond with structure effects observed for natural sentences. We also argue that jabberwocky parsing might have much more temporal variability from subject-to-subject or trial-to-trial than the parsing of natural sentences. This could explain why jabberwocky effects of structure occur in fMRI but are rare in MEG. Our reasoning is as follows.

There are (at least) three sources of syntactic category information that are likely used in syntactic structure building for normal stimuli: stored representations for function words, for content word roots, and for inflection on content words. Jabberwocky stimuli have two of these, as they retain function words and inflectional markings on content words, but not known content word roots. Any structure-building therefore has to be accomplished from function words and inflection alone. We know that these two sources apparently do not allow full approximation of normal syntactic structure-building, since jabberwocky structure effects in fMRI tend

to be smaller and less robust than natural structure effects (as in, for example, Goucha & Friederici (2015)), affirming that there is a role for syntactic category information gleaned from content word roots. We also see specifically for Matchin et al. (2017) that jabberwocky structure effects are elicited by the sentence/list but not the sentence/phrase contrast (while structure effects for the natural stimuli are elicited by both contrasts). The essential difference between the sentence/phrase contrast for jabberwocky stimuli and the sentence/phrase contrast for natural stimuli, in this design, is the availability of the identity of the main verb. It may be, then, that the structure effects elicited by a sentence/phrase contrast rely on information from the identity of that content word.

We note that this raises an interesting possibility that the magnitude of jabberwocky structure effects might vary across languages according to the distribution of syntactic cues in that language. For languages whose structure-building relies largely on function words, syntactic processing of jabberwocky might not diverge significantly from the processing of natural stimuli. For a language that does rely on syntactic category information from the identity of content words, the loss of a primary information source might have more of an impact.

However, we do not think it is simply the case that the processing of jabberwocky proceeds as usual and just with fewer cues to syntactic category, or that jabberwocky structure effects are a pure reflection of the contributions of function words and inflection. Instead, loss of a primary cue is liable to change the way in which syntactic processing is carried out with the remaining cues. In the extreme, consider the possibility that, for syntactic processing of normal stimuli,

comprehenders rely exclusively on syntactic category information from content word roots. Jabberwocky structure effects would not then reflect "what is left" of structure effects from normal stimuli, without the aspects that are accomplished with lexical information. They would reflect an entirely new processing mode driven by whatever information can newly be recruited.

This is why we believe that whatever drives jabberwocky structure effects is not something that we should assume is also occurring in manipulations of normal stimuli. Comprehenders might, for example, be pushed away from a top-down parsing mode because of the number of words for which syntactic category cannot be discerned until the next word provides more information (this would be any nonword without obvious inflection). Comprehenders might also vary amongst themselves in how they respond to the loss of a primary cue. Some might cease structure-building altogether. To our knowledge, there is no detailed theory of jabberwocky syntactic processing and its temporal properties, or to what extent it is subject to individual differences. ERP work on whether the detection of syntactic violations in jabberwocky differs from in natural stimuli seems to show mixed results (Hahne & Jescheniak, 2001; Yamada & Neville, 2007).

Importantly, temporal variation in syntactic processing of jabberwocky stimuli between participants could explain why jabberwocky structure effects from sentence/list contrasts are so far attested only in fMRI, and not in methods with temporal resolution: ECoG (Fedorenko et al., 2016), MEG (Matchin et al., 2019), or a related EEG design (Lau & Liao, 2018). This is particularly evident for Matchin et al. (2017, 2019) using essentially the same paradigm in fMRI and MEG. In fMRI, if all

participants engage in structure-building at some point in the processing of a sentence, the block design should manifest a structure effect. In MEG, however, where activity is compared at each millisecond, there will not be a structure effect in the group average if structure differences are temporally jittered across participants. This could also be a problem with between trial variation, if a study does not use the same syntactic structure for all trials, because different structures might vary in the extent to which the different syntactic cues are necessary.

In **Section 5.4.4.1**, we discuss the possibility that our nonsense stimuli were actually processed as if they were jabberwocky, and that the resulting temporal variability explains the lack of structure effects.

5.4.3.5 Disentangling the lexical and syntactic accounts in a new design

Fully crossing structure and lexical predictability

In the previous sections, we have considered how the syntactic prediction effect reported by Matchin et al. (2019) could be explained instead by a lexical prediction account, and what implications this would have. The fundamental problem we are faced with is that a lexical prediction account, a syntactic prediction account, and a syntactic-via-lexical prediction account all predict the difference between natural sentences and unpredictable lists that Matchin et al. (2019) observed.

To disentangle them, we would need, to start with, a study that fully crosses structure (sentences vs. lists) and lexical predictability (natural vs. nonsense), so that we can examine the conditions for which the different accounts do make different predictions. This could be done by essentially combining the design of our current experiment with the design of Matchin et al. (2019) (see **Table 9** below). Sentence

blocks would contain either natural or nonsense sentences, and phrase blocks would consist of the same stimuli without the verbs. The two determiner phrases occurring within each trial would, as in our current study, be matched between conditions.

Table 9. *Example stimuli for study crossing structure (sentence vs. list) and lexical predictability (natural vs. nonsense).*

	Sentence	List
Natural	the spacecrafts reached the stars.	the spacecrafts the stars
Nonsense	the spacecrafts measured the wolves.	the spacecrafts the wolves

In **Figure 25**, we provide a schematic of observed PTL activity in the two conditions in the Matchin et al. design (top left), as well as predicted PTL activity in the four conditions of the 2x2 manipulation under the syntactic account (bottom left), the lexical account (top right), and the syntactic-via-lexical account (bottom right). The observed difference between natural (lexically predictable) sentences and nonsense (not lexically predictable) lists occurs, as we have noted, for all three accounts, but there are other pairwise comparisons that can distinguish them.

For example, one difference predicted only by the syntactic account is between unpredictable sentences and unpredictable lists. Our study fails to find this difference. What difference would distinguish the lexical (top right) and syntactic-via-lexical (bottom right) accounts? Under the lexical account, there is no simple effect of structure when there is lexical predictability. There is, however, a simple effect of lexical predictability when there is no structure.

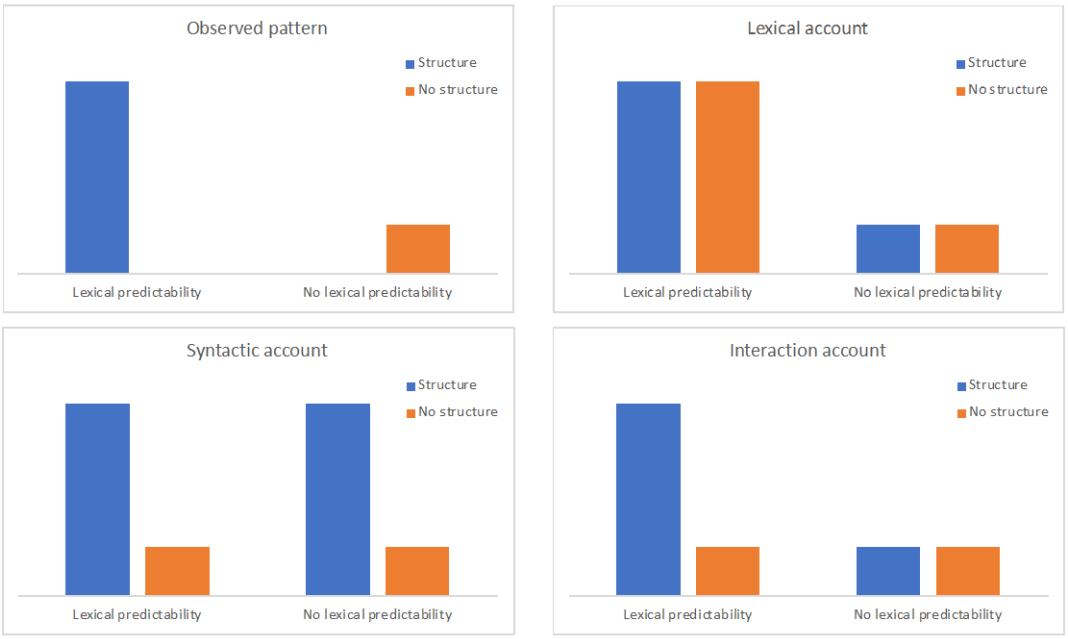


Figure 25. *Schematic of observed PTL activity (top left) and expected PTL activity under syntactic (bottom left), lexical (top right), or interaction (bottom right) accounts of the Matchin et al. (2019) prediction effect.*

Under the lexical-via-syntactic account, we expect a simple effect of structure when there is lexical predictability. As we have presented it in **Figure 25** there is no simple effect of lexical predictability when there is no structure, but whether or not this is true would depend on the specifics of the interaction. What is important for this account is that there is some structure effect in the lexically predictable conditions. The data point we most need then is the test for a structure effect when lexical predictability is present in both sentences and lists.

The importance of cues for engaging prediction

Beyond including these two conditions, there is one additional consideration that is crucial to a study being able to distinguish the different prediction accounts: comprehenders must know at the beginning of a trial whether prediction is warranted.

In all three of our theories under consideration, the assumption is that the reason for the contrast between a condition with and without structure or with and without lexical predictability is that prediction (whatever its nature) does not occur when it is not warranted. If comprehenders don't know whether prediction is warranted, they could not possibly modulate whether or not they engage in prediction, and then our expectations about what should or shouldn't happen in each condition have no basis.

In a block design, this is trivial, as participants are explicitly told what type of stimulus they will encounter in the block. In an event-related design, however, trials from different conditions are presented in a random order, and participants are not told at the onset of a trial what they are about to process. For prediction to be modulated, there must then be an early cue as to what type of trial it is. This is relatively simple for syntactic prediction, because whether or not the trial has syntactic structure is generally apparent in the first two to three words. In contrast, whether or not content words within the trial are lexically related and therefore predictable is not evident, we suspect, until most of the trial has been comprehended. Modulation of lexical prediction in an event-related design, then, has to rely on alternative cues as to whether it will be warranted.

What might participants use as a heuristic for whether lexical prediction is warranted? In a design that presents only natural sentences and unpredictable lists, the presence or absence of structure is in fact a perfect cue for the necessity of lexical prediction (even if the prediction is not syntactic in any way). In designs for which there are, for example, lexically predictable trials both with and without structure, or structured trials both with and without lexical predictability, structure is no longer a

perfect cue. However, we can expect from e.g. the finding of Lau et al. (2013) that lexical prediction occurs even when warranted on only a small percentage of trials. Therefore, if there are any trials with structure in which there is also lexical predictability, we expect lexical prediction to occur on all trials with structure. In effect, this means that in event-related designs, because structure is the only reliable cue to trial type, there are only two "conditions" from the perspective of the participant, and the condition will trigger prediction if predictability is ever observed for that condition.

If the predictions are purely syntactic, this collapsing of conditions doesn't matter. But if predictions are lexical, fully or in part, then within each structure condition, lexically predictable and unpredictable trials will be collapsed and both treated as lexically predictable. This means it is not in fact possible to have a condition in which lexical prediction does occur and a condition in which it does not occur, within the same experiment, unless structure is a perfect cue for lexical predictability. In other words, because there is only one cue to trial type for the participants (is there structure or not, and has structure been associated with lexical relatedness such that prediction is warranted?) we cannot modulate lexical predictability in a single event-related design without confounding it with structure. De-confounding the two would seem to require a block design, or an additional way to cue for trial type other than presence of structure. It is possible, however, to examine structure effects when in both conditions lexical predictability is present.

Other approaches

A variant on this design we have been describing, if conducted in EEG, would relate to our discussion of relevant work on the N400 (**Section 5.4.3.2**). In the design proposed here, one could present an unexpected sentence trial during or at the end of a phrase block; if prediction is indeed not occurring in the phrase block, the response to the verb in this sentence trial should be less facilitated than the response to a verb in sentence trials that occur in a sentence block. If this is manifesting on the N400, then we would take this to be evidence that lexical prediction for the verb is normally occurring.

Though completely unrelated to this design, a final useful data point in asking about lexical influences on apparent effects of structure would be a variant of naturalistic paradigms in which neural signal in response to reading or listening to a story is correlated with complexity metrics for a hypothesized syntactic parser (e.g., Brennan & Hale (2019)). It would be informative to know whether the same correlations are obtained for nonsense stimuli.

5.4.3.6 Disentangling the lexical and syntactic accounts in previous designs

In the following sub-sections, we describe several existing structure manipulations in fMRI and the extent to which they are able to distinguish the syntactic and lexical accounts, per the considerations we have outlined above. It is not within the scope of this chapter to survey the entire literature on this topic; the studies we cover are simply examples. Note that many of these studies include a parallel jabberwocky manipulation as their intended method for dissociating syntactic and

lexical effects, while what we are discussing is simply the extent to which this dissociation can occur or not within the manipulation of real-word conditions.

Fedorenko et al. (2012)

In the design used by Fedorenko et al. (2012) (more details about the stimuli are reported in Fedorenko et al. (2010)) lexical and syntactic predictability are perfectly correlated in exactly the way they are for Matchin et al. (2019). The study uses a block design, but sentences always have lexical predictability and lists never do, because lists are created by scrambling words across sentences rather than within them. Their activation maps for the pairwise comparison between the sentence and scrambled conditions indicates that there is a structure effect in PTL. However, a difference in activity in PTL for this contrast does not distinguish lexical and syntactic explanations.

The localizer task that we used in our study, though event-related, in fact has the same problem, as this is a very common design. In discussing the localizer results we raised the point that the structure effects observed in that dataset were a good fit with Matchin et al. (2019)'s structure effects, suggesting that those manipulations had something in common that is not shared with Experiment 4. The localizer, too, uses only natural sentences and random, unpredictable word lists, and the structure effect we found could just as well have been a lexical predictability effect. Our analysis of the localizer data collapsed over word position, so we do not know when that effect occurs in the sentence (i.e., whether it is predictive).

Humphries et al. (2006)

Humphries et al. (2006) is a very promising comparison point in fMRI because that design fully crosses syntactic structure (sentence or list) with semantic content. The three levels of their semantic manipulation are congruent stimuli, random stimuli, or pseudowords (jabberwocky). The sentence conditions are either natural sentences, nonsense sentences very similar to the current study in that they are created by randomly swapping in different content words of the same syntactic category at each position ("the freeway on a pie watched a house and a window"), or typical jabberwocky sentences in which content words are replaced with pseudowords but function words are retained in their original positions. List conditions are created by scrambling the words within the sentence. Therefore, a congruent list trial would contain words with the same degree of lexical association as in a congruent sentence trial, but without syntactic structure. A random list trial would contain words with the same lack of lexical association as in a random sentence trial. This is different from typical list conditions, which do not usually contain the same words in a given trial as a natural sentence because they scramble across the full set of sentences. Structure and lexical predictability are therefore not correlated in this design from the standpoint of the experimenter.

However, Humphries et al. (2006) use an event-related rather than a block design, so participants do not know on any given trial whether it will be semantically congruent or random. The onset of a trial makes clear only whether it involves pseudowords or real words, and whether there is syntactic structure or not. Lexical predictability is possible when structure is present and also when it is not, so from the

perspective of the participant this design collapses to a comparison of potentially lexically predictable sentences and potentially lexically predictable lists. Fortunately, as we identified earlier, this is the crucial comparison in distinguishing the lexical and syntactic-via-lexical accounts, as the lexical account predicts no difference in PTL between the sentence and list conditions, while the interaction predicts less activity in the lexically predictable list than the lexically predictable sentence. Though their design technically includes the comparison that we would like to see in fMRI to corroborate the null effect we observe in the current MEG study (sentence vs. list, when neither has lexical predictability), we have to assume that lexical prediction is occurring in those conditions since they are indistinguishable from the predictable stimuli at the onset of the trial.

Humphries et al. (2006) report no differences between their conditions in PTL, but whole brain contrast maps show activation in PTL in all conditions. They did find a main effect of syntactic structure in ATL, which likely reflects the semantic composition that can occur in sentence but not list conditions. The lack of a structure effect in PTL when lexical predictability is matched seems to support the possibility that in a typical sentence/list contrast, the reason for effects observed in PTL is actually not the presence of structure.

Goucha and Friederici (2015)

Goucha and Friederici (2015) use an event-related design with a predictable sentence condition, an unpredictable (nonsense) sentence condition, and a random list condition (as well as several jabberwocky conditions). Their random list condition has the advantage of being lexically matched with the nonsense sentence condition.

However, because participants can't know at the onset of a sentence whether it will be lexically predictable or not, this design, from the participant perspective, collapses to the same conditions as Matchin et al. (2017) and Fedorenko et al. (2012). Goucha and Friederici find a structure effect in PTL for the contrast between unpredictable sentences and unpredictable lists, but across the experiment, lexical prediction is warranted when there is structure, but not when there is not structure. This contrast would otherwise have been quite useful if the natural sentences were not included in the experiment.

Mollica et al. (2020)

A recent study by Mollica et al. (2020) is an interesting example because degree of structure is explicitly manipulated, in a way that would seem to be quite informative for our question. In an event-related design, they present natural sentences as well as several conditions of natural sentences in which an increasing number of word position swaps have been introduced, such that with more swaps, the trial becomes more like an unstructured list. They also include a random word list condition. A second experiment includes a swap condition specifically designed to break up local dependencies. In Experiment 1, they find no difference (in PTL, or any of the language areas they examine) between natural sentences and sentences with word position swaps, but all of these condition that have lexical predictability show more activity than the random word list condition. In Experiment 2, they find that the more extreme swap condition patterns with the random word list rather than the natural sentence and more mild swap conditions.

The problem with interpreting these results with respect to their implications for syntactic versus lexical prediction, however, is that we cannot know whether from the participant perspective the mild swap conditions presented as sentences or lists, or to what extent participants differentiated among conditions in any more detailed way than having structure or not having structure. The take-aways from this study vary widely according to the assumptions that we make about how participants classified the trials, and so we refrain from interpreting further here.

5.4.4 Other possible accounts

Moving beyond lexical predictability, in the following sub-sections we discuss four possible alternative explanations for the difference between our finding and Matchin et al. (2019): (1) nonsense stimuli are parsed differently from natural sentences, (2) syntactic structure was ignored due to the lack of semantic coherence, (3) the difference is due to MEG analysis differences, and (4) Matchin et al. (2019) observed a false positive. There are two additional possibilities that we will not consider in depth but that we do acknowledge. The first is that the prediction process is strategic rather than automatic, and largely governed by task demands. The effect would therefore be subject to variation across experimental settings, participants, and stimulus sets, such that we only happened to not observe it. The second is that our addition of an adjective within the determiner phrases was the crucial change, and the prediction effect observed by Matchin et al. (2019) only occurs in maximally simple sentence structures.

5.4.4.1 Syntactic parsing is different for nonsense stimuli

In designing this study, one of our motivations for using nonsense rather than jabberwocky stimuli was that all of the syntactic cues available in natural stimuli would still be available. We expected to be targeting syntactic processes, and so we hoped that our design would leave syntactic processing intact, as we could not expect would be fully the case in jabberwocky. And yet, we do not see structure effects for our nonsense stimuli in MEG. As we have described in previous sections, this suggests the possibility that the Matchin et al. (2019) effect instead reflects a process related to lexical predictability. However, we also consider the possibility that our nonsense stimuli are actually being processed as if they are jabberwocky, and the identity of content words is not available.

Some support for nonsense stimuli being processed like jabberwocky comes from the Payne et al. (2015) study showing that facilitation on the N400 for closed-class words increases with word position in a sentence for nonsense but not natural stimuli. This suggests that nonsense stimuli lead to increased attention on function words, which is exactly in line with what we understand comprehenders to be doing in syntactic processing of jabberwocky, where function words are the main source of syntactic information.

As we discuss in **Section 5.4.3.5**, above, jabberwocky manipulations tend not to show structure effects in methods with good temporal resolution, and we argue that this could be because jabberwocky parsing is subject to increased temporal variability. Thus, if the nonsense stimuli in our MEG study are parsed as jabberwocky, this could explain why we do not observe structure effects, even if the

Matchin et al. structure effects are in fact due to syntactic structure. If this account is right, our stimuli might yield structure effects in fMRI (though the stimuli would have to be altered slightly to match the number of words per trial in each condition). If our stimuli do not produce structure effects in fMRI, this would be stronger evidence that the presentation of nonsense stimuli leads comprehenders to process the input as if it is not structured at all, or to stop paying attention entirely.

A variant on the nonsense-as-jabberwocky explanation for the lack of structure effects in our study is that participants were uniformly pushed from a top-down to a bottom-up processing mode because of the constant lexical violations. As we discuss in **Appendix B**, we do see differences between the sentence and phrase conditions in our study, but only in response to the verb. There, it is impossible to say which effects are lexical only (i.e., they reflect the difference between seeing a verb and seeing nothing) and which might actually be effects of structure-building. If one of those effects is a structure-building effect, it could be that structure-building happens predictively in the Matchin et al. (2019) design but only in response to the bottom-up input that requires it in our design.

Both of these explanations (nonsense as jabberwocky, and nonsense being bottom-up) assume that the original Matchin et al. (2019) effect was in fact syntactic, but nonsense leads to changes in syntactic processing. This makes them different from the possibilities we described above in which the original Matchin et al. (2019) effect was due to a lexical or syntactic via lexical process, and the lack of lexical predictability led that process to stop. What they all converge on, however, is the

importance of lexical predictability in normal syntactic processing, and the difficulties in isolating a syntactic process from a lexical one.

5.4.4.2 Syntactic structure of nonsense stimuli is ignored

We also acknowledge the possibility that the syntactically licit but semantically nonsense sentences that we used were so strange, difficult to comprehend, and ill-suited to semantic composition that participants stopped processing them as structured input. This would neutralize the structural manipulation.

The fact that we observed structure effects in the localizer, which did not have a task, but not in this experiment, which used a memory probe task, suggests that the problem is not overall failure to attend in the experimental session. Accuracy in the memory probe task was also above chance. However, we would ideally have an independent indicator that composition or structure-building were occurring (such as, for example, an effect of the bigram probability of the noun given the adjective), or that the more structured input was processed differently in any way.

5.4.4.3 Analysis differences

Another possible concern about our failure to observe the syntactic prediction effect reported by Matchin et al. (2019) was that our MEG analyses differed slightly, as we chose to use baseline correction and a fixed orientation of the dipole during source localization. To address the possibility that our null result was due solely to this difference, we made a post-hoc decision to re-run our targeted ROI test using loose (rather than fixed) orientation of the dipole and no baseline correction, to align the analysis as closely as possible with Matchin et al. (2019). Repeating the *t*-tests we

had run on the participant averages in the Matchin et al. ROI, comparing the sentence and phrase list conditions during the adjective and noun time windows, we again failed to find effects of sentence > phrase. In fact, we found a significant effect in the opposite direction (phrase > sentence) in the adjective time window ($t(30) = -2.13$, $p = .041$), with no significant difference between the two conditions in the noun window ($t(30) = -1.71$, $p = .098$).

In **Figure 26**, we plot the time course of activity for these two conditions in this ROI, during the 1500 ms window of the first determiner phrase, using loose orientation and no baseline correction. The profile of activity appears roughly comparable to what is observed in Matchin et al. (2019), reproduced in **Figure 27**, except for the reversal of the sentence/phrase effect. Note that the comparison is between "the *adjective noun*" (sentence condition) and "the *adjective noun*" (phrase condition) in our data vs. "the *noun modal*" (sentence condition) and "the *noun* the" (phrase condition) in the Matchin et al. (2019) data.

5.4.4.4 Matchin et al. (2019) observed a false positive

Finally, if the prediction effect observed by Matchin et al. (2019) was a false positive, then a true negative in our study would indicate either that prediction of sentence-level syntactic structure does not occur in response to initial determiner phrases, or that it always occurs but is not modulated by our block design. In other words, it is automatic in such a way that the knowledge that the entire block of input will be unstructured does not allow participants to avoid generating structural predictions in response to what seems to be a subject determiner phrase.

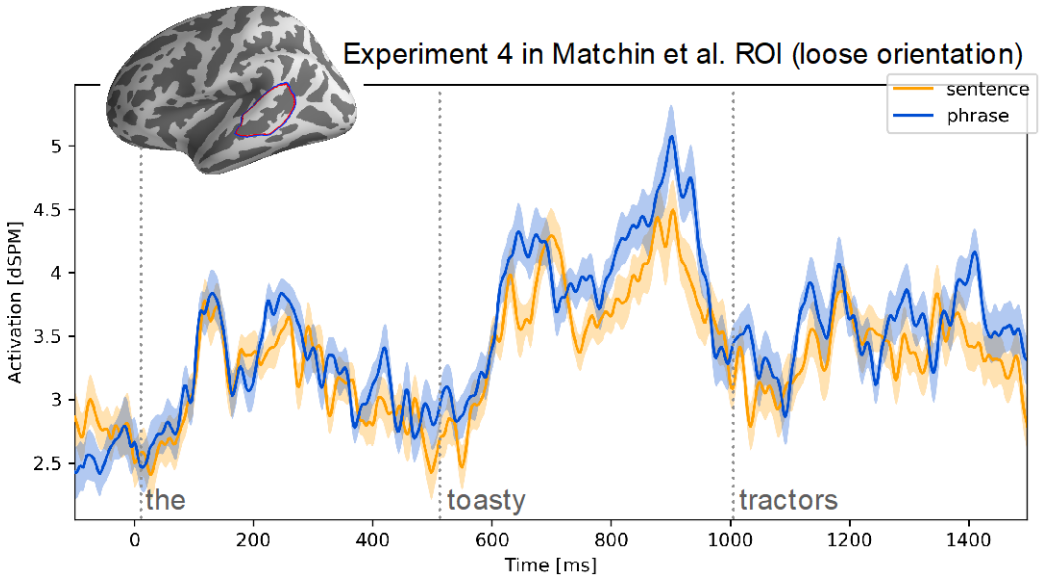


Figure 26. *Time course of neural activity in the PTL ROI based on Matchin et al. (2019), in 0-1500 ms window, using loose orientation. Example item shows onset of each word.*

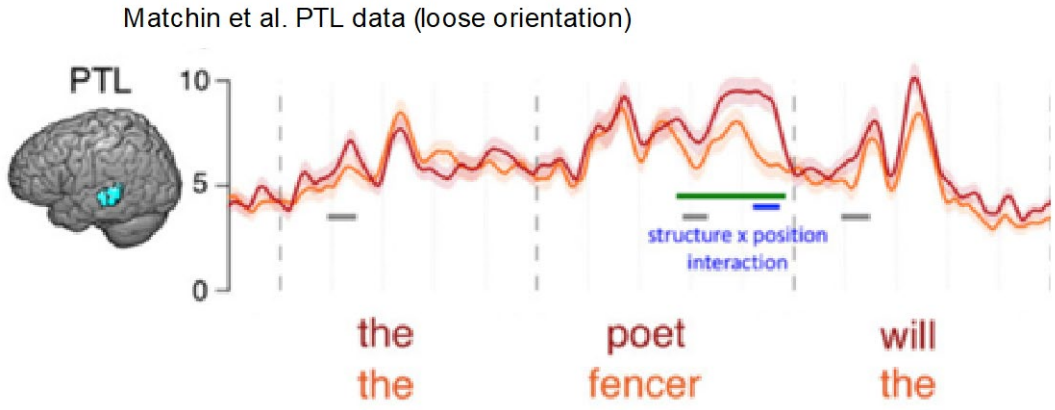


Figure 27*. PTL ROI and time course of neural activity from Matchin et al. (2019)'s Figure 4.*

5.5 Conclusion

Our study was intended to dissociate syntactic structural prediction from effects of simple lexical prediction in MEG. We compared the response to determiner phrases that were expected, by comprehenders, to be the subject of a sentence or the first item in a list. The stimuli lacked lexical predictability so that the only difference between our conditions would be whether or not syntactic structure could be predicted. A previous study using natural sentences had reported an effect that appeared to reflect syntactic prediction in a very similar design. We analyzed MEG data in the region of interest reported by that previous study, in regions of interest identified by a sentence localizer task in the same experimental session as our study, and using spatiotemporal cluster tests. However, in the absence of lexical predictability, we did not find evidence of a syntactic prediction effect.

This experiment was primarily concerned with differences between structured and unstructured stimuli in the time window of the first DP, when effects could only reflect prediction. However, this design also had the potential to demonstrate effects of syntactic structure later in the trial, once there is bottom-up support for it. As we report in **Appendix B**, we did observe many differences between the neural response to the verb, the "and" in the coordination condition, and the blank between DPs in the list condition. In this way, our dataset presents an unexpected opportunity to anchor potential structure-building effects within the well-documented temporal landscape of visual word recognition in MEG.

We have discussed several possible explanations for our lack of syntactic prediction effect, including the possibility that participants stop processing syntactic

structure when presented with semantically nonsensical stimuli, or that the lack of lexical predictability causes a shift from top-down to bottom-up parsing. We also consider that syntactic prediction may occur in natural sentences but is tightly connected with the process of lexical prediction. When lexical prediction is no longer warranted, syntactic prediction could then cease as well. This complicates our work toward a neural index of syntactic prediction. However, there is also a sense in which it bolsters the idea, fundamental to this dissertation, that syntactic information could influence word recognition in the first place. During listening, the activation of lexical candidates in response to initial phonemes is a form of prediction for what the incoming word will turn out to be, and in Chapters 2-4 I have been concerned with how this interacts with syntactic prediction. In the current study, what we have observed is that during reading, the prediction of upcoming words might be importantly linked to the prediction of syntactic structure. This raises many interesting questions about the relationships between predictions that occurs on the basis of auditory, lexical, or syntactic information, as well as any influence of the incrementality of the input on the timing of these interactions.

Chapter 6: General discussion

6.1 Summary of this dissertation

The focus of this dissertation has been a mechanistic account of how syntactic context influences auditory word recognition. In **Chapter 1**, I described a conflict in the past findings arising from different experimental methods. Some studies show that lexical candidates that do not fit with the syntactic context are still activated in response to auditory input, and some studies show that this kind of activation is prevented. I argued that investigating the potential syntactic constraint on word recognition requires an understanding of the mechanism for the constraint: does it function by facilitating good contextual fits or inhibiting bad ones? Studies in this area are generally not designed to be able to detect a constraint that operates by facilitating good contextual fits, and instead seek evidence that the activation of bad contextual fits has been prevented.

In **Chapter 2**, we explored one factor potentially contributing to the conflict observed in the literature, using an existing computational model of auditory word recognition. We could not examine contextual constraints directly, but as a starting point we simulated how quickly a change in lexical activation due to perceiving a new phoneme would lead to observable changes in a behavioral measure, when that behavioral measure allowed only a small set of response options (as in the visual world paradigm) or a large set of response options (as in gating). We theorized that if the size of the set of response options mattered for observing activation changes in response to a new phoneme, it might also matter for observing activation changes in response to syntactic category information. We found that the nature of the set of

response options can indeed have a large influence in simulations, not because of its size per se, but because smaller sets tend to have a skewed composition. These simulations illustrated the importance of accounting for the composition of the set of response options when making inferences about lexical activation from behavioral measures, and demonstrated how small-set measures can show earlier effects.

In **Chapter 3**, we presented a study on syntactic context in the visual world paradigm. Experiment 1 was designed to yield different predicted outcomes under facilitatory and inhibitory syntactic constraints on word recognition. We used insights from the simulations in **Chapter 2** to make sure that if wrong-category lexical competition were occurring, it would be detectable in fixation proportions, unlike in some previous studies in which we suspected that the activation of wrong-category lexical candidates had been obscured. Our study provided evidence that wordforms that can only be used as nouns are still activated when they are consistent with the auditory input in a verb-constraining context, a finding that is inconsistent with an inhibitory constraint. We did not, however, find positive evidence in favor of the facilitatory constraint.

In **Chapter 4**, we described an alternative approach to investigating the activation of lexical candidates, using neural effects of cohort entropy and phoneme surprisal. These are information-theoretic quantities that reflect, at each phoneme in a word, properties of the probability distribution of lexical candidates competing in response to auditory input. In Experiment 2, we established their effects in a MEG study of auditorily presented single words analyzed with temporal response functions. This type of analysis more accurately models the speech input than have previous

methods, and prior to our study had only been applied to continuous speech. We found effects of surprisal but not entropy, for non-word-initial phonemes, consistent with the small previous literature. Our findings bolster progress toward phoneme surprisal and cohort entropy as stable neural measures of lexical processing. Also in this chapter, we presented the design for Experiment 3, for which data could not be collected due to the COVID-19 pandemic. The goal for this study is to disambiguate the outcome of Experiment 1 and provide evidence for or against the facilitatory syntactic constraint. We intend to present structurally predictable sentences in which the syntactic category of upcoming words is always known, but content words are lexically unpredictable. We have calculated variants of entropy and surprisal that reflect a lexicon under the influence of a facilitatory or inhibitory syntactic constraint. Target nouns and verbs will be selected such that the facilitatory, inhibitory, and unconstrained variables are substantially de-correlated. We can then test which of these variables best predicts the neural response, and when.

Understanding top-down effects of syntactic category expectations on auditory word recognition will also require more precise knowledge of the nature of syntactic prediction. Therefore, in **Chapter 5**, we sought a reliable neural index of syntactic prediction, aiming eventually to examine both the cost of prediction and its impact on word recognition in the same sentence. In Experiment 4, we used MEG to test for differences in the neural response to visually presented determiner phrases (DPs) when participants knew that these DPs were the subject of a sentence or the first item in a list. We avoided lexical predictability in our stimuli so that any differences could more likely be attributed to syntactic prediction. However, we did

not observe the syntactic prediction effect that we expected on the basis of prior work. Our data suggest instead the possibility that syntactic prediction co-occurs with or is related to lexical prediction in such a way that it is no longer observable when lexical predictability is removed.

6.2 Conclusions

We do not yet have a full mechanistic account of how syntactic context influences auditory word recognition. In fact, I believe one of the important outcomes of the research undertaken in this dissertation is a better understanding of the true complexity of constructing such an account, and of what is still unknown. However, this dissertation has taken several steps toward the intended model. I have argued that the body of available evidence now points away from an inhibitory syntactic constraint that can prevent the activation of lexical candidates. We have further developed a promising approach to studying the phoneme-by-phoneme dynamics of lexical activation with neural data. We have also observed a new correspondence between lexical and syntactic prediction that underscores the richness of the problem initially posed. Along the way, I have argued for the importance of more detailed linking hypotheses between what we aim to measure and the data we have access to, including the influence of the experimental task and context. I have also prioritized the thorough accounting of patterns in previous literature. These two endeavors are closely linked, and are central to the contributions of this dissertation.

6.3 Additional questions raised

In each of the previous chapters, we have discussed connections to related questions and problems, and ideas for future work. In this section, I briefly raise

considerations that go beyond the concerns of any single preceding chapter. These considerations are related to (1) integration of other kinds of top-down information within the sentence, (2) context effects that go beyond the level of the sentence, or (3) uncertainties about the top-down or bottom-up information sources themselves.

First, the idea that a listener processing auditory input would not consider ungrammatical options is familiar from other areas of linguistics. In parsing, we don't generally assume that comprehenders are considering ungrammatical syntactic continuations. Why, then, in word recognition do we entertain the possibility that auditory input might activate lexical candidates that are incompatible with available syntactic information about the sentence being heard? One reason is that the determination of what an appropriate candidate is comes from two different systems. It might not be possible to integrate the two information sources quickly enough, or at all. Another reason is that syntactic category may not be represented on lexical items in the way that would be necessary to enable the kind of restriction we have been considering. As described in the **Introduction**, syntactic category, under Distributed Morphology, is not a stored feature but arises from the combination of category-less roots and categorizing affixes. It is not necessarily surprising, under this view, if category restrictions cannot be used to block consideration of candidates, as on some level these candidates are not inherently in conflict with the context. These kinds of issues are present for the narrow problem of syntax and then become amplified with every additional source of information within a sentence.

For example, the semantic context of the sentence should lead to specific wordform predictions that may or may not be compatible with the set of wordforms

compatible with the auditory input and the set of wordforms compatible with the syntactic context. How and when is this information integrated? What if the semantic context not only makes it likely that words related to, e.g., cats, will appear in the sentence but yields a very strong expectation for just a single word? This is both a syntactic and a semantic constraint. We did not see evidence in this dissertation that the activation of a contextually incompatible lexical candidate can be prevented. However, we should consider whether such prevention can occur in the types of sentences with high cloze probability completions used in many N400 studies. Experiment 4, in **Chapter 5**, has suggested a possible dependency between lexical-semantic prediction and prediction of syntactic structure

Second, there are potentially higher-level modulators of the way top-down and bottom-up information are used. The relative weight of each constraint could shift according to the reliability of different information sources, e.g., when speech input is noisy. Many have wondered whether the visual context of the visual world paradigm changes the normal course of lexical activation; the MEG measures of cohort competition described in **Chapter 4**, in combination with a visual world design, could potentially help us answer that question.

Finally, many questions remain about the information sources that feed top-down/bottom-up integration. Does the interaction of auditory and syntactic input occur when auditory input has been processed to the level of the phoneme, or earlier? In **Chapter 5**, we pursued a neural index of structural prediction to eventually enable more targeted questions about what exactly these predictions consist of. Are structural predictions and ensuing category expectations non-deterministic? Could category

expectations actually be captured by word-to-word co-occurrence statistics rather than abstract structure predictions? We can study these issues from the vantage of both the generation of the expectation and its impact, and the answers to these questions will jointly inform both syntactic parsing models and our understanding of word recognition.

Appendix A: Stimuli

A.1 Experiment 1

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 flowers.	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	They remembered the jellyfish that often	jellyfish	pineapple	eyebrow	motorcycle

	appeared in the summer.	appeared in the summer.				
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

A.2 Experiment 2

A.2.1 Targets

abroad	arbitrary	beta	bust	chart	consist
absolve	are	beverage	cab	chest	consolidate
abstain	aroma	bias	cadet	chic	construct
abstract	arrange	bid	caliber	chime	consume
absurd	arrive	binge	calorie	chimney	contain
accelerate	asphalt	bite	cancel	choir	contrast
accent	aspire	bitter	canoe	chomp	convey
achieve	aspirin	blatant	capacity	chore	convince
actual	associate	bleach	cardiac	circa	cough
adept	assume	bless	care	circle	cower
admit	athlete	blight	caribou	circus	crank
advance	atom	blip	caricature	citadel	crate
affair	attack	blitz	carnival	claim	crop
affiliate	autumn	blizzard	carton	clean	crowd
affinity	avail	blubber	cartoon	clear	crumb
affirm	avalanche	bluff	casket	clever	crusade
afraid	average	bolster	casual	clock	crux
agitate	awe	bore	catapult	clone	culinary
aisle	awkward	born	catastrophe	clove	culprit
alarm	baboon	botch	catch	coarse	cup
alcohol	bait	bother	cater	collapse	curd
ale	bake	bounty	cauldron	come	curse
algorithm	balcony	bovine	cavalry	commit	cushion
aloof	banquet	braid	ceiling	commute	custody
altar	bare	bran	celebrate	compete	custom
alumni	bark	brave	celery	compound	cycle
ambiguous	barnacle	breach	cemetery	conceal	cylinder
analog	barren	break	census	concern	dab
annoy	basket	brim	ceremony	condition	date
annual	batch	brittle	chagrin	confess	daughter
antique	bawl	broke	challenge	congratulate	decent
anvil	beep	brow	chameleon	congress	decree
apathy	behave	bruise	chant	conjure	dedicate
appeal	bend	bump	character	connect	defer
appreciate	bent	bundle	charity	consensus	defy
approximate	berserk	burst	charm	conserve	deity

democrat
denounce
deny
desert
desk
destroy
detain
develop
device
devise
diagram
dialogue
diaphragm
dice
dictate
did
digest
digress
diligent
dimple
din
dingy
diploma
direct
dish
disrupt
distant
distinct
document
dolt
domain
door
dozen
drab
drastic
dribble
drift
droop
drug
duet
duration
eager
ease
easel
eclectic
eight
elastic
electric
element
elevate
eliminate
eloquent
else
eminent
emphasis
empty
enamel
entertain
enzyme
equal
equity
eradicate
erase
errand
escort
essay
essence
eternal
evade
evolve
exam
example
exceed
excel
exempt
exert
expanse
expect
expense
expire
explicit
explode
exquisite
extinct
exuberant
exude
fable
fail
farce
fare
fast
fathom
fax
feat
fed
fee
feed
femur
ferocious
fervor
fest
feud
fiction
fierce
fig
file
fill
finale
finger
firm
fissure
five
flank
flog
fluid
flurry
flush
focal
fog
font
foray
forget
fortuitous
four
frantic
fraught
fray
fringe
fritter
fuel
fungus
furtive
fuse
fuss
gain
gallon
galore
garage
gargle
garment
gas
gear
geese
generous
gentle
genuine
giddy
gift
gimmick
gist
glaze
glimmer
glimpse
glitter
gloom
glue
gone
gorge
gossip
gouge
gown
grab
graduate
grail
gram
grammar
grape
grass
grate
grease
grew
grim
grip
grit
groove
ground
group
gruff
gurgle
gust
halibut
happy
hard
harp
harsh
hash
haunt
helium
hemp
hence
heron
hid
hideous
hit
hoarse
hog
hold
hole
horrid
horror
hostile
how
huddle
huge
human
humid
humility
hunch
hurricane
igloo
ignore
illuminate
image
imitate
immense
imperial
implement
imply
incessant
incinerate
include
incumbent
indict
indignant
induce
industry
inept
inferior
inflict
initial
inject
injury
ink
inning
instinct
interest
interfere
interlude
intricate
introduce

invite
iodine
iota
jacket
jade
jam
jargon
jaw
jest
join
jubilee
juncture
kennel
kernel
knee
knife
lab
language
lapse
lark
laser
late
latitude
laugh
launch
lavish
lawn
lax
leaf
leap
learn
lease
length
lesion
level
license
lid
lift
lilac
linoleum
lit
live
locate
loose
lose
loud
loyal
lumbar
mahogany
maintain
male
mammoth
manifest
map
marine
mast
mat
match
mayhem
meager
measure
meat
mellow
melt
menu
merry
mesh
message
metal
methane
microbe
middle
mile
milk
million
minimum
mint
mirth
mist
moat
modern
mogul
moist
molecule
monarch
mountain
mug
mule
multiple
mumble
mundane
muster
myriad
nail
naked
nap
neat
needle
nephew
net
nickel
nine
node
nominate
null
nun
nutrient
oasis
observe
obtain
obvious
occupy
octave
odd
offend
offer
onion
operate
opportune
optimum
orchid
ordeal
organize
ornery
orthodox
ounce
ovation
owl
pact
pad
paddle
paint
pair
palate
pale
paltry
pamper
pamphlet
panic
parade
paradigm
parakeet
parasol
parent
paste
pattern
pecan
pedestrian
pencil
perform
peril
perk
persist
petal
petty
phantom
pheasant
phosphorus
phrase
picket
pie
piece
pile
pinch
pioneer
pistol
piston
pitch
pity
plain
plant
platform
platter
please
pledge
plethora
plight
plot
plough
plunge
polymer
poor
portion
posse
posture
poultry
precinct
precious
predicament
premise
prerogative
present
pretend
pretty
pristine
problem
prod
produce
progress
prom
property
prophet
protect
proud
province
proxy
prune
pry
publish
pulp
pulse
pungent
punish
purchase
purge
putt
puzzle
quaint
quarantine
quarrel
quarry
quartet
quick
quilt
quote
race
radius
rag
rage
rain
raise
rally
ran
rang
rapt
rash
rate
rattle
raw
razor

reap
rear
reduce
reef
reflex
register
relax
relic
relieve
remedy
renaissance
report
reserve
reside
resilient
resplendent
rest
result
ride
rigid
rind
rinse
riot
ripe
ripple
risk
rival
roar
roast
rocket
rodent
role
roost
rope
roster
rotate
rove
rubble
ruin
rummage
run
rupture
salad
salary
sane
satin
sausage
scale
scalpel
scan
scare
scene
scoff
scrape
scrawny
screen
screw
scrounge
scuffle
sculpt
secret
sect
sediment
seek
seem
segment
seldom
self
seminar
sermon
serve
settle
severe
shabby
shake
shall
shame
shard
share
shawl
shear
shed
shift
shine
shirt
shoddy
should
shove
shrewd
shrimp
shroud
shudder
shuttle
shy
sigh
sign
similar
simmer
sinus
sit
site
sleek
sleeve
slender
slide
slight
slim
sling
slip
slit
sliver
slogan
sloth
slug
slurp
slush
sly
smack
smash
smear
smooth
smuggle
snack
snarl
snip
soak
soap
soft
solid
soliloquy
somber
soup
soy
span
spare
sparse
spectacle
spend
spigot
spin
splice
sprain
sprinkle
spur
spurt
squash
stab
staff
stake
stale
stalk
stamp
staple
starve
stash
staunch
steal
stellar
stink
stomach
stooge
store
straddle
strain
stress
strict
stroll
stub
stubborn
study
stuff
stutter
substance
succulent
sulk
sullen
summer
supplement
supply
suppress
surge
suspect
suspend
swamp
swear
sweat
sweep
swim
swing
switch
tactic
tail
taint
talent
tame
tangle
tar
taut
tax
tease
temperature
ten
tend
tense
tent
tenuous
terrain
theory
thigh
three
thrive
throttle
thrust
thyroid
tick
tile
timid
tire
tissue
toffee
token
ton
torment
torrent
tradition
traipse
trait
tram
trance
trap
treason
trespass
tribe
trick
trickle
trim
troop
trophy
trot
trough

truck
tunic
turbine
turmoil
turpentine
turquoise
twelve
twig
twill
ugly
umbrella
umpire
understand
vain
valet
valid
vanilla
vanish
vary
vast
vector
vegetable
vein
velocity
venture
venue
veranda
verdict
vertical
vet
vigor
vile
villain
vintage
violin
vow
wad
waltz
wand
wander
want
wary
wave
wax
weary
wedge
weep
weigh
wet
whine
whisk
whisper
wide
wilt
win
worship
wrap
wrath
wreak
wreath
wrinkle
wrist
yield
zest

A.2.2 Probes

Item	Probe	number in list	distance to next probe	match
organize	coordinate	21	19	1
foray	exterior	41	20	0
beverage	conceive	47	6	0
rally	threat	59	12	0
posture	position	78	19	1
coarse	rough	97	19	1
mile	rhinoceros	105	8	0
chime	bell	121	16	1
care	aid	133	12	1
proxy	lug	134	1	0
lilac	lavender	146	12	1
brittle	ambient	152	6	0
athlete	hut	162	10	0
opportune	concoct	169	7	0
offer	volunteer	183	14	1
sloth	tardy	191	8	0
construct	build	202	11	1
stooge	gallery	221	19	0
clear	open	241	20	1
live	survive	246	5	1
valet	elegant	250	4	0
palate	accept	264	14	0
charity	love	278	14	1
publish	print	281	3	1
crank	grouch	282	1	1
mint	pursue	289	7	0
pitch	throw	290	1	1
thrust	sad	299	9	0
pamper	cyst	306	7	0
blatant	worry	320	14	0
bust	raid	324	4	1
absolve	varmint	339	15	0
humid	weather	345	6	1
conjure	evoke	356	11	1
garment	clothe	374	18	1

latitude	debacle	394	20	0
eminent	people	395	1	1
rest	aunt	414	19	0
wilt	limp	426	12	1
afraid	fear	438	12	1
inning	recess	446	8	0
soap	rub	449	3	1
gas	note	461	12	0
asphalt	pave	480	19	1
choir	saloon	489	9	0
vet	farm	500	11	1
melt	pedestal	515	15	0
harsh	dunce	525	10	0
drug	medicine	538	13	1
cancel	grave	540	2	0
essay	literary	543	3	1
avail	cause	552	9	0
cushion	vague	566	14	0
turmoil	chamber	581	15	0
blubber	fat	586	5	1
rash	red	604	18	1
vintage	stubble	608	4	0
ran	away	628	20	1
affirm	lay	630	2	0
parasol	shade	639	9	1
tissue	cuss	659	20	0
stalk	affect	664	5	0
moist	damp	674	10	1
scan	read	677	3	1
vein	blood	689	12	1
lid	tartar	699	10	0
trick	joke	709	10	1
alumni	quota	722	13	0
consist	slate	732	10	0
smuggle	import	745	13	1
linoleum	figure	753	8	0
stellar	star	759	6	1
pie	petition	779	20	0

diagram	explain	784	5	1
lab	science	785	1	1
annual	slump	801	16	0
collapse	cheap	814	13	0
affair	muzzle	815	1	0
ounce	pound	821	6	1
hoarse	cry	824	3	1
element	product	832	8	0
shake	tremble	837	5	1
enzyme	protein	853	16	1
lark	podium	857	4	0
pretty	delicate	871	14	1
scuffle	brake	879	8	0
ton	magnet	884	5	0
huge	check	897	13	0
equal	peer	909	12	1
wreak	fugitive	918	9	0
digest	truce	926	8	0
cavalry	horse	929	3	1
din	lame	936	7	0
intricate	elaborate	945	9	1
onion	flavor	957	12	1
eradicate	kill	972	15	1
distinct	discrete	989	17	1

A.3 Experiment 4

Note: Each pair of DP's was presented once for each participant, and whether it occurred as a sentence, a coordinated phrase, or a list of DP's was counter-balanced across participants. Here we show the sentence version for each pair. The verb was replaced with "and" in the coordination condition or a blank screen in the DP list condition.

the	**ADJ**	**N**		**the**	**ADJ**	**N**
Set 1						
the	analytic	companies	planned	the	satirical	funerals
the	icy	spacecrafts	measured	the	silky	wolves
the	wicked	prisons	buried	the	green	officials
the	deadly	parasites	scavenged	the	atrocious	pizzas
the	controversial	lamps	delivered	the	nimble	servants
the	belligerent	vendors	described	the	pungent	cardigans
the	bubbly	merchants	visited	the	ferocious	brides
the	artful	balloons	eliminated	the	decent	technicians
the	merciful	boys	orbited	the	forlorn	employees
the	apathetic	prisoners	tested	the	honest	patients
the	scrappy	workers	searched	the	skittery	couriers
the	mysterious	tornados	conquered	the	frank	missionaries
the	verbose	wizards	trampled	the	opulent	baths
the	merciless	repairmen	deciphered	the	mediocre	squares
the	silly	giants	negotiated	the	jubilant	weddings
the	sick	zombies	espoused	the	communal	robots
the	powerful	sentries	baffled	the	kind	journalists
the	warm	disciples	defeated	the	hospitable	operators
the	chipper	pilots	carried	the	steady	tables
the	adept	sailors	perused	the	blurry	journals

the	juvenile	dictators	selected	the	coherent	anthropologists
the	soggy	lawyers	sent	the	woodsy	midwives
the	fatal	leggings	dismantled	the	octogonal	bridges
the	pure	pencils	captivated	the	corporate	bakers
the	stereotypical	rectangles	cleared	the	moldy	couches
the	fanatical	players	fought	the	plush	penguins
the	ceramic	tails	requested	the	sparkly	hooligans
the	fleshy	soldiers	corrupted	the	pale	bikes
the	eclectic	sunsets	appreciated	the	shrunken	politicians
the	masterful	entrances	bought	the	malicious	mosquitos
the	excitable	managers	feared	the	useless	tycoons
the	superb	robes	detected	the	pliable	butchers
the	algebraic	games	approved	the	formal	criminals
the	strong	sages	oppressed	the	transient	clerks
the	ornate	ladders	knitted	the	powdery	manifestos
the	fierce	colors	imprisoned	the	angry	blacksmiths
the	devilish	artists	defused	the	persistent	dilemmas
the	appropriate	lies	brushed	the	fiendish	bishops
the	wealthy	insurgents	stung	the	lovable	ambassadors
the	barbaric	puzzles	bypassed	the	quiet	tourists
the	prime	goalies	protected	the	gradual	discoveries
the	stylish	chocolates	understood	the	lacy	buildings
the	blunt	attackers	brought	the	sassy	villains
the	exquisite	kegs	ate	the	risky	crackers
the	bitter	orators	hated	the	foundational	computers
the	magic	rivers	felt	the	printable	rulers
the	nasal	witches	sang	the	squeamish	answers
the	dingy	demons	watched	the	alert	wardens
the	buff	heroes	remembered	the	tranquil	acrobats
the	herbal	winds	fetched	the	prominent	horses
the	quirky	referees	recited	the	cultish	catastrophes

the	springy	bees	plastered	the	convenient	volcanoes
the	smooth	billboards	forged	the	continuous	boulders
the	fastest	biologists	called	the	meddlesome	babies
the	familiar	beaches	improved	the	remorseful	dwarves
the	splashiest	medics	broke	the	squeezable	cups
the	negotiable	magicians	released	the	ornery	officers
the	thrifty	trainers	dried	the	grateful	novices
the	worthy	spiders	repaired	the	natural	titans
the	major	towns	imagined	the	pricey	kidnappers
the	toasty	tractors	entered	the	scenic	cathedrals
the	empty	hotels	changed	the	sterile	businesses
the	ticklish	architects	organized	the	rude	families
the	seasonal	kings	answered	the	accidental	sergeants
the	feminist	otters	resisted	the	wordy	inmates
the	malevolent	goats	crushed	the	poetic	athletes
the	public	pans	defended	the	frugal	clerics
the	fanciful	beetles	hid	the	comedic	seamstresses
the	violent	novels	tasted	the	crinkly	cheeses
the	demonic	chefs	covered	the	academic	fencers
the	lawyerly	questions	served	the	benign	sculptors
the	readable	bubbles	destroyed	the	structural	diamonds
the	egregious	feet	revealed	the	flirtatious	nuns
the	slanderous	armies	hoarded	the	informational	smiles
the	mindful	books	confused	the	sincere	editors
the	wily	scissors	trapped	the	artistic	stylists
the	preppy	priests	fixed	the	rough	newspapers
the	additional	bones	melted	the	billowy	boxers
the	wet	floors	taught	the	tropical	brats
the	respectful	chiefs	solved	the	atmospheric	earthquakes
Set 2						
the	angsty	birds	planned	the	lavish	candles

the	breathless	buzzards	measured	the	synthetic	guns
the	simple	nurses	buried	the	charitable	moms
the	questionable	enemies	scavenged	the	glittery	fishermen
the	spotty	recessions	delivered	the	abrasive	sandals
the	pointless	barbers	described	the	mature	attornies
the	pricey	calculators	visited	the	huge	murderers
the	forceful	headaches	eliminated	the	malleable	governors
the	clumpy	beets	orbited	the	negligible	presidents
the	ghastly	words	tested	the	euphoric	doctors
the	decrepit	sandwiches	searched	the	new	agencies
the	critical	conductors	conquered	the	regal	bushes
the	blind	monks	trampled	the	extendable	tulips
the	terrific	hoses	deciphered	the	promotional	sketches
the	photogenic	nerds	negotiated	the	blatant	strategies
the	moist	engines	espoused	the	buttery	mittens
the	impactful	hosts	baffled	the	stout	mothers
the	groovy	students	defeated	the	colorful	hamsters
the	squirrelly	witnesses	carried	the	tolerable	geniuses
the	idiotic	ostriches	perused	the	changeable	campaigns
the	groggy	mobs	selected	the	inflatable	whales
the	reptilian	frames	sent	the	admirable	professors
the	drab	gifts	dismantled	the	architectural	cans
the	blubbery	minotaurs	captivated	the	symmetrical	trainees
the	tiny	events	cleared	the	mousy	gardeners
the	incredible	trains	fought	the	stretchable	castles
the	ambivalent	butterflies	requested	the	fortuitous	journeys
the	giggly	maids	corrupted	the	optical	daggers
the	exhaustive	dinners	appreciated	the	defunct	tigers
the	humble	elephants	bought	the	fun	oracles
the	purposeful	ninjas	feared	the	fancy	machines
the	puffy	experts	detected	the	finicky	buyers

the	slothful	villagers	approved	the	interactive	warriors
the	vile	interrogators	oppressed	the	shameless	spectacles
the	interpretive	astrologers	knitted	the	autonomous	cigarettes
the	melancholic	militias	imprisoned	the	innate	spirits
the	lockable	shingles	defused	the	greedy	prophets
the	visual	messengers	brushed	the	democratic	bears
the	competitive	dragons	stung	the	economical	cowgirls
the	flimsy	napkins	bypassed	the	rare	guests
the	caustic	blowtorches	protected	the	perceptive	cops
the	residential	investors	understood	the	temporary	toddlers
the	agile	vandals	brought	the	diminutive	mirrors
the	commendable	translators	ate	the	spidery	camels
the	primal	lights	hated	the	fruitful	idiots
the	gallant	comedians	felt	the	significant	boxes
the	stellar	choirs	sang	the	tangential	poems
the	sullen	dads	watched	the	slimy	consumers
the	dense	odors	remembered	the	ropy	mutants
the	collapsible	busboys	fetched	the	creamy	mushrooms
the	administrative	factories	recited	the	respectable	therapies
the	desirous	planets	plastered	the	moody	mice
the	tiresome	apologies	forged	the	splendid	planes
the	weak	scribes	called	the	sickly	therapists
the	monthly	commanders	improved	the	catastrophic	bracelets
the	spineless	treasuries	broke	the	extreme	ponds
the	proverbial	actors	released	the	perceivable	teachers
the	basic	citizens	dried	the	inferior	aliens
the	abhorrent	daughters	repaired	the	exultant	androids
the	quaint	committees	imagined	the	serpentine	explorers
the	tidy	nephews	entered	the	horrible	fights
the	intentional	samurai	changed	the	smart	traitors
the	oceanic	umpires	organized	the	shaky	buckles

the	fervent	tyrants	answered	the	vehement	musicians
the	auspicious	churches	resisted	the	chronic	photographers
the	deepest	arguments	crushed	the	maniacal	monsters
the	stiff	warehouses	defended	the	tame	casinos
the	jumpy	bartenders	hid	the	diabetic	princes
the	feathery	monuments	tasted	the	gigantic	potatoes
the	loose	raccoons	covered	the	contemplative	cooks
the	complacent	beginnings	served	the	delicious	flowers
the	full	experiments	destroyed	the	mindless	memories
the	paternal	barons	revealed	the	worthless	movies
the	adjustable	websites	hoarded	the	indulgent	jackets
the	inspirational	ghosts	confused	the	bouncy	assassins
the	pensive	deputies	trapped	the	feeble	spoons
the	dry	keyboards	fixed	the	resolute	vampires
the	generous	boats	melted	the	pushy	generals
the	portable	breads	taught	the	manageable	bells
the	talkative	clowns	solved	the	ubiquitous	governments
Set 3						
the	outrageous	upstarts	planned	the	hybrid	yards
the	hazy	airports	measured	the	weary	pirates
the	timid	pioneers	buried	the	smelly	heretics
the	motherly	philosophers	scavenged	the	eligible	diseases
the	regular	zebras	delivered	the	gracious	princesses
the	luxurious	cakes	described	the	whiny	kangaroos
the	austere	scholars	visited	the	sweet	photographs
the	extensive	messages	eliminated	the	bogus	stews
the	velvety	lions	orbited	the	difficult	visitors
the	decisive	representatives	tested	the	shapeless	businessmen
the	lawful	mermaids	searched	the	evil	songs
the	packable	instructors	conquered	the	noisy	plumbers
the	extra	surgeons	trampled	the	spherical	pictures

the	bridal	choices	deciphered	the	crispy	threads
the	distraught	poets	negotiated	the	secular	customers
the	harmonious	debts	espoused	the	attractive	desks
the	wobbly	frogs	baffled	the	crazy	ministers
the	environmental	laborers	defeated	the	salient	contestants
the	grand	purses	carried	the	perceptible	mayors
the	tolerant	celebrities	perused	the	haughty	institutions
the	brassy	canyons	selected	the	conservative	sheriffs
the	sulfurous	umbrellas	sent	the	weird	windows
the	molecular	stoves	dismantled	the	dangerous	mistakes
the	horizontal	eagles	captivated	the	formulaic	curators
the	weighty	pickles	cleared	the	nasty	tunnels
the	skillful	models	fought	the	computational	astronauts
the	periodic	agents	requested	the	bulky	knights
the	snuggly	requests	corrupted	the	invincible	veterans
the	selective	dealers	appreciated	the	fragile	expeditions
the	ghoulish	bosses	bought	the	rickety	turtles
the	streaky	sweaters	feared	the	lamentable	jesters
the	putrid	carpets	detected	the	easy	fries
the	amnesiac	victims	approved	the	temperamental	girls
the	freakish	ideas	oppressed	the	dumb	drivers
the	adamant	performers	knitted	the	gangly	robins
the	prim	knees	imprisoned	the	pragmatic	hikers
the	mythical	rooms	defused	the	poignant	wars
the	mystical	travelers	brushed	the	ergonomic	shirts
the	managerial	systems	stung	the	crabby	waiters
the	slovenly	days	bypassed	the	royal	sofas
the	genuine	victories	protected	the	pliant	brains
the	intermediate	kids	understood	the	solar	gnomes
the	realist	gluttons	brought	the	blobby	diplomats
the	typical	janitors	ate	the	bucolic	steaks

the	provocative	curtains	hated	the	coastal	doors
the	tentative	dogs	felt	the	gooey	chairs
the	flashy	thieves	sang	the	expendable	directions
the	methodical	rabbits	watched	the	shameful	movements
the	jealous	fairies	remembered	the	absurd	reporters
the	classy	hordes	fetched	the	compliant	dukes
the	thankful	cities	recited	the	rebellious	members
the	dynamic	leaders	plastered	the	selfless	renegades
the	awkward	houses	forged	the	worrisome	panels
the	perpendicular	rugs	called	the	domestic	heirs
the	pinkish	cabins	improved	the	gratuitous	bachelors
the	profound	pandas	broke	the	ripe	pincers
the	ironic	directors	released	the	hopeful	stains
the	solid	hyenas	dried	the	rowdy	apartments
the	delectable	problems	repaired	the	splintery	squirrels
the	primitive	husbands	imagined	the	furious	stories
the	meek	pharmacies	entered	the	godless	parties
the	juicy	sounds	changed	the	optimistic	guards
the	accurate	cashiers	organized	the	wrinkly	programmers
the	stray	parents	answered	the	scary	secretaries
the	enormous	cleaners	resisted	the	paranoid	banks
the	ambiguous	hunters	crushed	the	expressive	lovers
the	carcinogenic	farmers	defended	the	clear	scientists
the	rhythmic	remarks	hid	the	gaudy	bystanders
the	measurable	goblins	tasted	the	active	eggs
the	exotic	maps	covered	the	abusive	assistants
the	joyous	deans	served	the	diligent	judges
the	mellow	teams	destroyed	the	workable	trees
the	convertible	ogres	revealed	the	bothersome	accordions
the	hysterical	mattresses	hoarded	the	esoteric	tailors
the	careless	queens	confused	the	rosy	tribes

the	confident	owners	trapped	the	presumptuous	specialists
the	salty	fires	fixed	the	deficient	friends
the	coppery	asteroids	melted	the	proud	advisors
the	hip	misers	taught	the	amateur	sisters
the	bloody	detectives	solved	the	bountiful	mysteries

Appendix B: Second time window in Experiment 4

In this Appendix, we report and discuss the results of Experiment 4 for tests conducted in the second time window of interest. This was the 1500-3500 ms window in the trial, time-locked to the onset of the item intervening between the two noun phrases (the verb, "and", or a blank) and encompassing both this item and the second noun phrase. We do not discuss these results in **Chapter 5** because our primary question is about prediction effects, during the time window when the only difference between conditions is what can be predicted to occur next. In the analyses that we report here, either the stimulus itself is different (verb vs. "and" vs. blank) or the response is to identical noun phrases that we would expect to be processed differently because of the preceding item.

B.1 Results

B.1.1 Planned analyses

B.1.1.1 Coordination vs. phrase in Window 2

We ran a spatiotemporal cluster test over the left temporal lobe and left inferior frontal gyrus for the coordination vs. phrase contrast, in the 1500-3500 ms window starting with the presentation of either "and" or a blank screen and then encompassing the second noun phrase. This was motivated by the finding of a sustained negativity following the coordinator in Lau and Liao (2018). We found three significant clusters. Along ventral temporal lobe, we found a cluster (p = .035) showing a negative-going peak for coordination > phrase from 78-198 ms after onset of the coordinator (**Figure 28**, top). In anterior superior temporal lobe, we found a

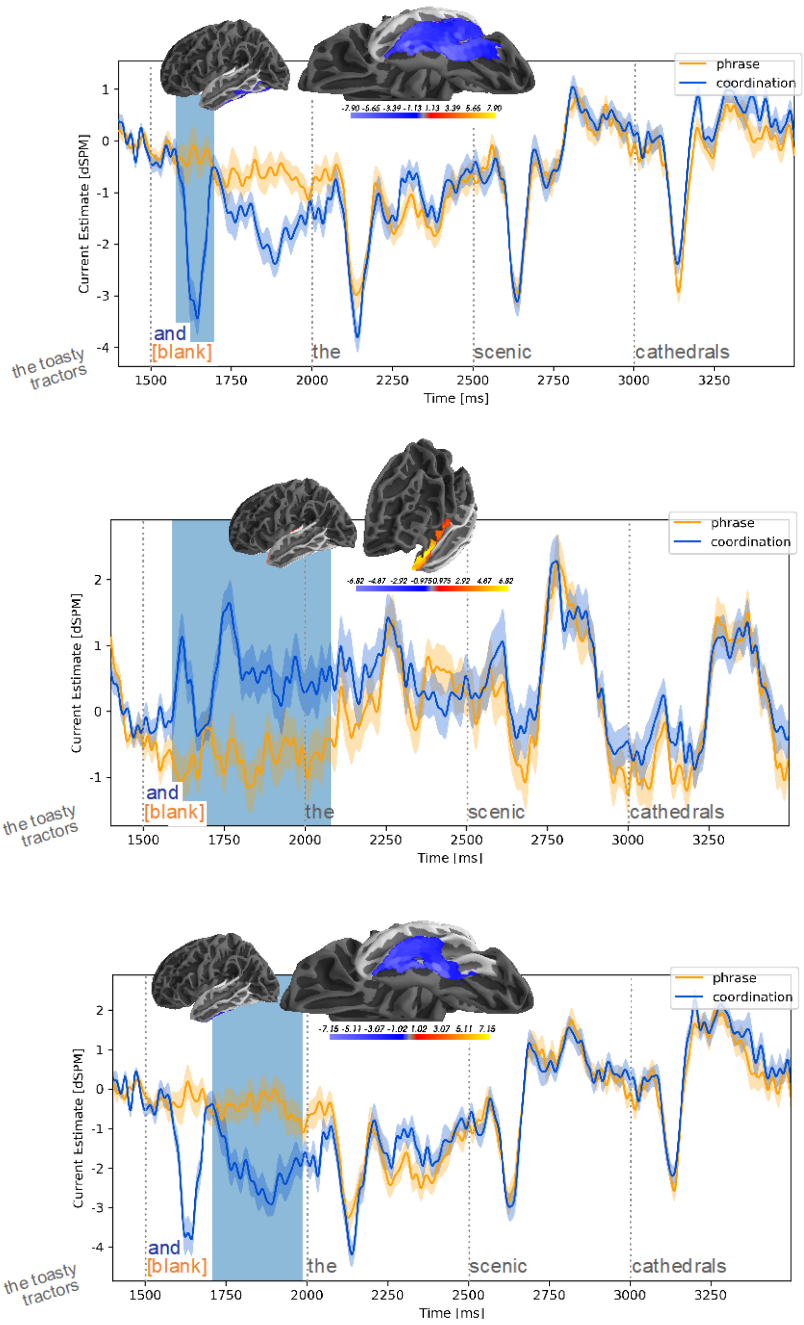


Figure 28. *Coordination vs. phrase clusters in 1500-3500 ms window, plotting time course and location of neural activity. Color bar shows maximum t-value at a source point. Example item shows onset of each word.*

cluster (p = .013) showing a positive-going peak and sustained effect of coordination > phrase from 90-582 ms after onset of the coordinator (**Figure 28**, middle). Finally, a more anterior ventral cluster (p = .024) again showed a negative-going and more sustained effect of coordination > phrase from 108-486 ms after onset of the coordinator (**Figure 28**, bottom). All three clusters reflected neural activity in response to presentation of "and" relative to presentation of a blank screen. We failed to observe the effect of coordination during the second noun phrase that we had expected on the basis of Lau and Liao (2018).

B.1.1.2 Sentence vs. phrase in Window 2

We had also planned an exploratory spatiotemporal cluster test for the sentence vs. phrase contrast in the 1500-3500 ms window following the onset of the verb/blank. This yielded four significant clusters, all in response to the verb/blank. The first three show nearly identical patterns of activity along TTS (112-524 ms; p = .002) (**Figure 29**, top), pSTS (156-532 ms; p = .004) (**Figure 29**, middle) and STS (164-522, p = .002) (**Figure 29**, bottom). All three show a double peak for sentence > phrase, with the first peak appearing to occur from roughly 150-250 ms, and the second from roughly 300-500 ms, with positive polarity for the pSTS cluster and negative polarity for the TTS and STS clusters. The fourth cluster (p = .018) shows a positive-going peak for sentence > phrase, from 226-518 ms after the onset of the verb/blank, in anterior temporal lobe (**Figure 30**).

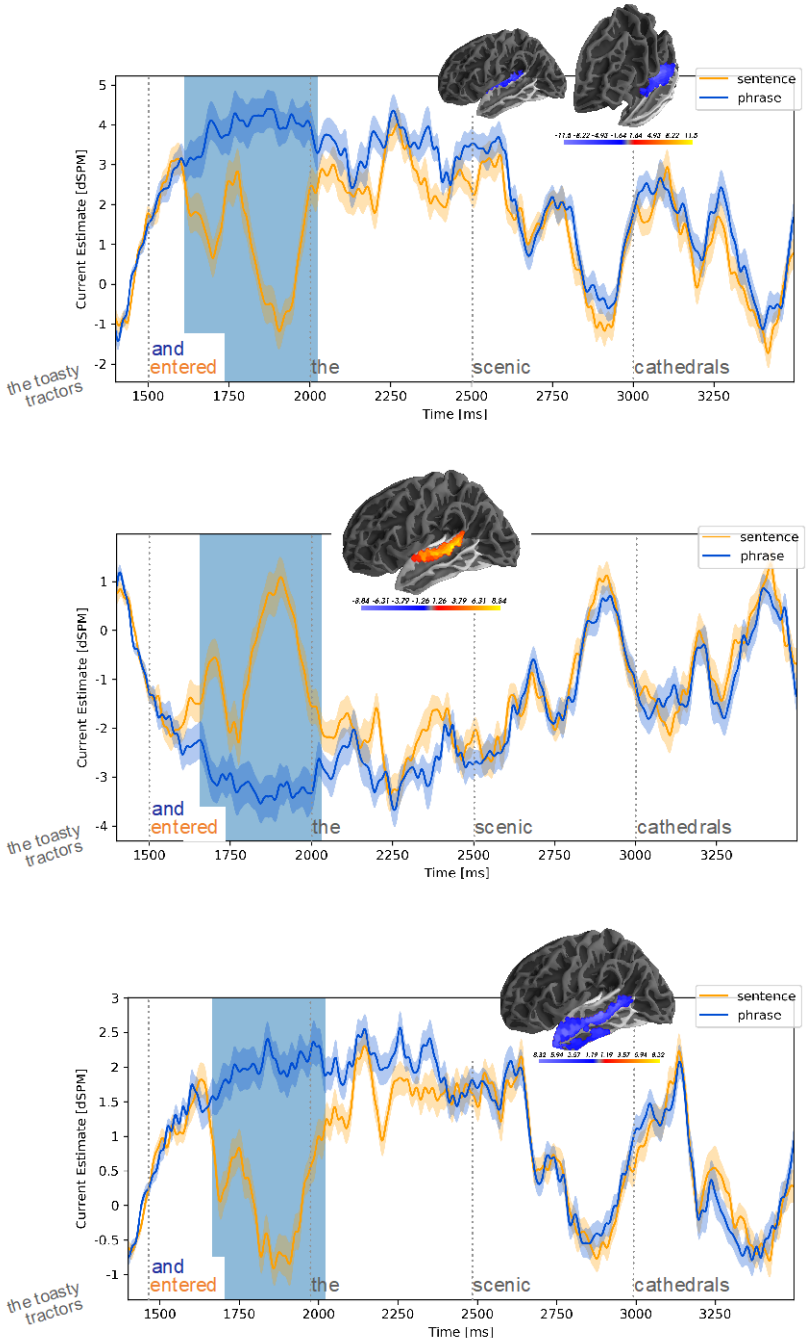


Figure 29. *Sentence vs. phrase clusters (TTS, pSTS, STS) in 1500-3500 ms window, plotting time course and location of neural activity. Color bar shows maximum t-value at a source point. Example item shows onset of each word.*

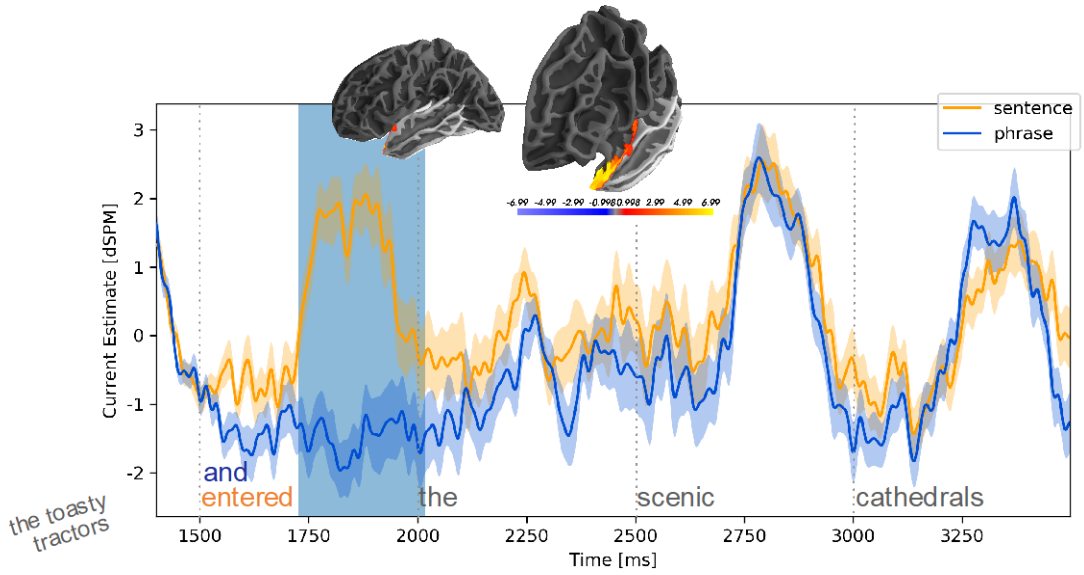


Figure 30. *Sentence vs. phrase cluster (anterior temporal lobe) in 1500-3500 ms window, plotting time course and location of neural activity. Color bar shows maximum t-value at a source point. Example item shows onset of each word.*

B.1.2 Data-driven exploratory analysis of sentence vs. coordination

Once it became clear that there were no effects in our dataset plausibly related to syntactic prediction, we decided to more thoroughly explore the unexpected effects that had occurred in the second window. Specifically, the significant clusters of activity we observed in response to the verb or "and" occurred on peaks that, in the time course of the full trial, appeared to recur either for all words or for open-class words only. To test this more directly, we ran spatiotemporal cluster tests directly comparing the sentence and coordination conditions. We caution that findings from these exploratory analyses should be considered preliminary, and would require replication for more serious consideration.

In the 1500 ms window of the first noun phrase we found no significant clusters for the sentence/coordination contrast. In the 2000 ms window beginning with the onset of the verb or "and" we found multiple significant clusters. In plotting these effects, though they are elicited by the sentence/coordination contrast, we also plot the phrase condition for reference.

In inferior and ventral temporal areas, we found a cluster (p = .000) from 138-530 ms showing sustained positive-going activity for the sentence condition and negative-going activity for the coordination condition **(Figure 31**). This negative-going activity matched a cluster we had observed from the coordination/phrase contrast. There was no such cluster for the sentence/phrase contrast. Plotting activity for this cluster for all three conditions shows the phrase condition intermediate between the sentence and coordination conditions.

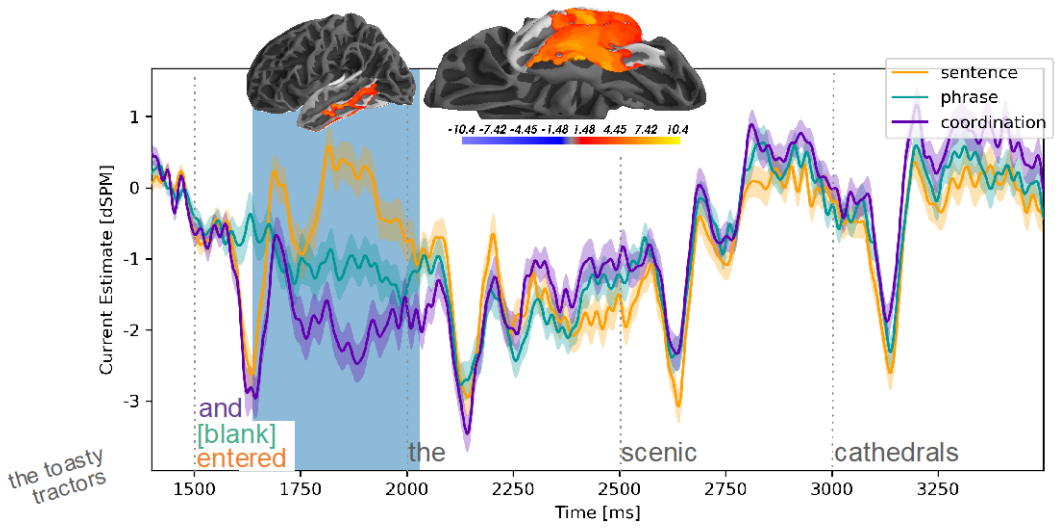


Figure 31. *Sentence vs. coordination cluster (inferior and ventral temporal lobe) in 1500-3500 ms window, plotting time course and location of neural activity. Phrase condition is plotted for reference. Color bar shows maximum t-value at a source point. Example item shows onset of each word.*

In striping activity along TTS, STS, and ITS, we found a cluster (p = .000) from 146-572 ms showing the same negative-going, apparently two-peaked effect that we observed in the sentence/phrase contrast but not the coordination/phrase contrast (**Figure 32**, top).

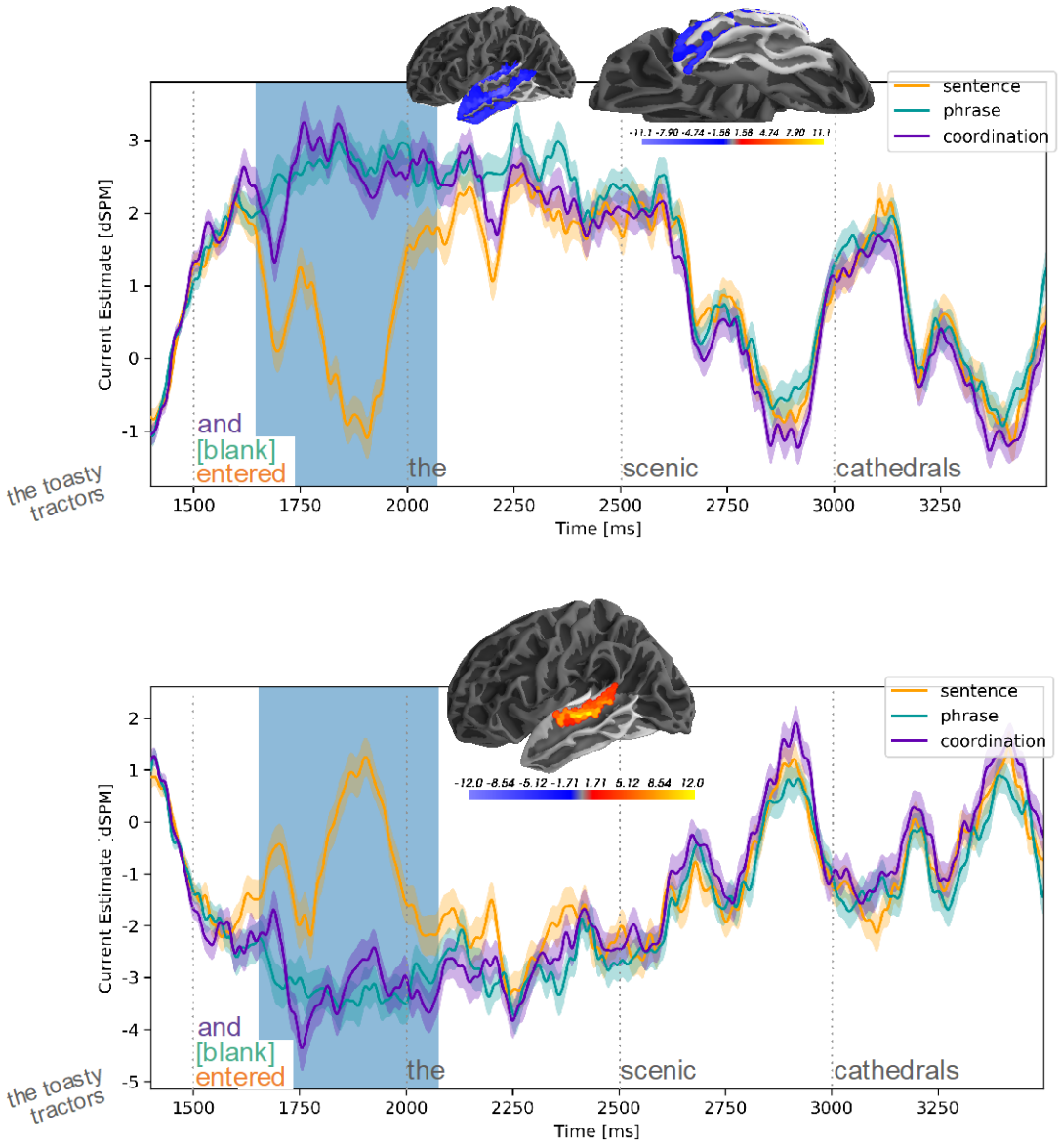


Figure 32. *Sentence vs. coordination clusters (TTS, pSTS, STS, ITS) in 1500-3500 ms window, plotting time course and location of neural activity. Phrase condition is plotted for reference. Color bar shows maximum t-value at a source point. Example item shows onset of each word.*

In pSTS, we found a cluster (p = .001) from 154-576 ms, mirroring the cluster along TTS, STS, and ITS (**Figure 32**, bottom). Plotting activity for these clusters for all 3 conditions, the coordination condition largely patterns with the phrase condition, with a small tendency in the direction of the sentence condition pattern.

Finally, a cluster in IFG (p = .036), from 252-476 ms, showed a positive peak for the coordination condition and a mirrored negative peak for the sentence condition (**Figure 33**). Plotting activity for this cluster for all three conditions shows the coordination condition patterning with the phrase condition. We had not found significant clusters for the sentence/phrase or coordination/phrase contrasts in IFG.

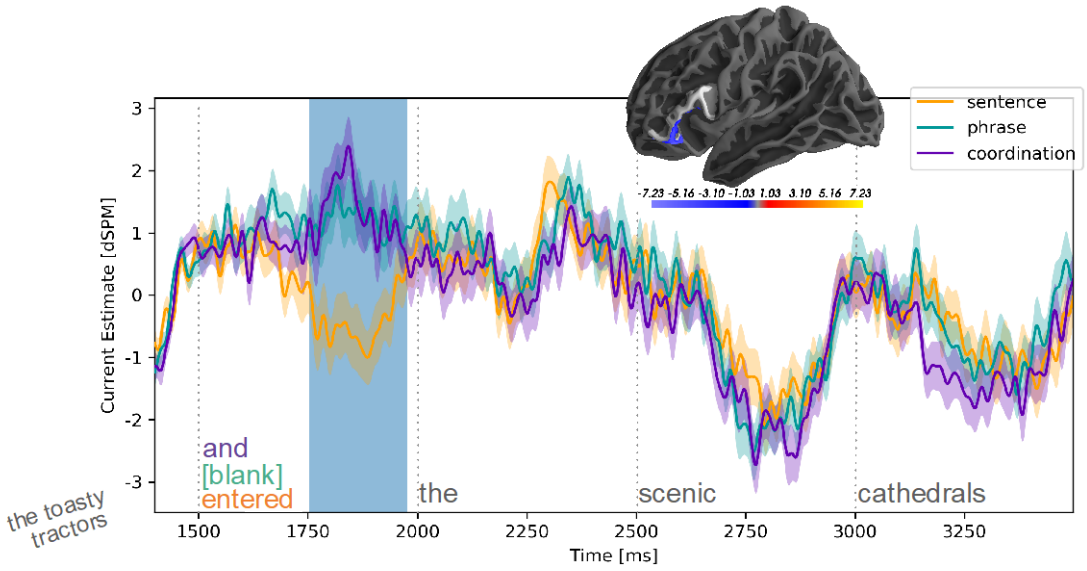


Figure 33. *Sentence vs. coordination cluster (IFG) in 1500-3500 ms window, plotting time course and location of neural activity. Phrase condition is plotted for reference. Color bar shows maximum t-value at a source point. Example item shows onset of each word.*

B.2 Discussion

In **Chapter 5** we focused on differences in comprehension of the subject noun phrase that might occur because of the anticipation of an upcoming verb or coordinator. In this Appendix, we examine the processing of the verb or coordinator itself, and the noun phrase that follows it. These analyses consisted of spatiotemporal cluster tests in a 2000 ms window that started with the verb, coordinator, or blank, and then encompassed that item and the ensuing noun phrase. We looked at all three pairwise comparisons between the conditions, and found that all significant differences occurred in the roughly 500 ms window following the onset of the verb or coordinator. We found no differences in the processing of the noun phrase when it was the object of a sentence vs. the second item in a coordinated pair vs. the second item in a list.

B.2.1 Lack of difference in the second noun phrase

B.2.1.1 Coordination vs. phrase

The lack of difference between conditions during the second noun phrase was, as for the first noun phrase, unexpected. For the coordination/phrase contrast, a difference was expected because Lau and Liao (2018) had found a sustained effect of coordination (in EEG) that began with the coordinator but persisted throughout the second noun phrase. This difference was interpreted as reflecting the maintenance of structure in memory, or semantic computations associated with that structure. Though we again acknowledge that it is difficult and problematic to reason from a null result, and our failure to observe the effect is quite not possibly not meaningful, we will enumerate some possible explanations for the manifestation of the effect in one case

but not the other, in the case that our effect is a true negative. Only follow-up and attempted replication can clarify whether either or both of the observed patterns are robust.

First, of course, we acknowledge that we are comparing two different methods—MEG and EEG—whose data do not necessarily correspond. Analysis of the EEG data that we have collected for the same paradigm will help address whether the method is a factor in the null effect.

Second, as in the comparison of the current study to Matchin et al. (2019), there are a variety of changes we made to the design that could potentially explain our failure to observe any difference. Lau and Liao (2018) used bare noun phrases, while we used determiner phrases. Their noun phrases were constructed to be "meaningful and plausible" (e.g., "sunlit ponds"), in contrast to our random combinations of adjective and noun. This means that their stimuli allow for some degree of expectation of the noun given the adjective, which is not possible in our randomly combined phrases. However, this type of lexical prediction is not consistent with the temporal profile of their coordination effect (a sustained negativity during the second noun phrase). Furthermore, the two studies are identical with respect to the fact that there was no relationship between the first and second noun phrase, so the Lau and Liao (2018) effect cannot be related to expectation for the second phrase given the first phrase.

Instead, the lack of lexical predictability in our study could have led to difficulty in semantic composition because of repeated nonsensical combinations, leading to the shutdown of such operations. If the sustained effect observed by Lau

and Liao (2018) reflected semantic interpretation processes, those processes would then be curtailed. This aligns with the fact that they did not observe the coordination effect in their jabberwocky condition. If this explanation is right, we might expect the coordination effect to be present at the start of the experiment, stopping as it becomes clear that the stimuli are semantically nonsense.

B.2.1.2 Sentence vs. phrase

For the sentence/phrase contrast, the lack of difference during the second noun phrase was also surprising, although in a less specific way. Given the many syntactic and semantic computations that are thought to be occurring with the integration of an object noun phrase into a sentence, we expected large differences whose multiple and distinct contributing causes would not be separable. Instead, we did not find any difference. In comparisons between the current study and Lau and Liao (2018), it is true in either case that the comparison between conditions is for the same stimuli presented in different contexts. This is however not the case for Matchin et al. (2019), for which the same phrases are not always appearing in both conditions, and for which the phrase condition is sometimes made up of verb phrases and sometimes noun phrases. It seems unlikely that e.g. the structure effects that Matchin et al. (2019) observed at all positions are due to this lack of lexical matching, but this possibility cannot be ruled out. We discuss this further in **Section B.2.3** below.

B.2.2 Verb vs. “and” vs. blank

If no difference was observable during the subject noun phrase indicating prediction of structure, and no difference was observable during the object noun phrase indicating structure-building or integration, then (again, with the caveat that

this only holds under the assumption of a true negative) the only potential locus of structure-building we can identify in our data is during the response to the verb or coordinator. This response also, of course, reflects the fact that very different words are being processed. In **Table 10**, we summarize the results of the three pairwise spatiotemporal cluster tests, including the nature of each response component with respect to how it applies to function vs. content words. We make this designation on the basis not only of differences between the response to a verb versus "and", but also, in looking at the full time course of activity in the cluster, whether the peak on which the effect occurs is present only for content words or for all words. We also include a handful of clusters with p between 0.05 and 0.1, which were not reported in the **Results** section, for a more complete picture.

This turns out to be an unusually clear dataset for examining the time course and response components of visual word recognition and lexical access. It is also unique in that our neutral comparison condition is an absence of input (the blank screen between phrases in the phrase condition) rather than making a comparison solely between different kinds of words or different kinds of visual/orthographic input.

The evoked response to visually presented words in MEG is characterized by well-established temporal landmarks. As recently reviewed by Dikker et al. (2019), pre-lexical response components include the M100, reflecting early visual processing of the stimulus; the M130, demonstrating sensitivity to orthographic features; and the M170, modulated by morphological properties. These three components are generally localized to occipito-temporal areas and the inferior temporal and fusiform gyri.

Table 10. *Summary and re-grouping of pairwise spatiotemporal cluster test results. Grey italicized text indicates clusters with p between 0.05 and 0.1, which were not reported in the Results section.*

	Contrast	Location	Window	*p*-value	Component, direction	Word type
1	Coordination /phrase	Ventral	78-198 ms	0.035	M130, negative	both
	Sentence /phrase	*Ventral*	*70-278 ms*	*0.056*		
2	Sentence /phrase	TTS	112-524 ms	0.002	M170 + M350, negative	open-class
	Sentence /phrase	STS/MTG +ITS	164-522 ms	0.002		
	Sentence /coordination	TTS +STS/MTG +ITS	146-572 ms	0.000		
	Sentence /phrase	pSTS	156-532 ms	0.004	M170 + M350, positive	
	Sentence /coordination	pSTS	154-576 ms	0.001		
	Sentence /coordination	ITS/ventral	138-530 ms	0.000		
3	Coordination /phrase	Anterior ventral	208-486 ms	0.024	negative	closed-class
	Sentence /phrase	*Anterior ventral*	*176-514 ms*	*0.098*	*positive*	*open-class*
4	*Sentence /phrase*	*IFG*	*222-500 ms*	*0.058*	negative	open-class
	Sentence /coordination	IFG	252-476 ms	0.036		
5	Coordination /phrase	ATL	90-582 ms	0.013	sustained, positive	graded (open > closed)
	Sentence /phrase	ATL	226-518 ms	0.018		
	Sentence /coordination	*ATL*	*252-458 ms*	*0.066*		

The 350-500 ms window in processing is understood to encompass lexical access and ensuing combinatorics; the M350, likely analogue to the N400 observed in

EEG, is likely generated in posterior middle and superior temporal areas. We now consider each group of clusters from the current dataset as we have summarized them in **Table 10**, and attempt to contextualize them within the above timeline, with the goal of identifying which could plausibly represent a structure effect and which are likely just due to the presentation of different words. In this section, where we hypothesize about the difference between the sentence/phrase and coordination/phrase contrasts, we need to highlight that these two comparisons are not independent, because the phrase condition is the same data in either case.

B.2.2.1 M130/early visual response

Along ventral temporal lobe (group 1 in **Table 10**), a cluster from the coordination/phrase contrast shows a large negative peak with latency ~140 ms that occurs in response to "and" but not in response to the blank. A marginally significant cluster from the sentence/phrase contrast shows the same effect for the verb relative to the blank. The sentence/coordination contrast did not show any differences between the two conditions in this spatial/temporal area. Examining the full time course of activity in the cluster, we see that the 140 ms peak recurs for every word in the trial, with no apparent difference between conditions, in either the coordination/phrase or sentence/phrase contrast. This is, of course, a null effect, but we note that we do not find support for differences in attention between the two conditions that modulate the early visual response.

This effect is consistent with the M130 in timing and in the fact that different kinds of words appear to evoke the same response. As for polarity and location, Tarkiainen et al. (1999) thoroughly characterized early visual responses to

orthographic input in a study that presented letter and symbol strings with varying degrees of visual noise. One of their findings was a response at occipito-temporal junction in the 140-170 ms window that showed increasing neural activity with decreasing visual noise, as well as increased activity for letters relative to symbols. Gwilliams, Lewis, and Marantz (2016), following up on this work, found two different clusters for these two patterns of response, both of which appear to have peak latency ~ 150 ms. The first cluster shows a negative peak at the occipito-temporal junction, increasing in amplitude with decreasing visual noise; when projected back into sensor space, this cluster has a peak latency of 149 ms. The second cluster shows a positive peak in the anterior fusiform area, increasing in amplitude for letters relative to symbols, with a peak latency of 172 ms when projected back into sensor space. These findings of Gwilliams et al. (2016) are consistent with the results of for example Solomyak and Marantz (2010), who examined visual word recognition with MEG, and considered their negative peak in the range of 140-150 ms in the posterior occipital area to be the M130 response, and a positive peak at 180-190 ms in the occipito-temporal area to be the M170. Cavalli et al. (2016) also find an orthographic effect in the M130 time window in posterior left ITG. Our negative peak at 140 ms, which does not appear to differ for different types of words, therefore corresponds well with this M130 response.

B.2.2.2 M170 and M350

We also see a peak in the time course data for the current study that we suspect corresponds to the M170 response observed in other studies. This response, as detailed above, is reported to be sensitive to letter strings, but has also been observed

in several cases to be modulated by morphological properties of the visually presented word (Fruchter et al., 2013; Lewis et al., 2011; Solomyak & Marantz, 2009, 2010; Zweig & Pylkkänen, 2009). In our data, this response manifests primarily alongside a later, larger response that is likely the M350. Thus, we cannot make strong claims about their separability, but in all of the clusters from group 2 reported in **Table 10**, we see an earlier peak at roughly 170 ms and a later peak at roughly 350 ms. These clusters are largely observed along the sulci and gyri of the lateral temporal lobe, though one extends more ventrally. Specifically, we observe clusters showing two peaks with positive polarity along pSTS, ITS and more ventrally, and clusters showing those same two peaks with negative polarity along TTS, STS, MTG, and ITS. As explained for the localizer data in **Chapter 5**, we consider these clusters to be the same effect as measured on either side of the gyri, and the data do not allow us to determine where the effect originates or is strongest.

All of these clusters manifest as activity for the sentence condition relative to the phrase or coordination condition. There are no differences for the coordination relative to the phrase condition. Given the apparent morphological sensitivity of M170, the early difference could be due to the morphological complexity of the verb relative to the coordinator. For the later peak (the M350), content versus function word is likely the right distinction.

Prior literature indicates that the M170 should localize to the fusiform gyrus and that the polarity of the peak should be positive. Prior literature on the M350 indicates that the polarity of the peak should be negative and the effect should localize to middle temporal areas (Fruchter et al., 2013; Halgren et al., 2002). Why,

then do we only detect the M170 in clusters in which it precedes the M350 in these lateral temporal areas? Given the nature of spatiotemporal cluster tests, we suspect the M170 is a weaker effect that is picked up in these lateral areas only because of the strength of the M350 and their being contiguous in time. It would otherwise be surprising to see a distinction between different words this early along lateral temporal lobe. It may be that the origin of the (positive) M170 is ventral, but is mirrored laterally (appearing negative), while the origin of the (negative) M350 is lateral, but is mirrored ventrally (appearing positive).

How would we distinguish a structure effect in pSTS from middle temporal M350 effects we are observing? That is of course impossible in our design, but might be possible with more targeted manipulation of single word versus sentence properties. Given the theory of lexicalized syntax described by Matchin and Hickok (2020), one consideration is that the M350 is inherently syntactic, in the sense that lexical retrieval also involves retrieval of associated structure.

B.2.2.3 Other effects in the lexical window

Besides this group of clusters showing the M170/M350 response to the verb but not the coordinator, we have several other effects in the time window of lexical processing, in different areas and showing different patterns with respect to the three conditions. If we could confidently identify which (if any) of these effects is associated with structure, one interesting implication would arise from whether such a structure effect is graded or does not occur for the coordinator.

The Group 3 anterior ventral clusters appear to show a peak at 350 ms, but the peak is positive for the coordinator and negative for the verb, with the blank for the

phrase condition patterning in the middle. Cavalli et al. (2016) also report an effect in this time window along ITG and the fusiform gyrus that they attribute to form and meaning access, but not in such a way that we would expect effects in opposite directions for our two word conditions. In IFG (Group 4), with a similar 350 ms (negative) peak, we see a response only for the verb. Again, Cavalli et al. (2016) report an IFG effect in a similar window that they attribute to orthographic and semantic recombination; this is actually consistent with our observing a response in that area only for the verb. Matchin and Hickok (2020) propose that IFG effects of structure might occur in comprehension because of top-down preactivation of lexical syntax.

In ATL (Group 5), a more sustained positive response occurs in both word conditions relative to the blank, but is much larger for the verb. The coordination/phrase difference may correspond to the coordination effect that Lau and Liao (2018) observed in response to the coordinator, though their effect continued for the rest of the noun phrase. Otherwise, we expect activity in this area to reflect some form of semantic composition (see Pylkkänen (2020) for review). Our effect is more sustained than is usually observed for simple adjective/noun composition, but it does sometimes appear more sustained in more complex combinatory situations (Brennan & Pylkkänen, 2017; Westerlund et al., 2015). Interestingly, a recent study from this group manipulating syntactic rather than semantic composition (Flick & Pylkkänen, 2020) finds a difference in PTL rather than ATL, around 200 ms, but we do not observe any corresponding effects.

B.2.3 Matchin et al.'s every-word structure effects

Finally, what do we make of the apparent effects of structure Matchin et al. (2019) observed on all open and closed-class words, as well as specifically on the determiner in the final DP? None of these effects occurred in our study with nonsense stimuli. In this section, we walk through several potential explanations.

Matchin et al. (2019) hypothesize that the effects occurring on all words arise because of increased attention or maintenance required for lexical-syntactic representations when they have to be integrated into sentence-level structure, as compared to when they do not. Our study would not invoke such attention or maintenance costs if participants did not parse the sentences as structured input.

A variant on this logic would be that the heightened attention in the sentence condition for Matchin et al. (2019) is due not to the necessity of integration into higher-level structure but due to the fact that the specific identity of the word is relevant for construction of meaning in a way that it is not in the phrase condition, and is not when the stimuli are nonsense, as in our study. This would make the processing of nonsense stimuli more akin to the processing of jabberwocky, and if this is the case then the lack of structure effects in our study could fall under the same explanation that we offered above for the general lack of jabberwocky structure effects in MEG. That is, when lexical identity is unavailable or ignored, syntactic structure-building loses one of its primary sources of information. Individual participants might then vary in their reliance on other cues (inflection, function words) as well as in whether this pushes them from a top-down to a bottom-up parsing mode. Such temporal variability would make it much more difficult to

observe structure effects in MEG than fMRI; it is therefore possible that our nonsense paradigm might have yielded structure effects if run in fMRI.

We acknowledge that these structure effects could also have arisen for Matchin et al. (2019) because the stimuli were not perfectly matched between conditions. Their phrase condition was sometimes a sequence of verb phrases and sometimes a sequence of noun phrases, and so the comparison was never between identical noun phrases with sentence-level structure as the only difference. However, we do not have any hypothesis as to why the lack of matching would have led to an effect specifically in this direction.

The final explanation we can consider is that these apparent structure effects were due to lexical predictability, in parallel with the apparent syntactic prediction effect on the subject noun. However, we think these effects are a poor fit with the lexical prediction account. First, the effect on open-class words is poorly aligned with the prediction effect on the subject noun because it is far less sustained in time (occurring from 284-332 ms as compared to 272-484 ms). In fact, it seems plausible that the prediction effect is comprised of first the open-class effect and then a separate effect that is later and longer-lasting.

The late effect of structure on the determiner before the object noun is also not at the right word position to be parallel to the subject noun effect; we would instead expect the prediction for the object noun to manifest on the verb.

The same issue holds for the effect occurring on all closed-class words, which is far earlier than the effect on the subject noun (92-148 ms rather than 272-484 ms). The potential contribution of a closed-class word to prediction of the open-class word

that follows it is (at least in this design) largely limited to the syntactic category of that upcoming word, but this possibility is the same in the sentence and phrase condition, and therefore should not manifest in a sentence/phrase contrast. It could potentially manifest in the phrase/list contrast, but Matchin et al. (2019) did not find such an effect in fMRI, and supplementary figures for the MEG data suggest it is not present there either. The Payne et al. (2015) finding that open-class N400 amplitudes do not reduce with word position in nonsense sentences, where the only possible prediction about each upcoming word is its syntactic category, is consistent with the null effects for the phrase/list contrast.

Prediction of at least the syntactic category of closed-class words, from the open-class words that precede them, is another effect that should only have manifested in the phrase/list contrast if occurring (but did not). Payne et al. (2015) showed that for closed-class words N400 amplitude does not decrease with word position in a natural sentence, suggesting that prediction of the sort we are concerned with may be occurring only for open-class words (and as triggered, apparently, also by only open-class words). This converges with recent claims that evidence for prediction of determiners before predictable nouns is weaker than previously thought (Kochari & Flecken, 2019; Nieuwland et al., 2018).

In summary, we do not believe that the other structure effects observed by Matchin et al. (2019) are good candidates for lexical prediction effects.

Bibliography

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.

Baayen, H., Piepenbrock, R., & Gulikers, L. (1995). *CELEX2 LDC96L14 [Web Download]*. Linguistic Data Consortium.

Baayen, H., Wurm, L. H., & Aycock, J. (2007). Lexical dynamics for low-frequency complex words: A regression study across tasks and modalities. *The Mental Lexicon*, *2*(3), 419–463. https://doi.org/10.1075/ml.2.3.06baa

Balling, L. W., & Baayen, H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition*, *125*(1), 80–106. https://doi.org/10.1016/j.cognition.2012.06.003

Bien, H., Baayen, R. H., & Levelt, W. J. M. (2011). Frequency effects in the production of Dutch deverbal adjectives and inflected verbs. *Language and Cognitive Processes*, *26*(4–6), 683–715. https://doi.org/10.1080/01690965.2010.511475

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

Brennan, J., & Hale, J. T. (2019). Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLOS ONE*, *14*(1), e0207741. https://doi.org/10.1371/journal.pone.0207741

Brennan, J., Lignos, C., Embick, D., & Roberts, T. P. L. (2014). Spectro-temporal correlates of lexical access during auditory lexical decision. *Brain and Language*, *133*, 39–46. https://doi.org/10.1016/j.bandl.2014.03.006

Brennan, J., & Pylkkänen, L. (2017). MEG Evidence for Incremental Sentence Composition in the Anterior Temporal Lobe. *Cognitive Science*, *41*, 1515–1531. https://doi.org/10.1111/cogs.12445

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

Brodbeck, C., Jiao, A., Hong, L. E., & Simon, J. Z. (2019). *Neural speech restoration at the cocktail party: Auditory cortex recovers masked speech of both attended and ignored speakers* [Preprint]. Neuroscience. https://doi.org/10.1101/866749

Brodbeck, C., Presacco, A., & Simon, J. Z. (2018). Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. *NeuroImage*, *172*, 162–174. https://doi.org/10.1016/j.neuroimage.2018.01.042

Brodbeck, C., Proloy Das, Teon L Brooks, & Reddigari, S. (2019). *Eelbrain 0.31* (v0.31) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.3564850

Brothers, T., Dave, S., Hoversten, L. J., Traxler, M. J., & Swaab, T. Y. (2019). Flexible predictions during listening comprehension: Speaker reliability affects anticipatory processes. *Neuropsychologia*, *135*, 107225. https://doi.org/10.1016/j.neuropsychologia.2019.107225

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

Brysbaert, M., New, B., & Keuleers, E. (2012a). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, *44*(4), 991–997. https://doi.org/10.3758/s13428-012-0190-4

Brysbaert, M., New, B., & Keuleers, E. (2012b). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, *44*(4), 991–997. https://doi.org/10.3758/s13428-012-0190-4

Cavalli, E., Colé, P., Badier, J.-M., Zielinski, C., Chanoine, V., & Ziegler, J. C. (2016). Spatiotemporal Dynamics of Morphological Processing in Visual Word Recognition. *Journal of Cognitive Neuroscience*, *28*(8), 1228–1242. https://doi.org/10.1162/jocn_a_00959

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

Davies, M. (2008). *The Corpus of Contemporary American English (COCA): 520 million words, 1990-present*.

Delaney-Busch, N., Morgan, E., Lau, E., & Kuperberg, G. R. (2019). Neural evidence for Bayesian trial-by-trial adaptation on the N400 during semantic priming. *Cognition*, *187*, 10–20. https://doi.org/10.1016/j.cognition.2019.01.001

Di Liberto, G. M., Wong, D., Melnik, G. A., & de Cheveigné, A. (2019). Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. *NeuroImage*, *196*, 237–247. https://doi.org/10.1016/j.neuroimage.2019.04.037

Dikker, S., Assaneo, F., Gwilliams, L., Wang, L., & Kösem, A. (2019). *Using Magnetoencephalography to Study the Neural Basis of Language*. 17.

Dikker, S., & Pylkkänen, L. (2013). Predicting language: MEG evidence for lexical preactivation. *Brain and Language*, *127*(1), 55–64. https://doi.org/10.1016/j.bandl.2012.08.004

Donhauser, P. W., & Baillet, S. (2020). Two Distinct Neural Timescales for Predictive Speech Processing. *Neuron*, *105*(2), 385-393.e9. https://doi.org/10.1016/j.neuron.2019.10.019

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D.,

& Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

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

Ettinger, A., Linzen, T., & Marantz, A. (2014). The role of morphology in phoneme prediction: Evidence from MEG. *Brain and Language*, *129*, 14–23. https://doi.org/10.1016/j.bandl.2013.11.004

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

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. *Journal of Neurophysiology*, *104*(2), 1177–1194. https://doi.org/10.1152/jn.00032.2010

Fedorenko, E., Nieto-Castañon, A., & Kanwisher, N. (2012). Lexical and syntactic representations in the brain: An fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia*, *50*(4), 499–513. https://doi.org/10.1016/j.neuropsychologia.2011.09.014

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., & Kanwisher, N. (2016). Neural correlate of the construction of sentence

meaning. *Proceedings of the National Academy of Sciences*, *113*(41), E6256–E6262. https://doi.org/10.1073/pnas.1612132113

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

Flick, G., & Pylkkänen, L. (2020). Isolating syntax in natural language: MEG evidence for an early contribution of left posterior temporal cortex. *Cortex*, *127*, 42–57. https://doi.org/10.1016/j.cortex.2020.01.025

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.

Francis, W. N., & Kučera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Houghton Mifflin.

Fruchter, J., Stockall, L., & Marantz, A. (2013). MEG masked priming evidence for form-based decomposition of irregular verbs. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00798

Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal Predictive Codes for Spoken Words in Auditory Cortex. *Current Biology*, *22*(7), 615–621. https://doi.org/10.1016/j.cub.2012.02.015

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

Goucha, T., & Friederici, A. D. (2015). The language skeleton after dissecting meaning: A functional segregation within Broca's Area. *NeuroImage*, *114*, 294–302. https://doi.org/10.1016/j.neuroimage.2015.04.011

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00267

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., & Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage*, *86*(Supplement C), 446–460. https://doi.org/10.1016/j.neuroimage.2013.10.027

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

Grosjean, F., Dommergues, J.-Y., Cornu, E., Guillelmon, D., & Besson, C. (1994). The gender-marking effect in spoken word recognition. *Perception & Psychophysics*, *56*(5), 590–598. https://doi.org/10.3758/BF03206954

Gwilliams, L., King, J.-R., Marantz, A., & Poeppel, D. (2020). *Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content* [Preprint]. Neuroscience. https://doi.org/10.1101/2020.04.04.025684

Gwilliams, L., Lewis, G. A., & Marantz, A. (2016). Functional characterisation of letter-specific responses in time, space and current polarity using

magnetoencephalography. *NeuroImage*, *132*(Supplement C), 320–333. https://doi.org/10.1016/j.neuroimage.2016.02.057

Gwilliams, L., & Marantz, A. (2015). Non-linear processing of a linear speech stream: The influence of morphological structure on the recognition of spoken Arabic words. *Brain and Language*, *147*(Supplement C), 1–13. https://doi.org/10.1016/j.bandl.2015.04.006

Hahne, A., & Jescheniak, J. D. (2001). What's left if the Jabberwock gets the semantics? An ERP investigation into semantic and syntactic processes during auditory sentence comprehension. *Cognitive Brain Research*, *11*(2), 199–212. https://doi.org/10.1016/S0926-6410(00)00071-9

Halgren, E., Dhond, R. P., Christensen, N., Van Petten, C., Marinkovic, K., Lewine, J. D., & Dale, A. M. (2002). N400-like Magnetoencephalography Responses Modulated by Semantic Context, Word Frequency, and Lexical Class in Sentences. *NeuroImage*, *17*(3), 1101–1116. https://doi.org/10.1006/nimg.2002.1268

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.

Hayes, B., & Wilson, C. (2008). A Maximum Entropy Model of Phonotactics and Phonotactic Learning. *Linguistic Inquiry*, *39*(3), 379–440. https://doi.org/10.1162/ling.2008.39.3.379

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

Humphries, C., Binder, J. R., Medler, D. A., & Liebenthal, E. (2006). Syntactic and Semantic Modulation of Neural Activity during Auditory Sentence Comprehension. *Journal of Cognitive Neuroscience*, *18*(4), 665–679. https://doi.org/10.1162/jocn.2006.18.4.665

Kemps, R. J. J. K., Wurm, L. H., Ernestus, M., Schreuder, R., & Baayen, H. (2005). Prosodic cues for morphological complexity in Dutch and English. *Language and Cognitive Processes*, *20*(1–2), 43–73. https://doi.org/10.1080/01690960444000223

Kocagoncu, E., Clarke, A., Devereux, B. J., & Tyler, L. K. (2017). Decoding the Cortical Dynamics of Sound-Meaning Mapping. *The Journal of Neuroscience*, *37*(5), 1312–1319. https://doi.org/10.1523/JNEUROSCI.2858-16.2016

Kochari, A. R., & Flecken, M. (2019). Lexical prediction in language comprehension: A replication study of grammatical gender effects in Dutch. *Language, Cognition and Neuroscience*, *34*(2), 239–253. https://doi.org/10.1080/23273798.2018.1524500

Lau, E., Almeida, D., Hines, P. C., & Poeppel, D. (2009). A lexical basis for N400 context effects: Evidence from MEG. *Brain and Language*, *111*(3), 161–172. https://doi.org/10.1016/j.bandl.2009.08.007

Lau, E., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 Effects of Prediction from Association in Single-word Contexts. *Journal of Cognitive Neuroscience*, *25*(3), 484–502. https://doi.org/10.1162/jocn_a_00328

Lau, E., & Liao, C.-H. (2018). Linguistic structure across time: ERP responses to coordinated and uncoordinated noun phrases. *Language, Cognition and Neuroscience*, *33*(5), 633–647. https://doi.org/10.1080/23273798.2017.1400081

Lau, E., & Namyst, A. (2019). FMRI evidence that left posterior temporal cortex contributes to N400 effects of predictability independent of congruity. *Brain and Language*, *199*, 104697. https://doi.org/10.1016/j.bandl.2019.104697

Lau, E., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933. https://doi.org/10.1038/nrn2532

Lewis, G., & Poeppel, D. (2014). The role of visual representations during the lexical access of spoken words. *Brain and Language*, *134*, 1–10. https://doi.org/10.1016/j.bandl.2014.03.008

Lewis, G., Solomyak, O., & Marantz, A. (2011). The neural basis of obligatory decomposition of suffixed words. *Brain and Language*, *118*(3), 118–127. https://doi.org/10.1016/j.bandl.2011.04.004

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

Matchin, W., Brodbeck, C., Hammerly, C., & Lau, E. (2019). The temporal dynamics of structure and content in sentence comprehension: Evidence from fMRI-constrained MEG. *Human Brain Mapping*, *40*(2), 663–678. https://doi.org/10.1002/hbm.24403

Matchin, W., Hammerly, C., & Lau, E. (2017). The role of the IFG and pSTS in syntactic prediction: Evidence from a parametric study of hierarchical structure in fMRI. *Cortex*, *88*, 106–123. https://doi.org/10.1016/j.cortex.2016.12.010

Matchin, W., & Hickok, G. (2020). The Cortical Organization of Syntax. *Cerebral Cortex*, *30*(3), 1481–1498. https://doi.org/10.1093/cercor/bhz180

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

Mollica, F., Siegelman, M., Diachek, E., Piantadosi, S. T., Mineroff, Z., Futrell, R., Kean, H., Qian, P., & Fedorenko, E. (2020). Composition is the Core Driver of the Language-selective Network. *Neurobiology of Language*, *1*(1), 104–134. https://doi.org/10.1162/nol_a_00005

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*(1), 1–25. https://doi.org/10.1002/hbm.1058

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D. J., Rousselet, G. A., Ferguson, H. J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E. M., … Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *ELife*, *7*, e33468. https://doi.org/10.7554/eLife.33468

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

Pallier, C., Devauchelle, A.-D., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, *108*(6), 2522–2527. https://doi.org/10.1073/pnas.1018711108

Payne, B. R., Lee, C.-L., & Federmeier, K. D. (2015). Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials: Effects of context on world-level N400. *Psychophysiology*, *52*(11), 1456–1469. https://doi.org/10.1111/psyp.12515

Pylkkänen, L. (2020). Neural basis of basic composition: What we have learned from the red–boat studies and their extensions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20190299. https://doi.org/10.1098/rstb.2019.0299

Rogalsky, C., & Hickok, G. (2009). Selective Attention to Semantic and Syntactic Features Modulates Sentence Processing Networks in Anterior Temporal Cortex. *Cerebral Cortex*, *19*(4), 786–796. https://doi.org/10.1093/cercor/bhn126

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

Sassenhagen, J., & Draschkow, D. (2019). Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology*, *56*(6), e13335. https://doi.org/10.1111/psyp.13335

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

Solomyak, O., & Marantz, A. (2009). Lexical access in early stages of visual word processing: A single-trial correlational MEG study of heteronym recognition. *Brain and Language*, *108*(3), 191–196. https://doi.org/10.1016/j.bandl.2008.09.004

Solomyak, O., & Marantz, A. (2010). Evidence for Early Morphological Decomposition in Visual Word Recognition. *Journal of Cognitive Neuroscience*, *22*(9), 2042–2057. https://doi.org/10.1162/jocn.2009.21296

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), 213–234. https://doi.org/10.1016/0010-0277(87)90010-2

Tarkiainen, A., Helenius, P., Hansen, P. C., Cornelissen, P. L., & Salmelin, R. (1999). Dynamics of letter string perception in the human occipitotemporal cortex. *Brain*, *122*(11), 2119–2132. https://doi.org/10.1093/brain/122.11.2119

Taulu, S., & Simola, J. (2006). Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Physics in Medicine and Biology*, *51*(7), 1759–1768. https://doi.org/10.1088/0031-9155/51/7/008

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2019). The Massive Auditory Lexical Decision (MALD) database. *Behavior Research Methods*, *51*(3), 1187–1204. https://doi.org/10.3758/s13428-018-1056-1

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

Van Petten, C. (1993). A comparison of lexical and sentence-level context effects in event-related potentials. *Language and Cognitive Processes*, *8*(4), 485–531. https://doi.org/10.1080/01690969308407586

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

Westerlund, M., Kastner, I., Al Kaabi, M., & Pylkkänen, L. (2015). The LATL as locus of composition: MEG evidence from English and Arabic. *Brain and Language*, *141*, 124–134. https://doi.org/10.1016/j.bandl.2014.12.003

Wurm, L. H., Ernestus, M. T. C., Schreuder, R., & Baayen, H. (2006). Dynamics of the auditory comprehension of prefixed words: Cohort entropies and Conditional Root Uniqueness Points. *The Mental Lexicon*, *1*(1), 125–146. https://doi.org/10.1075/ml.1.1.08wur

Yamada, Y., & Neville, H. J. (2007). An ERP study of syntactic processing in English and nonsense sentences. *Brain Research*, *1130*, 167–180. https://doi.org/10.1016/j.brainres.2006.10.052

Zweig, E., & Pylkkänen, L. (2009). A visual M170 effect of morphological complexity. *Language and Cognitive Processes*, *24*(3), 412–439. https://doi.org/10.1080/01690960802180420

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