Effects of context on neural oscillations during free-reading of stories

by

MAX CANTOR

B.A., University of Michigan, 2013

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Master of Arts

Department of Psychology and Neuroscience

2018

This thesis entitled:
Effects of context on neural oscillations during free-reading of stories
written by Max Cantor
has been approved for the Department of Psychology

_____
Albert Kim

_____
Tim Curran

Date_____

The final copy of this thesis has been examined by the signatories, and we
Find that both the content and the form meet acceptable presentation standards
Of scholarly work in the above mentioned discipline

Abstract

Cantor, Max (M.A., Psychology and Neuroscience)

Effects of context on neural oscillations during free-reading of stories

Thesis directed by Associate Professor Albert Kim

Context plays a fundamental role in language comprehension; however, much of what we know about the effects of context come from isolated sentences, and the role of context during free-reading of naturalistic stories is less well understood. It is plausible that experimentally manipulated, isolated sentences provide fewer predictive affordances than stories, where the topic of a story may provide top-down contextual information in addition to the linguistic constraints on sentences. These sources of contextual information have been shown to affect language comprehension, as measured by eye movements and electroencephalographic (EEG) scalp potentials. We examined the effects of context on the neural oscillatory activity at multiple frequency-bands during free-reading of naturalistic stories using time-frequency representations (TFRs), derived from coregistration of the EEG signal and eye movements. We compared words with low vs. high contextual fit, derived from the relationship between fixated words and the words preceding them (local contextual fit), and to the topic of the story (topic contextual fit). We found TFR effects of contextual fit, suggesting that similar neurocognitive mechanisms are engaged for language comprehension during free-reading of naturalistic stories as compared to reading of words in isolated sentences. However, the TFR effects were present only for local contextual fit, not topic contextual fit. Given the naturalistic and exploratory nature of the present study, we provide a speculative account of these results.

Acknowledgments

I would like to thank my advisor, Albert Kim, and the entire Kim Lab, for their support on this project. From the Kim Lab staff, I would specifically like to thank Don Bell-Souder and Sara Milligan, the research assistants who implemented the stimulus and recording paradigm for this study. I would also like to thank Shannon McKnight, another graduate student in the Kim Lab who originally wrote the stories used in the present study. Her research using these stories, as represented by her Master's thesis and an in-preparation paper, helped inform the direction of the present study. I would also like to thank our collaborators John Trueswell and Phillip Gilley, who contributed during the design and implementation stage of the present study.

CONTENTS

TABLES

Table

FIGURES

Figure

Introduction

The question: “How many animals of each kind did Moses take on the ark?” (Erikson & Mattson, 1981), is a strong demonstration of the power of context on language comprehension. People commonly respond to this question erroneously, failing to recognize that it was Noah, not Moses, who built the ark. A large body of prior research has shown that context provides constraints on the interpretation of a sentence, affecting whether a word is expected or unexpected, leading to large differences in brain responses. Something that is missing from this large body of work is an account of how context affects language processing in naturalistic contexts. Naturalistic texts are often rich in predictive affordances, but also contain very few words which are difficult to interpret or highly unexpected, as is often the case in experimental studies of sentence comprehension. The fact that such situations have been little studied raises important questions about the generalizability of the previous work in sentence comprehension to real world language processing.

In the present research, subjects read linguistically well-formed, naturalistic stories, while their eye movements were recorded, and scalp-recorded EEG was used to observe neural oscillations. Computational language models were used to derive measures of contextual fit, the relationship between a word and its context. Differences in the eye movements, EEG scalp potentials, and neural oscillatory dynamics were compared based on these measures of context. The focus of this research was to draw inferences about the functional cortical networks utilized for language comprehension during free-reading of naturalistic stories. Neural oscillations derived from the EEG using time-frequency representations (TFRs) have been suggested to reflect functional cortical networks engaged during language comprehension and other cognitive tasks, and may inform our interpretation of the on-line processes engaged during reading. To

draw these inferences, we used two models of contextual fit and contrasted the eye movements, fixation-related potential (FRP), and TFR response to words with low vs. high contextual fit. Whereas context in isolated sentences may be driven more so by the local, lexico-semantic properties of the preceding text, global topic-based knowledge related to story text may also contribute to naturalistic language comprehension, a distinction which has been under-explored in the prior experimental literature. It is plausible that the effects of context found to language anomalies or low probability words in isolated sentences are exaggerated as compared to effects of context in naturalistic texts, where readers can utilize topic-based contextual information.

In study 1, we examine lexical and context effects on the eye movements, which have been shown to reflect language comprehension difficulty. In study 2, we examine effects of local contextual fit on the FRPs and TFRs, to understand how the lexico-semantic properties of the context affect on-line language comprehension. In study 3, we examine effects of topic contextual fit on the FRPs and TFRs, to understand the role of topic knowledge on language comprehension during naturalistic story reading.

Neural oscillations

In the present study, TFRs of the neural oscillations, reflecting the rhythmic patterns of post-synaptic potentials (Buzsaki & Draguhn, 2004), were used to index the neurocognitive processes of on-line language comprehension during free-reading of stories. Oscillatory activity is often characterized in terms of specific clinically and experimentally-derived frequency-bands. In the present study, we examined activity in the theta-band (4-7 Hz), alpha-band (8-12 Hz), beta-band (13-30 Hz), and lower gamma-band (35-45 Hz). Activity at these frequency bands is thought to reflect the coupling and uncoupling of neurons, forming functional networks in the

brain. TFRs may be able to situate language processes within hierarchical functional cortical networks including other processes such as attention, memory, and executive function, as well as dissociate neural oscillations which are unique to language comprehension as compared to other cognitive mechanisms (Kielar, Panamsky, Links, & Meltzer, 2015).

Lexico-semantic retrieval and the theta-band. Increases in theta-band activity (4-7 Hz) have been found in response to word presentation, and have been suggested to correspond to lexico-semantic retrieval. Posterior theta-band increases have been found for words with overall greater semantic content compared to words with less overall semantic content. Greater occipital theta-band power increases have been found for open class words such as nouns, verbs, and adjectives, as compared to closed class words such as determiners, conjunctions, and prepositions (Bastiaansen, Linde, Keurs, Dijkstra, & Hagoort, 2005). Open class words contain more semantic information than closed class words, and so the theta-band increase in this case is suggested to reflect the lexico-semantic retrieval process. Context has also been shown to affect theta-band activity. When comparing semantically plausible control sentences such as “The wind swept through the trees”, to semantically anomalous sentences such as “the girl speaks three trees”, there was an increase in theta-band activity for the violation words vs. controls (Davidson & Indefrey, 2007). In terms of the predictive affordances of context, this suggests that readers make on-line predictions of what words they expect given the context, activating a functional lexico-semantic retrieval network. When a word is unexpected given its context, the relative increase in theta-band power as compared to more plausible words may reflect the increase in integration difficulty due to the incorrectly activated lexico-semantic retrieval network. Posterior theta-band increases have also been mapped to the hippocampus, and episodic memory encoding

and retrieval processes (Hasselmo, Bodelon, & Wyble, 2002), supporting the role of the theta-band in memory-retrieval networks more generally.

Sentence reanalysis and the theta-band. Theta-band increases to language violations have also been found at frontal electrodes, and are believed to correspond to increased demand on working memory due to semantic integration difficulty. Comparing words which were semantically anomalous given their context to control words, Hald, Bastiaansen, and Hagoort (2006), found an increase in theta-band power to semantic violations vs. controls at frontal electrodes, which they describe as an increase in working memory load due to the unexpectedness of the semantic violation. The frontal theta-band increase has also been found for words with low vs. high predictability, and specifically when the context was strongly constrained, but not when weakly constrained (Rommers, Dickson, Norton, Wlotko, & Federmeier, 2017). This may suggest that the demand on working memory varies by the properties of the context, such as how it constrains the interpretation of upcoming words. It has also been suggested that a general conflict resolution mechanism driven by prefrontal cortex is recruited in the context of language processing (Novick, Trueswell, & Thompson-Schill, 2005), and it is plausible that activity in the frontal theta-band reflects this same mechanism. This would converge with research from executive function suggesting that frontal theta-band increases reflect cognitive control on task representations (Cavanagh & Frank, 2014).

In sum, the theta-band has been implicated in semantic processes such as lexico-semantic retrieval and semantic integration driven by an increase in working memory load.

Task focus, word processing, and the alpha-band. Pertaining to psycholinguistics, decreases in alpha-band activity (8-12 Hz) have been shown in response to semantic processing (Klimesch, Doppelmayr, Pachinger, & Russegger, 1997), how word-like a stimulus is (Strauß,

Kotz, Scharinger, & Obleser, 2014), and to phrase structure and number violations vs. control words (Davidson & Indefrey, 2007). More broadly, the alpha-band has been suggested to reflect attention to tasks (Klimesch, Sauseng, & Hanslmayr, 2007; Klimesch, 2012). The semantic and word processing alpha-band decreases were suggested to reflect integration processes by suppressing competing lexico-semantic representations. The syntactic alpha-band decrease was suggested to reflect attention and grammatical processing. These findings suggest that activity in the alpha-band relates to attentional processes during language comprehension, by selecting language representations and suppressing competing representations.

Sentence-level representation maintenance and the beta-band. Beta-band (13-30 Hz) decreases have been found for words that are incongruent to their context compared to controls (Wang et al., 2012a), for category violations vs. controls, random word strings vs controls, and random word strings vs. category violations (Bastiaansen, Magyari, & Hagoort, 2010). In contrast, beta-band increases have been found in response to syntactic complexity, such as complex object-relative clauses vs. less complex subject-relative clauses (Bastiaansen & Hagoort, 2006), and preceding low vs. high probability words (Magyari, Bastiaansen, Ruiter, & Levinson, 2014). Additionally, beta-band activity has been shown to increase over the course of a sentence (Bastiaansen et al., 2010). These results have collectively been explained by a neural prediction and sentence-level representation maintenance account of the beta-band (Lewis & Bastiaansen, 2015; Lewis, Schoffelen, Schriefers, & Bastiaansen, 2016). According to the predictive perspective of beta-band oscillations, beta-band activity corresponds to the buildup of a neurocognitive network responsible for the sentence-level representation. When encountering language violations vs. controls, the beta-band decreases in the prior literature have been interpreted as reanalysis of the sentence-level representation. During sentence reanalysis, the

neurocognitive network that was being maintained for the prior sentence-level representation is disengaged, leading to the beta-band decrease. The beta-band increase over the course of the sentence was suggested to reflect a buildup of the neurocognitive network, and the increase to unpredictable vs. predictable words was suggested to reflect ambiguity in the length of the text, and therefore increased load on working memory to maintain the sentence-level representation.

Linguistic coherence and the gamma-band. The lower gamma-band (~35-45 Hz) was initially suggested to reflect semantic unification processes (Hald et al., 2006; Bastiaansen et al., 2010; Wang, Zhu, & Bastiaansen, 2012b). More recently, these results have been reinterpreted as reflecting prediction at the level of lexical-representation, or linguistic coherence (Lewis & Bastiaansen, 2015). Rather than facilitating integration between a word and the semantic representation built up from the preceding context, increases in lower gamma-band activity may reflect a predictive or matching mechanism. In other words, pre-activated lexical-representations derived from the context are compared to incoming linguistic stimuli, and if the bottom-up linguistic input is coherent given the predicted lexical-representation, this leads to a top-down synchronization effect of the neurons engaged in the lexical-representation, and lateral inhibition on competing neurons, ultimately passing along the predicted representation to higher levels of the cortical hierarchy (Lewis & Bastiaansen, 2015; 'match-and-utilization' model from Herrman, Munk, & Engel, 2004).

Global knowledge and the gamma-band. Comparing control sentences such as "The Dutch trains are yellow…", and world knowledge violations such as "The Dutch trains are white…", gamma-band increases were found to world knowledge violations (Hagoort, Hald, Bastiaansen, & Petersson, 2004). Data for this experiment were collected from Dutch participants, who were aware that Dutch trains are yellow, not white, so the violation in the

sentence “The Dutch trains are white…” is being driven by specific knowledge about Dutch trains, as opposed to being a language violation per se. Another study found a gamma-band increase to words embedded in linguistically well-formed ironic contexts vs controls, which they related to this world knowledge violation effect (Spotorno, Cheylus, Van Der Henst, & Noveck, 2013). In this case as well, the violation is not of the language per se, but of global knowledge pertaining to the meaning of the text. This interpretation of the lower gamma-band has been somewhat controversial. From a semantic unification account, these increases would reflect increased difficulty in semantic unification, as opposed to abandonment of the sentence-level representation in the case of semantic violations. However, given that lower gamma-band increases have not been found to words with low vs. high predictability (Wang et al., 2012a; Wang et al., 2012b), these two accounts of the lower gamma-band cannot be straightforwardly reconciled in this way. Whereas the lower gamma-band language coherence effect is generally found early in the sentence (onset latency <200ms post-stimulus), the world knowledge violation effect occurs later (onset latency ~400ms post-stimulus), so these mechanisms may not be mutually exclusive. It is plausible that they represent related neurocognitive processes, or that they represent separate functional cortical networks which operate on the same frequency-band but at different times during on-line language comprehension.

Fixation-related potentials

While the goal of the present study was to understand the neurocognitive mechanisms engaged during naturalistic free-reading using TFRs, given that much of the prior psycholinguistics work has focused on ERPs, we also wanted to examine the effects of words with low vs. high contextual fit on the ERPs.

N400. The N400 is a negative-going ERP which occurs ~400ms post-word onset, and has been found for words which are semantically unrelated to their contexts (Kutas & Hillyard, 1984; Kutas & Federmeier, 2011). It has also been suggested that the N400 is sensitive to compositional semantics driven by linguistic structure building (Kim & Osterhout, 2005; Kim, Oines, & Sikos, 2016). Whereas experimentally manipulated, isolated sentences in the prior literature often control for linguistic structure, this is not the case in naturalistic texts.

P600. The P600 is a positive-going ERP which occurs ~600ms post-word onset, which has been suggested to reflect syntactic processing and linguistic structure building (Osterhout & Holcomb, 1992; Osterhout & Holcomb, 1994).

In the present study, our models of contextual fit are derived from computational word vector models which are thought to capture the lexico-semantic relatedness between words. For this reason, we expect to find N400 FRPs to words with low vs. high contextual fit for both local and topic contextual fit. However, systematic differences between the linguistic structure of words with low vs. high contextual fit could also lead to P600s. Additionally, experimental work manipulating word animacy in short stories which give animacy to normally inanimate objects found N400s to animacy violations early in the stories, but not later in the stories (Nieuwland & Van Berkum, 2006). Note that it appears they also found a P600, but did not report this finding. This suggests that over the course of the story, the subjects grew to associate the object with animacy given the story, even though it is a language violation. In other words, this suggests that topic context and local context both affect language comprehension, and that sufficiently high contextual fit in one domain can override unexpectedness in the other domain. Importantly, these stories were intentionally designed to maximize the ERP by experimentally manipulating animacy. It is plausible that in naturalistic texts, where there are no language violations and

where local and topic context have not been designed to be in opposition, these two sources of contextual information will interact differently.

Much of the prior research on both the ERPs and neural oscillatory dynamics of language processing involved subjects reading one word at a time at a controlled rate, and without the ability to move freely through the text. It is plausible that this artificial rate of presentation puts greater demand on attention, memory, and executive function systems, influencing functional cortical network engagement. We tested two frequency-bands which may relate to working memory processes, the frontal theta-band increase (Hald et al., 2006) and the beta-band increase (Lewis & Bastiaansen, 2015; Lewis et al., 2016). By only seeing one word at a time and not being able to move freely through the context, frequency-band effects corresponding to working memory processes pertaining to language comprehension may be exaggerated in the prior experimental literature. The functional cortical network corresponding to the alpha-band, which has been suggested to reflect task-related attention (Klimesch, 2012), may also operate differently during free-reading. Readers may engage in task reset or wrap-up processes between experimental stimuli or trials which operate on the alpha-band, but these processes may not be systematically engaged to each fixation during free-reading. Likewise, while subjects were asked comprehension questions between stories, there were no task interruptions within stories, as opposed to psycholinguistics experiments which often intersperse tasks between trials. Therefore, to understand how these oscillatory dynamics generalize to naturalistic reading, it is important to understand the role of eye movements on language comprehension during free-reading.

Free-reading

Experimental design vs. free-reading. The time-course of language comprehension as established by eye movements during free-reading has been shown to be faster than the time-course of language comprehension established using ERPs (Sereno & Rayner, 2003). This may be due to the method of stimulus presentation used in many EEG studies; rapid serial visual presentation (RSVP), a method whereby words are presented one at a time, at the center of the screen, at a controlled rate. It is therefore plausible that our understanding of the time-course of the neurocognitive mechanisms engaged during reading is not representative of naturalistic reading. In addition to the time-course, the neurocognitive mechanisms engaged during RSVP may be different or behave differently than in naturalistic reading. As previously mentioned, functional cortical networks involved in working memory and task-related attentional processing may be activated more so for language comprehension during RSVP than during free-reading. It is plausible that our understanding of the neurocognitive mechanisms engaged during reading is at least partially driven by the idiosyncratic nature of RSVP as compared to naturalistic reading, and so it is important to understand the neural oscillatory dynamics of language comprehension during free-reading.

Eye movements during free-reading. As a means of constraining our interpretation of the EEG, we also examined effects of the eye movements. Eye movements have been shown to vary by properties of the language, including lexical properties as well as predictability (Kliegl, Grabner, Rolfs, & Engbert, 2004), and are thought to reflect language processing difficulty (Van Gompel, Pickering, & Traxler, 2001; Staub & Rayner, 2007).

Methodological considerations. One reason RSVP has traditionally been used with EEG studies is to control for noise in the EEG signal due to eye movements, however, recent advancements to statistically correct for these eye movement-related artifacts have made

coregistration of the EEG and eye movements viable. Another consideration for EEG and eye movement coregistration research is the effect of overlapping fixations on the ERP (Nikolaev, Meghanathan, & Van Leeuwen, 2016). While we matched for fixation duration across many of our analyses to account for systematic differences in fixation overlap between conditions, prior research would suggest that eye movements are predictive of language processing difficulty, and therefore we thought it important to also examine effects of the eye movements independent of the EEG.

EEG and eye movement coregistration. Coregistration has been used to reproduce psycholinguistic effects during free-reading of sentences with semantic violations (Metzner, Malsburg, Vasishth, & Rosler, 2017), as well as effects due to semantic predictability (Dimigen, Sommer, Hohlfeld, Jacobs, & Kliegl, 2011). Coregistration has also been used to show how eye movements may be used to constrain the interpretation of the EEG. Metzner et al. (2017) compared the FRPs to control words, semantic anomalies, and syntactic anomalies, and further by whether the target fixated words were followed by a forward or regressive saccade, the sharp eye movements which separate fixations. They found an FRP to both semantic and syntactic anomalies vs. controls, but this effect was only present for fixated words followed by regressive saccades, not forward saccades. This finding supports the interpretation of eye movements as reflecting processing difficulty or reanalysis, and shows how processing differences can occur word-by-word even when the linguistic properties of the words are similar.

Additionally, in one study examining the neural oscillatory dynamics of language processing during EEG and eye movement coregistration of experimentally-manipulated sentences, the authors failed to find a beta-band decrease for random word strings vs. control sentences (Vignali, Himmelstoss, Hawelka, Richlan, & Hutzler, 2016), which has been found in

the prior RSVP literature (Bastiaansen et al., 2010). They suggest that this replication failure between the original RSVP study and their coregistration study may be due to lower demand on working memory for free-reading than RSVP. While in RSVP subjects must maintain the context in working memory as the words are presented one-by-one, during free-reading subjects can return to prior words in the context at any time. This again stresses the importance of understanding the relationship between eye movements and the EEG response, and how the ability to move through the text freely may engage different neurocognitive mechanisms than those engaged during RSVP.

Present Research

Prior research suggests that context plays a role in language predictions, and furthermore that global topic-based knowledge may uniquely affect language comprehension during story reading as compared to local lexico-semantic context. However, it is less well understood to what extent context effects from experimental designs generalize to linguistically well-formed stories, where variation in the context is derived solely from natural variation in the language, and where the stories may provide strong predictive affordances due to topic knowledge pertaining to the stories. Although our goal was to understand the functional cortical networks contributing to naturalistic language comprehension, to test the generality of experimental context effects to naturalistic reading we also examined effects on the eye movements and on the FRPs. The first study tested the effects of the lexical properties and the measures of contextual fit on the eye movements, the second study tested the effects of local contextual fit on the FRPs and TFRs, and the third study tested the effects of topic contextual fit on the FRPs and TFRs.

Hypotheses

We make the following hypotheses for the TFR to words with low vs. high contextual fit, for both local and topic context. We predict two effects in the theta-band. The first is a posterior increase to words with low vs. high contextual fit, which we interpret as increased lexico-semantic retrieval and integration difficulty to words with low contextual fit. Second, we predict a frontal increase, which we interpret as an increased demand on working memory for lexico-semantic integration, which may relate to a general cognitive control mechanism. In the alpha-band we predict a decrease to low vs. high contextual fit reflecting attentional processes to the task of story reading. In the beta-band, we would interpret an increase as increased load on working memory to maintain the sentence-level representation, and a decrease as reflecting sentence-level representation reanalysis and a change to the neurocognitive network contributing to language comprehension. We discuss two accounts of the lower gamma-band, the linguistic coherence account and the global knowledge violation account. According to the linguistic coherence account, we would expect gamma-band decreases to low vs. high contextual fit driven by increased gamma-band activity to coherent, high contextual fit words. Given the global knowledge violation account, we would expect a lower gamma-band increase to words with low vs. high topic contextual fit, reflecting unexpectedness of the topic content of the low fit words. These two accounts are not necessarily mutually exclusive, as the coherence effect should occur early after fixation onset, whereas the effect of low topic contextual fit should occur later.

Methods

Participants

Data were collected from 26 participants, with a mean age of 21.19 (SD = 2.97), 13 males and 13 females from the University of Colorado Boulder community or the greater Boulder, Colorado area. 6 subjects were excluded from analyses due to obstructive noise in the EEG signal, or trackloss of the eyetracker, where eye movements were outside the range of the eyetracker or desynchronized. Subjects were all neurotypical and right-handed according to the Edinburgh Inventory (Oldfield, 1971). Subjects were compensated for their participation.

Measures

Eye movements, computational language models, and behaviorally-normed topic knowledge measures were used to characterize the neural oscillatory dynamics in the EEG during story reading. Eye movements are characterized in terms of first fixation duration, and regression-path time (see eye movements subsection of measures). Language measures include word length, and lexical frequency from the SUBTLEX corpus (Brysbaert & New, 2009), a large corpus of subtitles from movies. Local contextual fit was measured by an arithmetical operation of the cosine distance between the word vector for a fixated word and preceding words. Topic contextual fit was measured through a similar cosine distance metric informed by a behavioral norming study (see context subsection of measures).

Eye movements. Only first fixations on any instance of a word within the stories were analyzed in the present study. While a given word may appear multiple times within a story, each instance is considered a unique item, and the first fixation on each instance would be included in the analyses. The EEG data were matched for first fixation duration, a standard practice in EEG

and eye movement coregistration studies to account for systematic differences between conditions in fixation overlaps (Devillez et al., 2015; Nikolaev et al., 2016). However, given that eye movements have been shown to reflect language processing difficulty (Van Gompel et al., 2001; Kliegl et al., 2004; Staub & Rayner, 2007), we also examined effects of lexical properties and of contextual distance on the eye movements. Specifically, we examined first fixation durations, and regression-path time. This measure includes the first fixation duration, plus the duration for all subsequent regressive fixations (fixations going back to previous words), until the subject has forward saccaded from the target fixated word. While first fixation duration may be a better predictor of first-pass language processing, regression-path time may be more reflective of deeper language comprehension (Staub & Rayner, 2007). Because regression-path time includes the duration of the first fixation and all regressive fixations on the current or previous words until forward saccading, this measure is thought to increase sensitivity to processing difficulty as reflected by regressive eye movements.

Lexical properties. The stories were not controlled for lexical properties such as word length or lexical frequency, however in some of the analyses these variables were matched by condition. For lexical frequency, a log-scaled version of the SUBTLEX movie subtitle corpus was used (Brysbaert & New, 2009).

Context. We measured the relationship between fixated words to their preceding contexts (local context) and to the topic of the story (topic context), as measures of contextual fit. These measures were derived computationally using word vectors trained on a large corpus using the global vector (glove) system (Pennington, Socher, & Manning, 2014), where the relationship between words within the 50-dimensional word vector space is thought to reflect the lexico-semantic relationship between those words.

In the glove system, the cosine distance between word vectors are thought to represent similarity between those words, where similarity is inversely proportional to cosine distance. However, the word vector space is computed such that more complex relationships between words can be represented through various arithmetical operations. For instance, the vector for queen can be estimated by taking the difference of the vectors of king and woman. In other words, beyond encoding the similarity between e.g. man and woman, woman and queen, etc., the model is thought to also encode the shared relational structure between e.g. man and woman as compared to king and queen. See table 1 for an example of local and topic contextual fit across sequential words within a subset of one story.

For local contextual fit, the cosine distance between the word vector of each fixated word and each of the prior 10 words to the target fixated word were calculated and summed (for content words, N = 1951, M = 4.46, SD = 1.53). The higher the summed cosine distance between a given fixated word and its prior local context, the less similar it is to the words in its context, or in other words the lower its contextual fit. A similar method has been implemented in prior research (Elvevåg, Foltz, Weinberger, & Goldberg, 2007). They used averaged cosine similarity between word vectors derived from latent semantic analysis as one measure of what they called coherence, but is similar to what we call contextual fit in the present study. In other words, local contextual fit measures the relationship between a fixated word and preceding words, independent of pragmatic information, most notably topic context.

A measure of topic contextual fit was produced by taking the summed cosine distance between a fixated word and each topic word from a behavioral norming study, weighted by frequency in the norming study and z-scored within story (for content words, N = 1956, M = 0.01, SD = 1.5). In the behavioral norming study, subjects read the stories used in the EEG and

eye movement coregistration study, and chose five or more words which they thought best represented the topic of the stories. They were also told that the topic words did not need to be present within the story. This was a preliminary measure with a small sample size (N = 14). The first five participants read all stories, while subsequent participants read only four stories each, randomly but evenly distributed across subjects. Additionally, these word selections were in some cases manually pruned, for example, “cross-pollination” and “pollination” were consolidated into just “pollination”. For each story, several words were chosen which accounted for ~40% of all submissions for that story (~3-5 words per story). The cosine distance between each fixated word and each of the chosen topic words for a given story were weighted by the proportion for each topic word within the story, summed, and z-scored for each story to normalize across all stories for different number of topic words and weightings. For instance, in a story about the evolution of whales, the topic words were “whale”, “evolution”, and “adaptation”. The word “whale” accounted for 17% of the words submitted for that story, “evolution” also accounted for 17%, and “adaptation” accounted for 8%, totaling 42%. The cosine distance between each fixated word in the story about the evolution of whales and the word “whale”, “evolution”, and “adaptation” were multiplied by 17, 17, and 8, respectively, summed, and then all summed contextual fit scores within the story about the evolution of whales were z-scored. Whereas local contextual fit measures the lexico-semantic relationship between fixated words and their preceding contexts, topic contextual fit measures the lexico-semantic relationship between fixated words and the topic, independent of local linguistic constraints. In other words, it serves as a measure of the pragmatic contextual cues provided by global knowledge of the story topic.

Table 1

A subset of words from the practice story about fruits

Word	Local contextual fit	Topic contextual fit
'fruits'	2.6	-2.01
'and'	6.09	-0.19
'vegetables'	3.56	-2.17
'appears'	4.91	0.42
'very'	5.59	-0.11
'clear'	5.12	0.61
'cut'	5.42	-0.53
'in'	6.31	0.06
'reality'	4.81	1.69
'though'	6.91	0.3
'there'	6.78	0.72
'are'	6.95	-0.04
'a'	6.54	-0.03
'few'	7.65	-0.1
'different'	7.17	0.19
'ways'	6.95	0.08
'in'	7.14	0.06
'which'	7.45	0.36
'fruits'	3.68	-2.01
'and'	7.77	-0.19
'vegetables'	4.33	-2.17
'can'	6.71	-0.13
'be'	6.99	0.35
'grouped'	4.15	0.8
'together'	6.59	-0.02
'this'	6.69	0.09
'debate'	4.07	1.47
'can'	6.46	-0.13
'actually'	6.05	0.1
'get'	6.02	0.32
'rather'	6.68	0.07
'heated'	3.94	0.67
'among'	6.02	0.43
'the'	6.37	0.51
'culinary'	2.26	0.6
'elite'	3.77	1.89

Note. Words in red have low contextual fit, words in blue have high contextual fit, and words in dark gray are function or auxiliary words. These stop words were included in the calculation of local contextual fit, but fixations on these words were not included in the analyses and were not included in the percentile bins determining low and high contextual fit.

Procedures

Nine stories were written at the 10th-grade Flesch-Kincaid grade level of readability (Flesch, 1948) by Shannon McKnight, a graduate student at the University of Colorado Boulder, for her research, currently in-preparation for publication. The topics of the stories could be characterized as the culinary distinction between fruits and vegetables, the evolution of whales, bees, eclipses, seasons, Louis Pasteur, ants, plate tectonics, and Marie Curie. All of the stories were intended to be about a scientific topic or the history of science. While wearing the EEG cap, subjects read each story page by page as the eye tracker recorded eye position and eye movements. There was no limit on reading time; subjects would press the space bar to proceed from page to page. The first story was excluded from analyses, and was considered a practice story. See *figure 1* for a single page from the experiment with simulated eye movements. Of the eight stories included in analyses, there was a mean of 5.25 pages per story (SD = 0.83), and a mean of 7.02 lines per page (SD = 1.25). Subjects were also asked true/false or multiple-choice comprehension questions after each story.

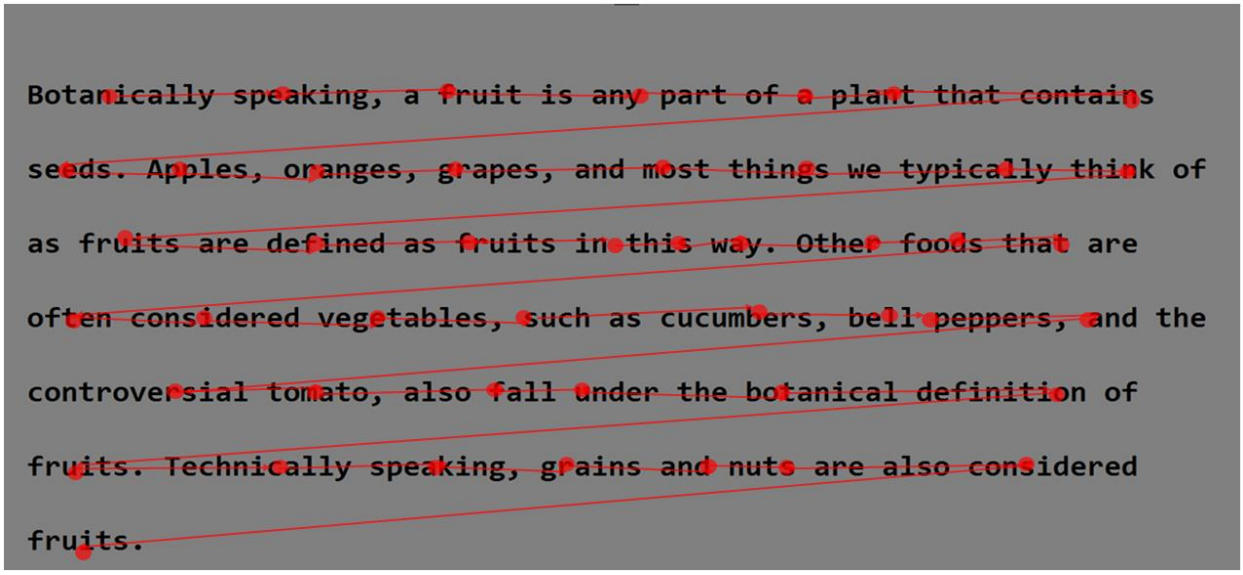

Figure 1. The first page of the practice story. The red dots are simulated examples of fixations, and the red lines are simulated examples of saccades.

Preprocessing and Coregistration. The EEG data were processed in matlab 2016b, scripted using the EEGLAB toolbox (DeLorme & Makeig, 2004), and the EYE-EEG plugin (Dimigen et al., 2011) to co-register the EEG and eye movement signals. The EEG signal was recorded using a neuroscan EEG amplifier at a sampling rate of 1000 Hz, and the tobii eye tracker recorded gaze samples at 333.33 Hz, which was upsampled to 1000 Hz during co-registration. Besides upsampling and aligning the two signals, linear regression was used to improve estimation of the latency of the start and end events, improving accuracy of the co-registration. Prior to preprocessing, electrodes visually identified as noisy were removed and replaced with data interpolated by neighboring channels (M = 3.8, SD = 5.32). The data were mastoid referenced and high-pass filtered to 1 Hz for preprocessing. Segments of continuous data with trackloss, as well as a 50ms window around these trackloss segments, were removed (M = 329524.3, SD = 61608.30) to avoid biasing independent components analysis (ICA), as trackloss and ocular artifacts may be correlated (Plöchl, Ossandón, & König, 2012). Visual inspection was used to remove noisy segments of continuous data (M = 591,395.8, SD = 2,341,652.64), and to derive cleaner ocular artifact independent components (ICs). Saccades were defined from an eye movement velocity factor of lambda = 3, with a minimum saccade duration of 15 ms, and a cluster distance of 25 samples, and these saccade clusters were used to define fixations. The EEGLAB runica method was utilized for ICA decomposition of the data, and ICs visually identified as containing mostly blink or saccadic activity were removed (M = 2.2, SD = 0.41). Manually removed ICs were compared to the pop_eyetrackerica function of the EYE-EEG plugin, which utilized a variance ratio criterion (Plöchl et al., 2012). The variance of the activity time course of each IC's saccade and fixation events were compared, and if the variance ratio

exceeded a certain threshold (in this case, the default 1.1 was used), that IC was removed from the data. The variance ratio criterion was only effective at removing saccadic ICs, but results were comparable to ICs manually removed for saccadic activity. All of the above cleaning procedures were applied to a backup of the original unfiltered data, which were then high-pass filtered to 0.1 Hz and low-pass filtered to 70 Hz, using the default "basic FIR (finite impulse response) filter (new)" implementation of the pop_eegfiltnew function from EEGLAB. The high- and low-pass filtering were performed separately rather than as a band-pass filter, as suggested by prior research (Widmann, Schröger, & Maess, 2015). The data were epoched to each fixation, from -1000ms pre-fixation to 2000ms post-fixation. The large epoch duration was chosen to accommodate the time-frequency tradeoff to measure neural oscillations at lower frequencies. Trials which exceeded a threshold of -130-130 μV were considered noisy and automatically rejected. Fixations less than or equal to 80ms or greater than or equal to 2000ms, fixations not on a word or on function words, repeat fixations on an instance of a word, fixations on words which were not included in the various language measures corpi, and fixations on words where the context measures could not be computed were not included in analyses.

Conditions. Fixations were sorted into conditions comparing the top and bottom third of contextual fit (i.e. low vs. high contextual fit) in separate procedures for local and topic contextual fit. Across conditions, fixations were matched for fixation duration for one set of analyses, and for fixation duration, word length, and lexical frequency for another set of analyses.

For matching fixation duration, a distribution of fixation durations discretized into 15 bins was created for each condition. If the distributions between the two conditions were unequal, a random fixation was removed from whichever condition had more fixations in the bin

with the greatest difference between conditions. This process was repeated until the discretized fixation duration distributions between the two conditions were equal. Matching by discretized distributions has been shown to control for effects of the matched variable, while retaining a greater number of fixations overall as compared to pairwise matching (Devillez, Guyader, & Guerin-Dugue, 2015).

For analyses matched across fixation duration, lexical frequency, and word length, 15 univariate bins were defined for fixation duration, 10 bins for lexical frequency, and 10 bins for word length, based on visual inspection. A multivariate distribution across all matching parameters was created for each condition, where the dimensions of the distribution were determined from the bins of the univariate distributions. The multivariate distributions were matched between conditions until exactly equal, in the same manner as described previously. This multivariate distribution approach matches not just for the discretized variables, but also for the interactions between them.

Fixation-related potentials. FRPs, equivalent to ERPs in an RSVP paradigm but time-locked to fixations, were calculated by re-epoching the EEG data to 200ms pre-fixation onset to 1000ms post-fixation onset, and fixations within each condition were averaged for each subject. The conditions were baselined separately from 200ms pre-fixation onset to 100ms pre-fixation onset. This window was chosen as opposed to a window enveloping up to fixation onset in order to avoid artifacts driven by fixation onset.

Time-frequency analysis. TFRs were produced at the single-trial level and averaged for each subject using the fieldtrip toolbox (Oostenveld, Fries, Maris, & Schoffelen, 2011). Frequency-band power was binned linearly from 4 to 70 Hz in 1 Hz bins, and then convolved with seven cycle Morlet wavelets to produce power estimates at each sample-point x frequency-

bin combination. Time-frequency power was baselined by the mean for each condition at each frequency bin, and converted to units of decibels.

Study 1: Eye movements

Eye movements have been used to measure reading comprehension difficulty driven by lexical properties and word predictability (Van Gompel et al., 2001; Kliegl et al., 2004; Staub & Rayner, 2007). Given that the ability to freely navigate through the context may affect the functional cortical network activation for language comprehension, it is important to understand how eye movements vary based on properties of the text. Specifically, for lexical properties, we expect decreases in lexical frequency and increases in word length to lead to increases in eye movement durations, as suggested by prior research, corresponding to differences in language comprehension difficulty. Additionally, if our measures of local and topic context contributed to language comprehension, this should also be reflected by increases in eye movement durations to low vs. high contextual fit (as characterized by cosine distance). We tested effects of lexical properties and context on first fixation duration and regression-path time, a measure which includes first fixation duration and all subsequent fixations prior to saccading forward in the story.

Statistical Analyses

To test effects of lexical properties and contextual fit on the eye movements, Pearson correlations and multiple regression were used. Analyses included all first fixations on content words for each subject. Eye movements were z-scored for each subject to account for subject-level variance. For correlation analyses, Pearson r-statistics, p-values, and β-coefficients derived from linear models are reported. For multiple regression, omnibus F-statistics, p-values, and adjusted R^2-statistics are reported, as well as the t-statistics and p-values for main effects and interactions, and β-coefficients for effects which are statistically significant at a threshold of $\alpha =$

0.05. Because our contextual fit measures are actually cosine distances, in this study we refer to contextual fit as contextual distance for ease of interpretation.

Results and Discussion

First fixation duration

34030 first fixations on content words were included in these analyses. To account for the non-normality of the first fixation durations, first fixation duration (in ms) was log-transformed (M = 5.32, SD = 0.49). Additionally, to account for between-subject variance, the log-transformed first fixation durations were also z-scored for each subject, such that the first fixation duration measure analyzed for each fixation is the log-transformed duration in units of standard deviations within the subject. To avoid overfitting of the multiple regression models, we tested the effects of lexical properties separately from the effects of the context measures on first fixation duration.

Lexical properties. In a model including lexical frequency, word length, and the frequency:length interaction on first fixation duration ($F(3, 34026) = 134$, $p < 0.001$, $R^2 = 0.01$, see table 2), there was a significant main effect of lexical frequency ($t(34026) = -8.69$, $p < 0.001$, $\beta = -0.02$), but not for word length ($t(34026) = 1.00$, $p = 0.32$) or the frequency:length interaction ($t(34026) = 1.73$, $p = 0.08$). As lexical frequency increased, first fixation duration decreased, consistent with our prediction that high frequency words would be less difficult to comprehend than low frequency words. Additionally, because lexical frequency and word length are strongly correlated, and given how little variance in first fixation duration is explained by the full model, the failure to find an effect of word length or a frequency:length interaction suggests that lexical frequency out-competes word length to explain mostly the same variance.

Table 2
Study 1, first fixation duration regressed on lexical properties

Predictor	t	p	β (if $p < 0.05$)
Lexical Frequency	-8.69	< 0.001	-0.01
Word Length	1.00	0.32	NA
Frequency:Length	1.73	0.08	NA

Note. β-coefficients only reported for significant effects.

Contextual distance. There was a significant relationship between first fixation duration and local contextual distance ($r = -0.07$, $p < 0.001$, $\beta = -0.05$), and between first fixation duration and topic contextual distance ($r = 0.03$, $p < 0.001$, $\beta = 0.03$), see table 3. In other words, as local contextual distance increases, first fixation durations decreased, and as topic contextual distance increases, first fixation durations increased. In contrast to our predictions, this suggests that words which were further from their local context were easier to comprehend than words which were closer to their local context. However, the relationship between topic contextual distance and first fixation duration was in the expected direction- the further a word was from its topic context, the more difficult it was to comprehend. One possible explanation for this finding could be that topic contextual distance was more predictive of language comprehension difficulty in these stories than local contextual distance. However, in a model containing both local and topic contextual distance and the local:topic contextual distance interaction as predictors of first fixation duration (($F(3,34026) = 56.8$, $p < 0.001$, $R^2 = 0.005$, see table 4), only the effect of local contextual distance was significant ($t(34026) = -11.63$, $p < 0.001$), but not topic contextual distance ($\beta = -0.05$) or the local:topic distance interaction ($t(34026) = 1.55$, $p = 0.12$; $t(34026) = 0.15$, $p = 0.88$). Alternatively, first fixation duration may not be an accurate measure of comprehension difficulty driven by contextual distance. Regression-path time, which includes regressive eye movements in the duration, may be a more accurate measure of language comprehension difficulty driven by contextual distance.

Table 3
Study 1, correlations between local and topic contextual distance and first fixation duration

Measure	Local contextual distance	Topic contextual distance
1. First fixation duration	-.30***	.64***

Note. * $p < .05$. ** $p < .01$. *** $p < .001$.

Table 4
Study 1, first fixation duration regressed on local and topic contextual distance and the interaction

Predictor	t	p	β (if p < 0.05)
Local contextual distance	-11.63	< 0.001	-0.05
Topic contextual distance	1.55	0.12	NA
Local:Topic	0.15	0.88	NA

Note. β-coefficients only reported for significant effects.

Regression-path time

For the same set of content words, regression-path time was calculated, log-transformed (M = 5.44, SD = 1.64) and z-scored for each subject.

Lexical properties. From a model regressing lexical properties on regression-path time ($F(3, 34026) = 127$, $p < 0.001$, $R^2 = 0.01$, see table 5), there was a significant main effect of lexical frequency ($t(34026) = -6.84$, $p < 0.001$, $\beta = -0.02$) and word length ($t(34026) = 3.56$, $p < 0.001$, $\beta = 0.01$), and a marginally significant frequency:length interaction ($t(34026) = 1.95$, $p = 0.051$, $\beta = 0.0008$). Given the similarity between the effects of the full models of the lexical effects on first fixation duration and on regression-path time, but the greater number of significant effects of the individual predictors for regression-path time, this may suggest that regression-path time is a somewhat more accurate measure of language comprehension difficulty than first fixation duration.

Table 5
Study 1, regression-path time regressed on lexical properties

Predictor	t	p	β (if p < 0.05)

Lexical Frequency	-6.84	< 0.001	-0.02
Word Length	3.56	< 0.001	0.01
Frequency:Length	1.95	0.051	NA

Note. β-coefficients only reported for significant effects.

Contextual distance. There was a significant relationship between regression-path time and local contextual distance ($r = -0.06$, $p < 0.001$, $\beta = -0.05$), and between regression-path time and topic contextual distance ($r = 0.01$, $p < 0.01$, $\beta = 0.01$), see table 6. In a model containing both local and topic contextual distance and the local:topic contextual distance interaction as predictors of regression-path time ($F(3,34026) = 47.7$, $p < 0.001$, $R^2 = 0.004$, see table 7), only the effect of local contextual distance was significant ($t(34026) = -11.06$, $p < 0.001$, $\beta = -0.05$); topic contextual distance and the local:topic contextual distance interaction were not significant $t(34026) = 1.58$, $p = 0.11$; $t(34026) = -1.01$, $p = 0.31$). These results are similar to those of first fixation duration, and do not elucidate the unexpected effect of local contextual distance. Additionally, local and topic contextual distance were not correlated (see supplemental material), so these results cannot be explained by co-linearity. However, variation in on-line comprehension as reflected by the FRPs and TFRs to the context measures may shed further light on these results.

Table 6
Study 1, correlations between local and topic contextual distance and regression-path time

Measure	Local contextual distance	Topic contextual distance
1. Regression-path time	-.06***	.01**

Note. * $p < .05$. ** $p < .01$. *** $p < .001$.

Table 7
Study 1, regression-path time regressed on local and topic contextual distance and the interaction

Predictor	t	p	β (if p < 0.05)

Local contextual distance	-11.06	< 0.001	-0.05
Topic contextual distance	1.58	0.11	NA
Local:Topic	01.01	0.31	NA

Note. β-coefficients only reported for significant effects.

Study 2: Effects of local contextual fit on the EEG

While eye movements may reflect overall differences in language processing difficulty, it is difficult to make inferences about the on-line processes engaged after encountering a word with low vs. high contextual fit, only the result. It is also not obvious from the eye movements which neurocognitive mechanisms are engaged. On the other hand, differences in the amplitude and latency of the FRPs may provide further insights into the time course of on-line language comprehension, and TFRs may provide further insights into the neurocognitive mechanisms involved.

For the FRPs, we expected to find an N400, although plausibly also a P600, to words with low vs. high local contextual fit. In this case, the N400 would reflect contextual fit driven by lexico-semantic relatedness, and the P600 could reflect systematic differences in linguistic structure to low vs. high fit words, or sentence reanalysis.

In terms of the neural oscillations, we expected effects in the theta-, alpha-, beta-, and lower gamma-band. In the theta-band we expected an increase at posterior electrodes reflecting lexico-semantic integration difficulty between low contextual fit words and the lexico-semantic retrieval network, and an increase at frontal electrodes reflecting increased demand on working memory due to increased lexico-semantic integration difficulty. In the alpha-band, we expected a decrease reflecting increased attention to reading due to semantic integration difficulty. In the beta-band, we expected either a decrease reflecting sentence-level representation reanalysis driven by low contextual fit words, or an increase reflecting increased demand on working memory to maintain the neurocognitive network, where low contextual fit words lead to added complexity in the sentence-level representation. Given the interpretation of the lower gamma-

band as reflecting linguistic coherence, we expected a gamma-band decrease to low vs. high contextual fit, driven by a gamma-band increase for words with high contextual fit.

Fixations were either matched for fixation duration between conditions, or matched for fixation duration, lexical frequency, and word length between conditions. There were few cases in which matching for lexical frequency and word length affected the statistical clusters, so results and figures will refer to only effects matched for fixation duration unless noted otherwise.

Statistical Analyses

Given the naturalistic and exploratory nature of the present study, nonparametric cluster-based permutation testing (Maris & Oostenveld, 2007) was used to test for significant FRP and frequency-band TFR effects across time and channels, to account for the type-1 error rate and family-wise error rate of these multiple comparisons. For the frequency-band statistics, the TFRs were averaged over the frequency-bands of interest, the theta-band (4-7 Hz), alpha-band (8-12 Hz), beta-band (13-30 Hz) and lower gamma-band (35-45 Hz). The design of the analysis takes subject-level averaged TFRs as observations. Using monte carlo permutation, the observations were resampled 1000 times and randomly assigned a condition within-subject for each permutation, creating a simulated empirical distribution from the data. A dependent samples t-test was tested for each time x channel EEG sample, and if the t-value exceeded a critical threshold derived from the simulated empirical distribution, it was considered significant. Temporally or spatially adjacent significant couplets form a cluster, and the sum of the t-values within the clusters were used as the cluster statistic. If there was a significant difference between conditions as reflected by a model which included the clusters which exceeded the cluster threshold vs. a model which did not include the clusters, then the clusters may be used to

describe the significant difference. In sum, this allows for the inference of frequency-band power across the whole epoch and the whole scalp in a way that avoids the multiple comparison problem and family-wise error rate without overcorrecting, as may be the case with Bonferroni post-hoc correction (Maris & Oostenveld, 2007). P-values for clusters which exceed the cluster threshold of $\alpha = 0.05$ (or 0.025 in each tail, where clusters are separated as positive and negative and therefore only encompass one tail) in a significant cluster permutation test are reported.

Results and Discussion

FRPs. There were no significant effects of low vs. high local contextual fit on the FRPs. It is possible that the ICA-based saccade correction reduced the amplitude of the EEG signal such that these relatively small effects were no longer significant (see limitations).

TFRs.

Theta-band (4-7 Hz). From 600-1130ms post-fixation onset over central and right-parietal electrodes there was a theta-band increase ($p < 0.01$), see *figure 2*. This result was predicted, and is consistent with prior research, where theta-band increases have been found for semantic violations vs. controls at posterior electrodes (Hagoort et al., 2004; Hald et al., 2006; Davidson & Indefrey, 2007) and has been suggested to reflect lexico-semantic retrieval (Bastiaansen et al., 2005). While this increase appears over centro-parietal electrodes, given the late onset latency of the effect, it is possible this effect instead reflects the increased demand on working memory for semantic integration, despite typically being associated with activity at frontal electrodes (Hald et al., 2006; Rommers et al., 2017).

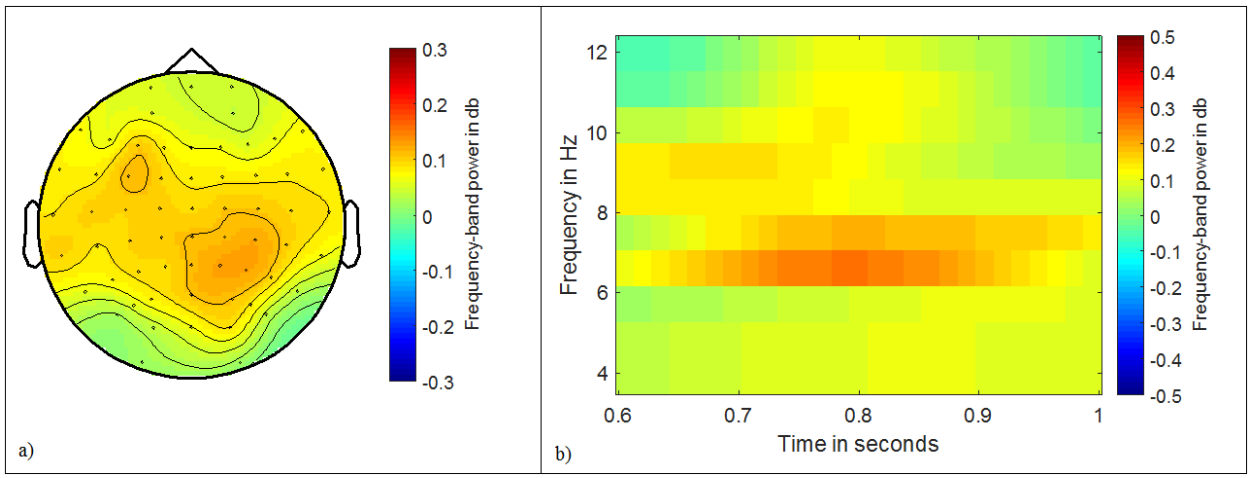

Figure 2. (a) Topographic plot of the theta-band (4-7 Hz) increase to low vs. high local contextual fit from 600-1130ms post-fixation onset. (b) TFR averaged over centro-parietal electrodes representing the theta-band increase. Given the time window, this theta-band increase may reflect increased working memory demand for semantic integration of words with low local contextual fit. However, the working memory-related theta-band effects are usually more frontal than the present finding, and so this may instead reflect lexico-semantic retrieval.

Beta-band (13-30 Hz). From 155-380ms post-fixation onset over frontal-midline and right parietal electrodes there was a beta-band increase ($p < 0.025$), see *figure 3*. We had predicted either a beta-band decrease reflecting deactivation of the neurocognitive network associated with the sentence-level representation, or in other words sentence reanalysis, or a beta-band increase reflecting the increased demand on working memory to maintain the sentence-level representation in the face of local lexico-semantic contextual complexity. This finding supports the maintenance account, suggesting an increase in working memory load during reading of words with low local contextual fit.

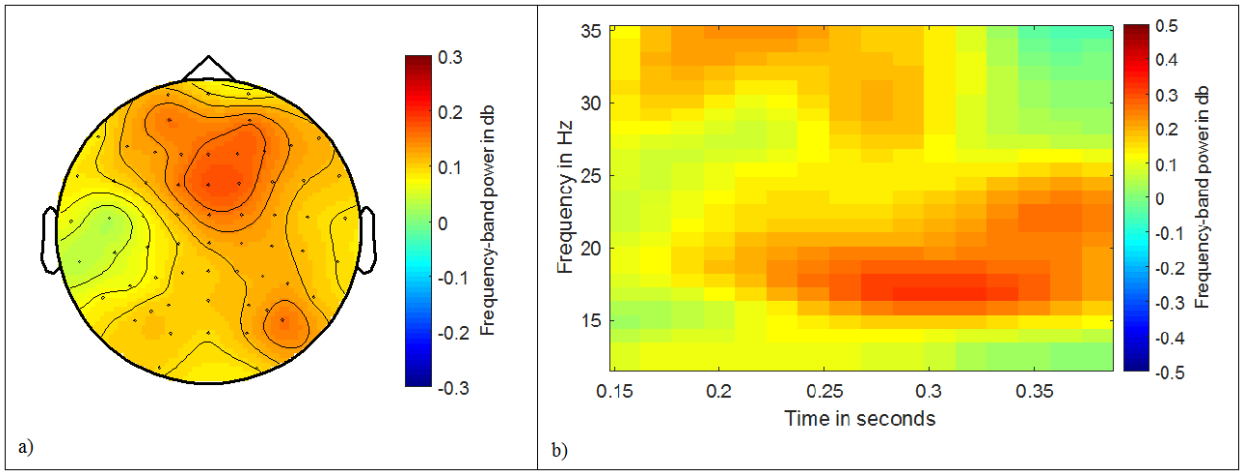

Figure 3. (a) Topographic plot of the beta-band (13-30 Hz) increase to low vs. high local contextual fit from 155-380ms post-fixation onset. Despite what appears to be a secondary right-posterior peak, this effect is the result of a single cluster. (b) TFR averaged over frontal-midline electrodes representing the beta-band increase. This beta-band increase may reflect an increased demand on working memory to maintain the sentence-level representation to low local contextual fit words, driven by lexico-semantic complexity.

Lower gamma-band (35-45 Hz). From 170-305ms broadly distributed across electrodes there was a lower gamma-band increase ($p < 0.001$), see *figure 4*. However, this effect was not present when additionally matched for lexical frequency and word length. This effect was not predicted, and given that it was not present in both analyses, this effect may be driven by the lexical properties of the language which systematically vary between conditions of low vs. high local contextual fit, rather than being an effect of local contextual fit per se.

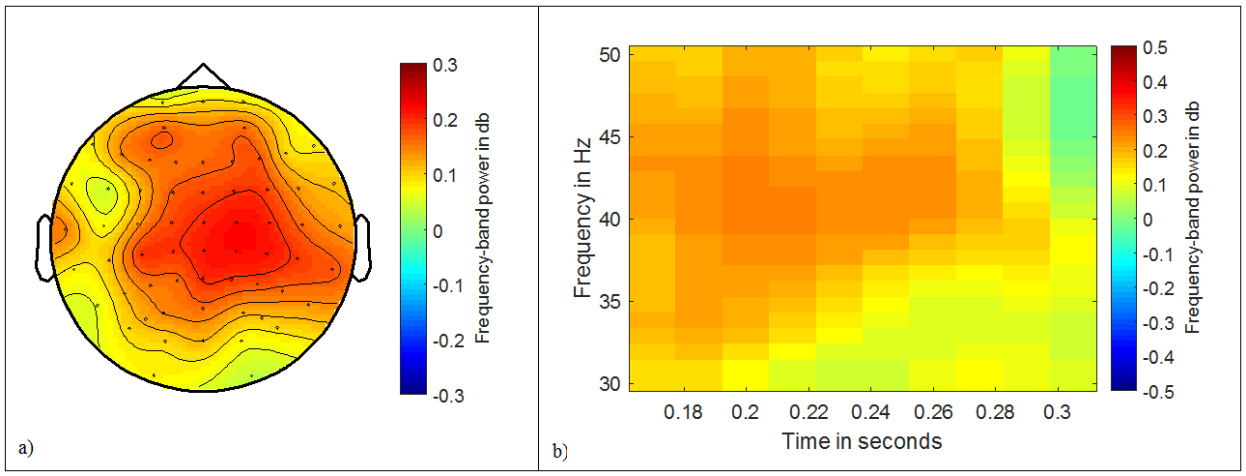

Figure 4. (a) Topographic plot of the lower gamma-band (35-45 Hz) increase to low vs. high local contextual fit from 170-305ms post-fixation onset. (b) TFR averaged over fronto-centro-parietal electrodes representing the lower gamma-band increase. This lower gamma-band increase was not predicted, and is not present when low vs. high local contextual fit are matched for lexical properties. This lower gamma-band effect may be driven by systematic differences in the lexical properties of words with low vs. high local contextual fit, rather than local contextual fit per se.

The significant TFR effects of local contextual fit across multiple frequency bands, time windows, and electrode regions, suggests that local contextual fit affects language comprehension during naturalistic story reading, and that multiple neurocognitive mechanisms are engaged. These effects occurred even when fixations on words with low vs. high local contextual fit were matched for fixation duration, suggesting that late effects in the TFR epoch were not driven by systematic differences in fixation overlap. Additionally, the theta-band and beta-band effects were present when words with low vs. high local contextual fit were also matched for lexical frequency and word length, suggesting that these effects were not driven by lexical properties, but specifically by local contextual fit.

Study 3: Effects of topic contextual fit on the EEG

To dissociate effects of context driven by the lexico-semantic relatedness of a word to the preceding words as compared to effects of context driven by global knowledge, we also tested effects of topic contextual fit. Because much of the prior research has focused on experimentally manipulated words in isolated sentences, effects of topic context have been underexplored, and may represent an important difference in experimental vs. naturalistic language processing.

For the FRPs, we expected an N400 to words with low vs. high topic contextual fit. Being derived from the relationship between target fixated words and the behaviorally-normed story words, topic contextual fit should be independent of local context and reflect global knowledge (Hagoort et al., 2004; Nieuwland and Van Berkum, 2006).

In the TFRs, we expected the same theta-, alpha-, and beta-band responses as with local contextual fit; a posterior theta-band increase reflecting lexico-semantic retrieval processes, a frontal theta-band increase reflecting working memory during semantic integration, and either a beta-band decrease reflecting sentence reanalysis or a beta-band increase reflecting a working memory increase to maintain the sentence-level representation given increased linguistic complexity. However, whereas with local contextual fit we only expected a lower gamma-band decrease to low vs. high contextual fit, driven by an increase in lower gamma-band activity to words with high contextual fit reflecting linguistic coherence, for topic contextual fit, we also expected a global knowledge effect. Words with low topic contextual fit should show increased lower gamma-band activity vs. words with high topic contextual fit, reflecting topic knowledge-related comprehension difficulty.

Statistical Analyses

Exploratory cluster-based permutation analyses were used to test the effect of low vs. high topic contextual fit for FRPs and frequency-band effects of the TFRs. This implementation was identical to that used in Study 2.

Results and Discussion

FRPs. There was a significant negativity to low vs. high topic contextual fit, broadly distributed over the scalp, from 170-400ms post-fixation onset ($p < 0.01$), see *figure 5*. We interpret this as an N400 effect, which may suggest either prediction or semantic integration difficulty for low topic contextual fit words. This is consistent with the finding that world knowledge violations vs. controls produce N400 effects (Hagoort et al., 2004).

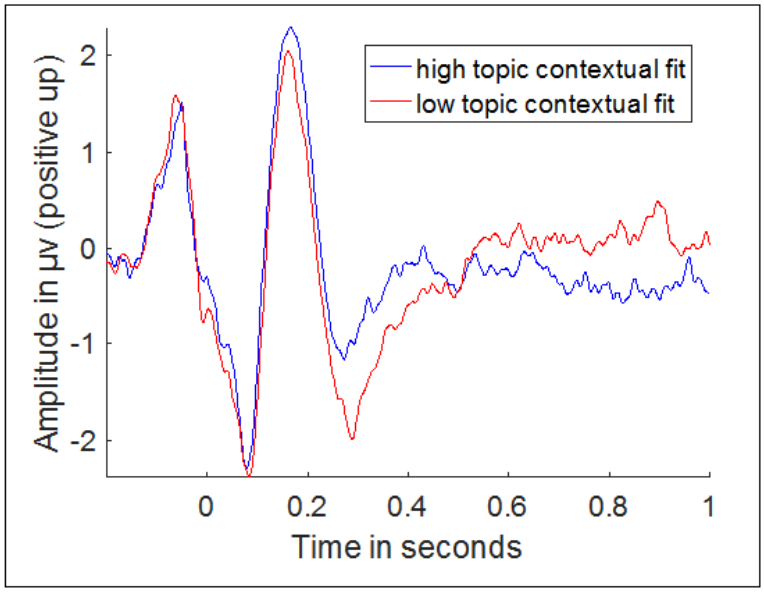

Figure 5. FRP of low vs. high topic contextual fit at electrode Cz. There was a significant negativity for words with low vs. high topic contextual fit from 170-400ms. We interpret this as an N400 effect, consistent with prior research showing N400s to world knowledge violations vs. controls. There is also an apparent P600, which would be consistent with prior research showing P600s to words in ironic contexts vs. controls, suggested as an effect of pragmatics or global knowledge, but this effect was not significant.

TFRs. In contrast to local contextual fit, we did not find any significant frequency-band effects of low vs. high topic contextual fit. Possible explanations for this will be discussed in the general discussion and limitations sections.

General Discussion

In the present study, we were interested in understanding the role of context on language comprehension during free-reading of naturalistic stories. Leveraging the information present in stories, we wanted to dissociate the role of local lexico-semantic information and topic knowledge on language comprehension, a distinction which has been under-explored in the prior literature. We tested these effects on the eye movements, the FRPs, and the TFRs, with the goal of understanding the functional cortical networks associated with natural language processing, and the extent to which they are similar to effects of experimental language manipulations.

Eye movements

Lexical properties. We found regression-path time to be somewhat more predictive of differences between the lexical properties and contextual distance of words than first fixation duration. For lexical properties, as lexical frequency increases, eye movement durations decreased, which we interpret as high frequency words being more quickly comprehended. Word length and the frequency:length interaction were not significant, suggesting that lexical frequency mediates word length. The finding that lexical properties contribute to eye movement duration is consistent with the prior research, and suggests that subjects read the stories in a typical manner.

Contextual distance. We were surprised to find that as local contextual distance increased, eye movement durations decreased. While we found the opposite (and expected) relationship between topic contextual distance and eye movement durations, this effect was not significant in a multiple regression model which included both local and topic contextual distance and the local:topic contextual distance interaction. This may suggest that the on-line

neurocognitive processes driven by contextual fit are not effectively captured by latency in eye movements, which we hoped to elucidate in the EEG analyses. Alternatively, it may be the case that when a word with low local contextual fit was encountered, rather than fixating for longer or regressing to previous words, readers instead saccade forward in the text for clarification.

Local contextual fit

FRPs. We did not find a difference in the FRPs for fixated words with low vs. high local contextual fit. For methodological reasons, we needed to control for the effect of eye movements on the EEG, so it is possible that in doing so we removed meaningful variance between words with low vs. high contextual fit. In fact, in one study, FRP effects of semantic and syntactic violations vs. controls were only present for fixations on violation words which were followed by regressive eye movements (Metzner et al., 2017). This suggests that only when the language violation led to comprehension difficulty (as predicted by the eye movements) were effects of the FRP present.

TFRs. We found several TFR effects, including a theta-band increase, beta-band increase, and lower gamma-band increase. It is possible that changes in frequency-band activity reflect differences in neurocognitive engagement to words with low vs. high local contextual fit which were not captured by the FRPs.

Theta-band. Increases in the theta-band (4-7 Hz) have been found over posterior electrodes (Hagoort et al., 2004; Hald et al., 2006; Davidson & Indefrey, 2007) and have been associated with lexico-semantic retrieval processes (Bastiaansen et al., 2005). Additionally, theta-band increases have also been found over frontal electrodes and have been associated with increased demand on working memory driven by semantic integration difficulty (Hald et al.,

2006; Rommers et al., 2017). This working memory effect may be part of a general conflict resolution mechanism driven by activity in prefrontal cortex recruited during language processing (Novick et al., 2005), further suggested by findings in executive function literature of a frontal theta-band increase associated with cognitive control (Cavanagh & Frank, 2014). In the present study, when comparing fixations on words with low vs. high local contextual fit, we found a theta-band increase over centro-parietal electrodes from 600-1130ms post-fixation onset (see *figure 2*). This was more posterior in terms of its scalp distribution than reported frontal theta-band effects but occurred later than reported posterior theta-band effects. Given that lexico-semantic retrieval is necessarily an early process for sentence-level language comprehension, the time window of this effect is more likely driven by a working memory load increase to semantic integration. This suggests that words with low local contextual fit imposed a greater working memory demand on comprehension, driven by semantic integration difficulty for the low-fit word.

Beta-band. Activity in the beta-band (13-30 Hz) during language comprehension has been suggested to reflect the maintenance of sentence-level representations and projection of these representations across hierarchical cortical networks recruited during language comprehension (Lewis & Bastiaansen, 2015; Lewis, Schoffelen, Schriefers, & Bastiaansen, 2016). According to this neural prediction perspective, increases in the beta-band correspond to increased syntactic complexity (Bastiaansen & Hagoort, 2006) or accruement of language context over the course of a sentence (Bastiaansen et al., 2010). Beta-band decreases, such as those found to words which were incongruent with their context compared to controls (Wang et al., 2012a), have been suggested to reflect disengagement of the preceding sentence-level representation and restructuring of the neurocognitive network driven by sentence reanalysis.

Therefore, if words with low contextual fit in our stories were more likely to lead to sentence reanalysis, we would expect a beta-band decrease, whereas if these words add complexity, but do not generally lead to sentence reanalysis, we would expect a beta-band increase to maintain the sentence-level representation in the face of increased complexity driven by low contextual fit. In the present study, when comparing fixations on words with low vs. high local contextual fit, we found a beta-band increase from 155-380ms post-fixation (see *figure 3*). This suggests that on average, words with low compared to high local contextual fit add linguistic complexity, without leading to disconfirmation of the sentence-level representation, and so readers maintain the neurocognitive network facilitating the sentence-level representation. If these stories contained intentional anomalies or greater linguistic complexity, it is plausible that we could instead have seen a beta-band decrease, reflecting sentence reanalysis. It is also plausible that certain words did elicit beta-band decreases driven by reanalysis, but that more often words with low local contextual fit drove complexity without requiring reanalysis. This effect did not change when the TFRs were matched for lexical frequency and word length, suggesting that differences in beta-band activity between words with low vs. high local contextual fit are not being driven by these lexical properties.

Lower gamma-band. There are two theories about activity in the lower gamma-band (35-45 Hz) which we discussed in the present study, a linguistic coherence account (Lewis & Bastiaansen, 2015), and a global knowledge violation account (Hagoort et al., 2004). We expected a lower gamma-band decrease for fixations on words with low vs. high local contextual fit, driven by an increase in gamma-band activity for words with high contextual fit, which are more linguistically coherent with their context. On the other hand, we expected a gamma-band increase to fixations on words with low vs. high topic contextual fit, reflecting recognition of low

contextual fit between the fixated words and the story topic, independent of the local context. In the present study we found a lower gamma-band increase to words with low vs. high local contextual fit from 170-305ms post-fixation onset over fronto-centro-parietal electrodes (see *figure 4*). However, this effect was opposite of what we predicted, and was the only TFR effect of low vs. high local contextual fit to not be significant when matched for lexical frequency and word length. This suggests that the lower gamma-band effect may be driven by systematic differences in the lexical properties between words with low vs. high local contextual fit. Additionally, this frequency-band was where we expected the most theoretically interesting differences between the effects of local and topic context, but we did not find any significant TFR effects of low vs. high topic contextual fit, meaning we cannot meaningfully compare the lower gamma-band activity between study 2 and study 3.

Alpha-band. Activity in the alpha-band (8-12 Hz) has been associated with semantic (Klimesch et al., 1997), lexical (Strauß et al., 2014), and syntactic processing (Davidson & Indefrey, 2007), and collectively these results have been suggested to reflect attentional processes during language tasks (Klimesch et al., 2007; Klimesch 2012). For these reasons, we expected an alpha-band decrease to words with low vs. high contextual fit, reflecting suppression of competing lexico-semantic representations for words which may have been pre-activated or would have higher contextual fit than the low contextual fit word they encountered. We did not find any alpha-band effects of contextual fit. If alpha-band activity reflects attentional processes, it is plausible that free-reading does not require task re-engagement in the same way as RSVP designs, where stimuli are separated by inter-stimulus intervals and trials by inter-trial intervals, which may put greater demand on task-representation. In other words, in RSVP designs, readers may engage certain neurocognitive mechanisms in anticipation of the trial- or even word-onset,

whereas the neurocognitive mechanisms behind free-reading may be more continuous, given the continuous nature of free-reading. However, we can only speculate from a null-result.

Topic contextual fit

FRPs. For topic contextual fit, we found an N400 FRP to words with low vs. high topic contextual fit. This suggests that global knowledge contributes to semantic processing during on-line language comprehension. However, it is not clear why FRP effects would be present when comparing low vs. high topic contextual fit, but not when comparing low vs. high local contextual fit. It is plausible that the FRP effects of topic contextual fit are more robust and therefore less susceptible to the effects of saccade correction, but then this does not explain why TFR effects were present for local contextual fit but not topic contextual fit.

TFRs. Surprisingly, there were no effects of topic contextual fit on the TFRs. The behavioral norming survey consisted of only 14 subjects, and only five subjects read all stories (the rest read only four stories each), so it is possible that the sample size for this survey was insufficient to derive reliable topic words. Additionally, the approach used to weight cosine distances and normalize summed cosine distances across stories may have been ineffective for measuring topic contextual fit. However, this would not explain why we found a significant N400 FRP to words with low vs. high topic contextual fit. It is also plausible that this discrepancy between local and topic contextual fit reflects some meaningful functional distinction between the FRP and neural oscillatory dynamics of naturalistic language comprehension. It is possible that in naturalistic stories, the topic context is so highly constrained that words which we described as having low and high fit varied little from each other, or in other words, the text is so constrained that topic contextual fit did not significantly affect the

neurocognitive mechanisms engaged for language comprehension as compared to local contextual fit. However, this still does not explain the significant N400 effect to words with low vs. high topic contextual fit.

Overall, these results could generally be interpreted as support for prior experimental psycholinguistic research on context in that we found differences between words with low vs. high local contextual fit across several of the predicted frequency-bands. However, there were some results which could not be easily explained, such as the failure to find FRPs for local contextual fit while finding TFRs, and the finding of FRPs for topic contextual fit but failure to find TFRs, as well as the opposite effects of local and topic contextual distance on eye movement durations. Further analyses or future research may clarify these discrepancies.

Limitations

Context measures. While our measure of local contextual fit is similar to measures used in prior research (Elvevåg et al., 2007; Paczynski & Kuperberg, 2012), and word vector-derived semantic similarity has been shown to be comparable to behaviorally-normed word predictability (Van Petten, 2014), it is possible that our measure of local contextual fit was in some way not appropriate for our stories. For topic contextual fit, our measure was weighted by results from a behavioral norming study, but this study included only 14 subjects, and of those 14, only five made selections for every story. The other nine subjects only made selections for four stories each. Given the small sample size, and the fact that the weighting of cosine distances contributing to the topic contextual fit measure was derived from the proportion of selections from this sample, our measure of topic contextual fit may be biased by idiosyncrasies in our sample or weighting implementation.

Saccade Correction. It is possible that our ICA-based implementation of saccade correction introduced artifacts in the EEG waveform, or reduced the amplitude in such a way as to minimize important differences between conditions in the FRPs. This may explain why we did not find effects of local contextual fit on the FRPs. In fact, a small difference between low vs. high contextual fit in the N400 time window was observed, but was not significant. While prior research has shown significant FRPs to word predictability (Dimigen et al., 2011; Metzner et al., 2017), in these cases the words being compared were in isolated sentences where word predictability was manipulated, whereas in the present study, the stories were not experimentally manipulated, and so even a word with relatively low contextual fit in the present stories may be more predictable than in the prior research, or alternatively, the topic context or lexical properties of the context may have provided predictive affordances which impacted the FRP. On the other hand, the TFRs, reflecting frequency-power within the EEG over sliding time windows, may be robust to the effects of saccade correction on the amplitude of the EEG waveform.

Baseline. Unlike RSVP designs, where the ERPs and TFRs can be baselined from an inter-trial interval, effects of fixation overlap both in the baseline window and late in the epoch need to be accounted for, and it is possible that our approach to circumventing these issues introduced artifacts in the data.

Exploratory Approach. There were no experimental manipulations in the study, nor were lexical properties such as lexical frequency or word length controlled for in the stories, meaning that there are potentially many confounding variables. The distributional approach to matching fixations between conditions by other variables has been empirically validated (Devillez et al, 2015; Nikolaev et al., 2016), but it is still plausible that the multivariate distributional matching approach used for fixation duration, lexical frequency, and word length in the present study was

insufficient, or that additional variables that were not matched were impacting the results. The EEG analyses were derived from nonparametric cluster-based permutation testing to find effects across time and channels. This method is useful for exploratory analyses in that it accounts for the family-wise error rate of so many multiple comparisons made in the EEG time series across adjacent timepoints and electrodes, however planned comparisons using parametric statistics of the EEG averaged over specific time windows and scalp regions may have produced more robust effects.

Future Directions

Model-based alternatives to distributional matching. The distributional matching approach was utilized to control for as many variables as possible while throwing out as little data as possible. While this approach has been implemented in the prior literature to control for eye movement effects (Devillez et al., 2015; Nikolaev et al., 2016), as more variables are accounted for in the distribution, more fixations must be removed to match the conditions. We found that anything more than three matching variables was not practical given the number of fixations which would have to be removed. An alternative approach would be to use general linear regression model estimates of the FRP or TFR controlling for these variables, such as the rERP framework (Smith & Kutas, 2015a; Smith & Kutas, 2015b) or using partial correlations (Dufau, Grainger, Midgley, Holcomb, 2015). This would allow for more variables to be accounted for, without sacrificing fixations.

Additional variables. Variables such as Orthographic Levenshtein Distance (OLD) and semantic concreteness have been shown to affect visual word processing (Dufau et al., 2015), as well as affecting the N400 in a multiple regression model including word-vector measures of

context (Van Petten, 2014). While it would not be reasonable to match for OLD and concreteness in addition to fixation duration, lexical frequency, and word length, these variables and others could be included in a GLM model estimate of the FRPs and TFRs. To account for overfitting, a penalization parameter such as a ridge regression lambda parameter could be used to up-weight more influential predictors and down-weight less influential predictors, allowing for many predictors to be included while minimizing overfitting.

Alternative approaches to contextual fit. It is plausible that our measures of contextual fit were inaccurate. For local context, it may be necessary to weight the impact of a word within a context by its distance to the target word. It may also be necessary to log-scale or in some other way regularize the cosine distances within the context so that abnormally low or abnormally high fit between a target word and one word within its context doesn't bias the contextual fit as determined by the entire 10-word preceding context. Alternatively, a prediction error-based updating model, such as the word2vec skipgram model (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Goldberg & Levy, 2014), may be more predictive as a measure of local contextual fit than the glove system. The glove system derives lexico-semantic relatedness from the co-occurrence statistics of words in documents, but does not account for linguistic structure such as word order, instead treating the documents as a "bag of words". However, from a compositional semantics perspective, this information may be important for understanding language prediction (Kim et al., 2016). For topic contextual fit, rather than deriving a measure from the glove system using behaviorally-normed topic words, a latent dirichlet allocation (LDA) probabilistic topic model could be used to create posterior probabilities for each word given each story (Griffiths & Steyvers, 2004; Griffiths, Steyvers, & Tenenbaum, 2007). While certain implicit assumptions are made for LDA models, such as the number of topics and the hyper-parameters which affect the

distributions for the probabilities of words given topics and topics given documents, given the appropriate assumptions, the probability of a target fixated word given the story it is present in may be a more empirically efficacious way of measuring topic contextual fit.

Time- and region-averaged trial-level EEG analyses. Rather than exploratory nonparametric cluster-based permutation testing, we could average over specific time windows and electrode regions of the EEG z-scored by subject, and regress these averaged responses at the trial-level, controlling for confounds statistically rather than matching. An alternative approach would be to use mixed effects models, allowing for subject-level random intercepts, rather than z-scoring the averaged EEG response. In the mixed effects models, words would not be treated as a random or fixed effect, since what we are interested in is understanding what properties predict comprehension, not differences in comprehension between the words per se. These approaches would be more powerful than the exploratory analyses, and in the latter case would more precisely account for individual differences between subjects.

Total-time of fixations. We found an unexpected relationship between local contextual distance and eye movement durations, where an increase in local contextual distance led to a decrease in both first fixation duration and regression-path time. This would suggest that when words are less related to their local context, they are comprehended more easily. However, it may be the case that rather than fixating longer or regressing to prior words in the context, instead readers saccade forward to collect more information. Total-time is a measure which includes regression-path time, as well as the duration of all subsequent fixations on an instance of a word, even after the forward saccade from the first fixation. If subjects forward saccade after encountering a difficult word for clarification and then regress back, this measure may be more indicative of the effect of local contextual distance on eye movements.

REFERENCES

Bastiaansen, M. C., Linden, M. V. D., Keurs, M. T., Dijkstra, T., & Hagoort, P. (2005). Theta responses are involved in lexical—Semantic retrieval during language processing. *Journal of cognitive neuroscience*, *17*(3), 530-541.

Bastiaansen, M., & Hagoort, P. (2006). Oscillatory neuronal dynamics during language comprehension. *Progress in brain research*, *159*, 179-196.

Bastiaansen, M., Magyari, L., & Hagoort, P. (2010). Syntactic unification operations are reflected in oscillatory dynamics during on-line sentence comprehension. *Journal of cognitive neuroscience*, *22*(7), 1333-1347.

Buzsáki, G., & Draguhn, A. (2004). Neuronal oscillations in cortical networks. *science*, *304*(5679), 1926-1929.

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.

Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in cognitive sciences*, *18*(8), 414-421.

Davidson, D. J., & Indefrey, P. (2007). An inverse relation between event-related and time–frequency violation responses in sentence processing. *Brain Research*, *1158*, 81-92.

Devillez, H., Guyader, N., & Guérin-Dugué, A. (2015). An eye fixation–related potentials analysis of the P300 potential for fixations onto a target object when exploring natural scenes. *Journal of vision*, *15*(13), 20-20.

Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., & Kliegl, R. (2011). Coregistration of eye movements and EEG in natural reading: analyses and review. *Journal of Experimental Psychology: General*, *140*(4), 552.

Dufau, S., Grainger, J., Midgley, K. J., & Holcomb, P. J. (2015). A thousand words are worth a picture: Snapshots of printed-word processing in an event-related potential megastudy. *Psychological science*, *26*(12), 1887-1897.

Elvevåg, B., Foltz, P. W., Weinberger, D. R., & Goldberg, T. E. (2007). Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia research*, *93*(1), 304-316.

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 540-551.

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, *32*(3), 221.

Goldberg, Y., & Levy, O. (2014). word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Griffiths, T., & Steyvers, M. (2004). Finding Scientific Topics. *Proceedings of the National Academy of Sciences, 101 (suppl. 1)*, 5228-5235.

Griffiths, T.L., Steyvers, M., & Tenenbaum, J.B.T. (2007). Topics in Semantic Representation. *Psychological Review*, 114(2), 211-244.

Hald, L. A., Bastiaansen, M. C., & Hagoort, P. (2006). EEG theta and gamma responses to semantic violations in online sentence processing. *Brain and language*, *96*(1), 90-105.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *science*, *304*(5669), 438-441.

Hasselmo, M. E., Bodelón, C., & Wyble, B. P. (2002). A proposed function for hippocampal theta rhythm: separate phases of encoding and retrieval enhance reversal of prior learning. *Neural computation*, *14*(4), 793-817.

Herrmann, C. S., Munk, M. H., & Engel, A. K. (2004). Cognitive functions of gamma-band activity: memory match and utilization. *Trends in cognitive sciences*, *8*(8), 347-355.

Kielar, A., Panamsky, L., Links, K. A., & Meltzer, J. A. (2015). Localization of electrophysiological responses to semantic and syntactic anomalies in language comprehension with MEG. *Neuroimage*, *105*, 507-524.

Kim, A. E., Oines, L. D., & Sikos, L. (2016). Prediction during sentence comprehension is more than a sum of lexical associations: the role of event knowledge. *Language, Cognition and Neuroscience*, *31*(5), 597-601.

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of memory and Language*, *52*(2), 205-225.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*(1-2), 262-284.

Klimesch, W., Doppelmayr, M., Pachinger, T., & Russegger, H. (1997). Event-related desynchronization in the alpha band and the processing of semantic information. *Cognitive Brain Research*, 6(2), 83-94.

Klimesch, W., Sauseng, P., & Hanslmayr, S. (2007). EEG alpha oscillations: the inhibition–timing hypothesis. *Brain research reviews*, *53*(1), 63-88.

Klimesch, W. (2012). Alpha-band oscillations, attention, and controlled access to stored information. *Trends in cognitive sciences*, *16*(12), 606-617.

Lewis, A. G., Schoffelen, J. M., Schriefers, H., & Bastiaansen, M. (2016). A predictive coding perspective on beta oscillations during sentence-level language comprehension. *Frontiers in human neuroscience*, *10*, 85.

Lewis, A. G., & Bastiaansen, M. (2015). A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex*, *68*, 155-168.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of neuroscience methods*, *164*(1), 177-190.

Magyari, L., Bastiaansen, M. C., de Ruiter, J. P., & Levinson, S. C. (2014). Early anticipation lies behind the speed of response in conversation. *Journal of Cognitive Neuroscience*, *26*(11), 2530-2539.

Metzner, P., Malsburg, T., Vasishth, S., & Rösler, F. (2017). The importance of reading naturally: Evidence from combined recordings of eye movements and electric brain potentials. *Cognitive science*, *41*(S6), 1232-1263.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Nieuwland, M. S., & Van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of cognitive neuroscience*, *18*(7), 1098-1111.

Nikolaev, A. R., Meghanathan, R. N., & van Leeuwen, C. (2016). Combining EEG and eye movement recording in free viewing: Pitfalls and possibilities. *Brain and cognition*, *107*, 55-83.

Novick, J. M., Trueswell, J. C., & Thompson-Schill, S. L. (2005). Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. *Cognitive, Affective, & Behavioral Neuroscience*, *5*(3), 263-281.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, *9*(1), 97-113.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational intelligence and neuroscience*, *2011*, 1.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of memory and language*, *31*(6), 785-806.

Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: evidence of the application of verb information during parsing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 786.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Plöchl, M., Ossandón, J. P., & König, P. (2012). Combining EEG and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data. *Frontiers in human neuroscience*, *6*.

Rommers, J., Dickson, D. S., Norton, J. J., Wlotko, E. W., & Federmeier, K. D. (2017). Alpha and theta band dynamics related to sentential constraint and word expectancy. *Language, cognition and neuroscience*, *32*(5), 576-589.

Sereno, S. C., & Rayner, K. (2003). Measuring word recognition in reading: eye movements and event-related potentials. *Trends in cognitive sciences*, *7*(11), 489-493.

Smith, N. J., & Kutas, M. (2015). Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology*, *52*(2), 157-168.

Smith, N. J., & Kutas, M. (2015). Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. *Psychophysiology*, *52*(2), 169-181.

Spotorno, N., Cheylus, A., Van Der Henst, J. B., & Noveck, I. A. (2013). What's behind a P600? Integration operations during irony processing. *Plos One*, *8*(6), e66839.

Staub, A., & Rayner, K. (2007). Eye movements and on-line comprehension processes. *The Oxford handbook of psycholinguistics*, *327*, 342.

Strauß, A., Kotz, S. A., Scharinger, M., & Obleser, J. (2014). Alpha and theta brain oscillations index dissociable processes in spoken word recognition. *Neuroimage*, *97*, 387-395.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, *30*(4), 415-433.

Van Gompel, R. P., Pickering, M. J., & Traxler, M. J. (2001). Reanalysis in sentence processing: Evidence against current constraint-based and two-stage models. *Journal of Memory and Language*, *45*(2), 225-258.

Van Petten, C. (2014). Examining the N400 semantic context effect item-by-item: Relationship to corpus-based measures of word co-occurrence. *International Journal of Psychophysiology*, *94*(3), 407-419.

Vignali, L., Himmelstoss, N. A., Hawelka, S., Richlan, F., & Hutzler, F. (2016). Oscillatory brain dynamics during sentence reading: a fixation-related spectral perturbation analysis. *Frontiers in human neuroscience*, *10*, 191.

Wang, L., Jensen, O., Van den Brink, D., Weder, N., Schoffelen, J. M., Magyari, L., Hagoort, P., & Bastiaansen, M. (2012). Beta oscillations relate to the N400m during language comprehension. *Human brain mapping*, *33*(12), 2898-2912.

Wang, L., Zhu, Z., & Bastiaansen, M. (2012). Integration or predictability? A further specification of the functional role of gamma oscillations in language comprehension. *Frontiers in psychology*, *3*, 187.

Widmann, A., Schröger, E., & Maess, B. (2015). Digital filter design for electrophysiological data–a practical approach. *Journal of neuroscience methods*, *250*, 34-46.

APPENDIX

Supplemental: Context Measures

While we matched our EEG effects for lexical frequency and word length in addition to fixation duration and found few differences between these and the effects when only matched for fixation duration, we nonetheless wanted to ensure that our measures of contextual fit were measuring unique properties of the language.

Statistical Analyses

To test effects of lexical properties on each measure of contextual distance, Pearson correlations and multiple regression were used. See methods for how local and topic contextual fit measures were calculated. For correlation analyses, Pearson r-statistics, p-values, and β-coefficients derived from linear models are reported. For multiple regression, omnibus F-statistics, p-values, and adjusted R^2-statistics are reported, as well as the t-statistics and p-values for main effects and interactions, and β-coefficients for effects statistically significant at a threshold of $\alpha = 0.05$. Because our contextual fit measures are actually distance measures, in this supplemental material we refer to contextual fit as contextual distance for ease of interpretation.

Results and Discussion

Lexical properties and local contextual distance

We calculated local contextual distance for 1951 content words in the stories (M =4.46, SD = 1.53). In a model regressing lexical frequency, word length, and the frequency:length interaction on local contextual distance ($F(3, 1947) = 851$, $p < 0.001$, $R^2 = 0.57$, see table supplemental 1), we found a significant effect of lexical frequency ($t(1947) = 21.62$, $p < 0.001$, β

= 0.77), word length ($t(1947) = -2.78$, $p < 0.01$, $\beta = -0.06$), and the frequency:length interaction ($t(1947) = 11.17$, $p < 0.001$, $\beta = 0.07$). The finding that as lexical frequency increases, local contextual distance increases, can be interpreted as the more frequently a word is used, the less likely it is to specifically relate to any word in its context, and therefore the greater its distance to its context. Likewise, the finding that as word length increases, local contextual distance decreases, can be interpreted in a similar fashion. Longer words are less frequent, and therefore are more likely to have a shorter distance to their context. The significant frequency:length interaction does run somewhat in contrast to this mediation account of lexical frequency on word length, and suggests that word length, or some other lexical property correlated with word length, may uniquely predict contextual distance as compared to lexical frequency. However, while these effects are significant, the full model only explains 57% of the variance in local contextual distance after adjusting for the number of predictors, suggesting that this measure of local contextual distance captures some unique property of the language independent of these lexical properties.

Table supplemental 1
Supplemental, local contextual distance regressed on lexical properties

Predictor	t	p	β (if $p < 0.05$)
Lexical frequency	21.62	< 0.001	0.77
Word length	-2.78	< 0.01	-0.06
Frequency:length	11.17	< 0.001	0.07

Note. β-coefficients only reported for significant effects.

Lexical properties and topic contextual fit

We calculated topic contextual distance for 1956 content words in the stories, which were z-scored within each story. In a model regressing lexical frequency, word length, and the frequency:length interaction on topic contextual distance ($F(3, 1952) = 3.82$, $p < 0.01$, $R^2 =$

0.004, see table supplemental 2), there was a significant effect of lexical frequency ($t(1952) = 2.87$, $p < 0.01$, $\beta = 0.12$) and word length ($t(1952) = 2.73$, $p < 0.01$, $\beta = 0.07$), but no significant frequency:length interaction ($t(1952) = -1.29$, $p = 0.20$). The effect of lexical frequency on topic contextual distance was in the same direction as local contextual distance, albeit much smaller, but the effect of word length was in the opposite direction- as word length increases, topic contextual distance increases. It is plausible that longer words have a shorter distance to some of the topic words than average, but longer distance to the other topic words than average, appearing as an overall increase in topic contextual distance. However, the full model only explains 4% of the variance in topic contextual distance after adjusting for the number of predictors, so it may be reasonable to say that overall, these lexical properties have a fairly minimal impact on topic contextual distance.

Table supplemental 2
Supplemental, topic contextual distance regressed on lexical properties

Predictor	t	p	β (if $p < 0.05$)
Lexical frequency	2.87	< 0.01	0.12
Word length	2.73	< 0.01	0.07
Frequency:length	-1.29	0.2	NA

Note. β-coefficients only reported for significant effects.

Relationship between local and topic contextual fit

There was no significant relationship between local and topic contextual distance for the 1948 content words which had both a local and topic contextual distance measure ($r = 0.009$, $p = 0.68$). This suggests that local and topic contextual fit provide unique predictive affordances to language comprehension.