

THE UNIVERSITY OF CHICAGO

AUDITORY UNDERSTANDING OF SPEECH AND NON-SPEECH

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

INTERDISCIPLINARY SCIENTIST TRAINING PROGRAM: COMPUTATIONAL

NEUROSCIENCE

BY

SOPHIA UDDIN

CHICAGO, ILLINOIS

JUNE 2018

This dissertation is dedicated to my two favorite sources of sound waves: Gabriel Lucas Gitin Riccio and Willow Uddin-Riccio.

Table of Contents

List of Tables.....	vii
List of Figures.....	viii
Acknowledgements.....	ix
Abstract.....	x
Chapter 1: Introduction	1
1.1 Specialized language systems vs. powerful general cognitive systems.....	2
1.2 Non-language stimuli and language mechanism extensibility	6
1.3 Context Effects on Word Recognition	9
1.3.1 Context facilitates word recognition and interpretation	9
1.3.2 Implications of context effects for language-specific vs. general models.....	12
1.3.3 Unanswered questions in context effects on non-language.....	13
1.4 Overview of the studies in this dissertation.....	16
Chapter 2: Does spoken sentence context aid recognition of words and environmental sounds equally?	19
2.1 Introduction	20
2.2 Methods	27
2.2.1 Participants	27
2.2.2 Stimuli	28
2.2.3 Task.....	29
2.2.4 Data collection.....	30
2.2.5 Data Analysis	30
2.3 Results.....	31
2.4 Discussion	33

Chapter 3: Can non-linguistic sounds be quickly and easily understood in linguistic context?	35
3.1 Introduction	35
3.2 Methods	35
3.2.1 Participants	35
3.2.2 Stimuli	36
3.2.3 Task	36
3.2.4 Data collection and analysis	37
3.3 Results.....	38
3.4 Discussion: Implications of Experiments 1 and 2	40
 Chapter 4: Do the neural responses to speech and non-speech reveal important processing differences related to understanding them in linguistic context?	 47
4.1 Conceptual Introduction	48
4.2 Methods	53
4.2.1 Participants	53
4.2.2 Stimuli	54
4.2.3 Testing procedure	54
4.2.4 EEG Setup	55
4.2.5 Data preprocessing	55
4.2.6 Analyses	56
4.3 Results.....	59
4.4. Discussion	77
4.4.1 Conclusions	83

Chapter 5: Are differences in the ERP's observed for speech and non-speech in speech context due purely to acoustic differences?	85
5.1 Conceptual Introduction	85
5.2 Methods	88
5.2.1 Participants	88
5.2.2 Stimuli	89
5.2.3 Testing procedure	89
5.2.4 EEG Setup	89
5.2.5 Data preprocessing	90
5.2.6 Analyses	90
5.3 Results	93
5.3.1 Experiment 3 Follow-up Analyses	93
5.3.2 Exploratory analyses—talker normalization	103
5.3.3 Do WM-related mechanisms mediating talker normalization also mediate the understanding of ES in sentence context?	108
5.4 Discussion	111
5.4.1 Discussion – follow-up to ERP work on ES vs. words	111
5.4.2 Discussion – follow-up to previous talker normalization work	115
5.4.3 Discussion – WM and talker-normalization-related sources in the processing of ES	120
Chapter 6: General Conclusions	126
6.1 Purpose of work	126
6.2 Explaining the present results: developing a new framework that can incorporate ES and words	128
6.2.1 Accomplishing early acoustic signal processing without phonemes	128
6.2.2 Syntactic analysis and its meaning in a framework that accommodates ES	132

6.2.3 Semantic processing in speech and ES understanding, and the role of context	134
6.2.4 Interpretation of the meaning of a sentence as a whole, even when it includes ES	137
6.3 Final Conclusions.....	138
References	142
Appendices.....	156
Appendix 1: Stimuli.....	156
Appendix 2: Detailed methodology for topographical randomization analyses	157
Appendix 3: AAHC microstate analysis	159
Appendix 4: Questionnaires	161

List of Tables

Table 2.1 Participant <i>n</i> 's, Expt 1.....	29
Table 3.1 Coefficients for factors in a logit mixed model for percent correct data.....	39
Table 4.1 Electrodes included in ROIs.....	58
Table 4.2 Raw topography TANOVA windows.....	62
Table 4.3 Difference topography TANOVA windows.....	66
Table 4.4 Summary of statistics for randomization analyses of topographical maps.....	66
Table 4.5 Summary of statistics for repeated measures ANOVA examining N400 latency	67
Table 5.1 Raw topography TANOVA windows.....	94
Table 5.2 Difference topography TANOVA windows.....	97
Table 5.3 Summary of statistics for randomization analyses of topographical maps.....	99
Table 5.4 Summary of statistics for repeated measures ANOVA examining N400 latency	100
Table 5.5 TANOVA including WM.....	104
Table 5.6 Time windows examined in source analysis.....	106
Table 5.7 Residual variance for source analysis models.....	106
Table 5.8 TANOVA including WM for Expt 3.....	109
Table 5.9 Residual variance for source analysis models.....	111

List of Figures

Figure 1.1 A predictive coding framework.....	10
Figure 2.1 Duration of target needed for recognition.....	32
Figure 2.2 Cloze probability and recognition point.....	33
Figure 3.1 Percent correct for RT task.....	39
Figure 3.2 RTs for meaningful vs. nonsense judgment.....	40
Figure 4.1 Match vs. mismatch N400.....	60
Figure 4.2 ES and word N400 difference topographies.....	64
Figure 4.3 General vs. specific difference topographies.....	68
Figure 4.4 N400 for general vs. specific, words vs. ES.....	71
Figure 4.5 ACC topographies for ES.....	73
Figure 4.6 ES and words match vs. mismatch voltage traces.....	75
Figure 4.7 P600 topographies for match vs. mismatch.....	77
Figure 5.1 N400 and P600 match vs. mismatch topographies.....	95
Figure 5.2 General vs. specific N400 difference topographies.....	98
Figure 5.3 Time course of GD for match vs. mismatch.....	98
Figure 5.4 N1 and P2 window topographies across talker.....	101
Figure 5.5 Microstates.....	105
Figure 5.6 Correlation between source analysis model fit and WM.....	108
Figure 5.7 Scalp topographies dependent on WM.....	109

Acknowledgements

Many thanks to my PI Howard Nusbaum and the rest of my committee Leslie Kay, Nicholas Hatsopoulos, and Dan Margoliash, for scientific mentorship and moral support. I also thank the MSTP (especially Alison Anastasio and Shanetha Thomas) for support along the graduate school path. This work would have been impossible without Steve Van Hedger, Serena Klos, and Shannon Heald filling critical roles as teachers, role models, and co-authors. Thomas Koenig, the mind behind the RAGU topographic analysis software, graciously made himself available to answer questions over email. Thanks for project feedback, collaboration on piloting and/or data collection, and general encouragement goes to present and past Nusbaum lab members Katie Reis, Nina Bartram, Peter Hu, Heather Mangelsdorf, Jean Matelski-Boulware, Michael Stubna, and Brendan Colson. I also thank my MSTP class and associated folks (Eliza Frost, Sofija Canavan, Jeff Bunker, Alex Guzzetta, Deanna Arsala, Dan Camacho, Chad Sell, Victoria Lee, Ben Casterline, Regina DePietro, and Chanie Howard) for moral support, general science advice, and the frequent sharing of good food. Thanks to my Folklear Nuclei (Tahra Eissa, Kyler Brown, Haley Stinnett, Bo Osinski, and Geoff Brookshire) for sharing with me a rare combination of neuroscience expertise, music, and the ability to have a good time anywhere, anytime. Thanks to my Mandelbrots (Amy Langdon, Ben Dair, Leland Kusmer), Sacha Mullin, my prog brethren (Dino Zaigirdar, Bill Banks, Matt Tate, and Nick Tate), and everyone in MEME for making music with me and encouraging me to take a break from science once in a while. Finally, thanks to my life partner Gabe Riccio and my little darling Willow Uddin-Riccio for making our home a productive writing-haven, and my family (Alice Uddin, Haseen Uddin, and Olivia Uddin) for unconditional support from afar.

Abstract

There is a longstanding debate about the specificity and encapsulation of the mechanisms behind humans' ability to understand speech. Some scholars argue that these abilities arise from evolutionarily specialized language modules, while others hold that they are a result of the interaction between cognitive-general mechanisms and extensive experience. In the present work, we use meaningful non-speech sounds, e.g. dog barks and train whistles, in the context of spoken sentences, to probe the extent to which language understanding mechanisms can quickly and easily adapt to non-speech. We find that behaviorally, listeners can recognize and understand non-speech sounds in sentence context quickly and easily, with processing costs much smaller than those that would be expected from a "translation" or covert naming process. Moreover, neurally, the N400 and P600 potentials are remarkably similar between non-speech sounds and matched spoken words, suggesting similar processing that is, in both cases, equally sensitive to sentential constraint and congruency with the preceding sentence context. These results closely mirrored the results from a final study in which we examined neural responses to spoken words said by a different talker. Analysis of relationships between working memory and scalp topography, as well as source analysis, indicated that both the comprehension of a different talker and the comprehension of an environmental sound involve early recruitment of working memory in the N1 and P2 time windows. These processes are mediated by parietal and temporal speech areas that have previously been identified in studies of talker normalization. Though the degree of recruitment of these resources may be greater and/or last longer for non-speech sounds than for a changing talker, the underlying mechanisms appear to be qualitatively the same. These results are discussed in light of a new model of auditory understanding, in which working memory, attention, and experience

interact to allow flexible and rapid understanding of many different types of auditory stimuli, including (but not limited to) speech.

Chapter 1: Introduction

Although many species use vocal communication in complex ways, human spoken language is thought to be unique (cf. Hockett, 1959) in its generative ability—called “duality of patterning”—of combining old units to make new meanings (i.e., through syntax) and to create novel words (i.e., through phonology, phonotactics, and morphology) that can name or describe objects or events (Trask, 2003). This kind of capacity for meaningful unit creation and unit combination in a biological communication system has been theoretically attributed to specialization of systems through evolution (Nowak & Komarova, 2001; Nowak, Plotkin, & Jansen, 2000). This kind of theory postulates that such specialized systems must be tuned specifically for the sounds of speech in order to be able to combine such sounds in speaking and in comprehension (e.g., Liberman, 1970). However, alternate theories hold that the sophisticated capabilities of the human language system are not a result of a highly specialized language system, but of more cognitive-general processes. In this alternative view, it is conceivable that the language system could also process non-linguistic, but meaningful sounds, such as environmental sounds (e.g., the sound of a dog barking). This dissertation tests hypotheses regarding the extensibility of the human language understanding system in just this way using behavioral measures and brain electrophysiology. If the language system is limited to combining spoken words into legal (syntactically structured) and meaningful sequences, sound patterns that are not words but that are still meaningful (e.g., environmental sounds) should not be understood in the same way, or indeed by the same system, as spoken words.

In this introduction, I will first describe modular-specialized versus general-cognitive accounts of language understanding in more detail (section 1.2). I will then explore the idea of

non-language stimuli such as environmental sounds as meaningful stimuli that can be used to probe the ability of language systems to extend to non-language (section 1.3). I will then address the well-documented effects of context on language processing to lay the groundwork for my examination of the implications of context effects on non-language stimuli (section 1.4). Finally, I will introduce the individual experiments (section 1.5). The first experiment tests whether context affects recognition of non-speech stimuli in the same way it is known to affect speech. The second experiment asks how well listeners can understand non-speech auditory stimuli in the context of a spoken sentence. The third experiment uses EEG to compare the neural processing of non-speech and speech stimuli in the context of a spoken sentence. Finally, the last experiment compares neural responses to speech said by the same or a different talker as the preceding sentence to assess the extent to which responses could be affected by acoustic differences.

1.1 Specialized language systems vs. powerful general cognitive systems

Is the celebrated generative power of human language a consequence of a powerful general cognitive system or a specialized language system? Fodor (1983) proposed a view about the specialization of the language system resulting in a dedicated language module that specified particular features that define language in this way. This theoretical framework emerged from a proposal in which Chomsky argued that language depends on a “language acquisition device”—an innately encoded module responsible for language acquisition, which is present from birth (Chomsky, 1969). Chomsky’s reasoning was that given young children’s impressive ability to quickly learn their native tongue, strong predispositions, even including knowledge about grammar, must be naturally present at birth given that there is putatively insufficient information in the spoken language environment to “induce” language from listening experience (Chomsky, 1969, 1986). Hand in hand with Chomsky’s ideas were theories that the neural substrates for

understanding speech evolved to match the speech production system. Liberman et al. (1967) recognized that in natural speech, phonemes that people perceive as being in the same category (e.g., all “b’s”) actually vary widely acoustically, depending on many factors including the surrounding phonemes, the speaker, and the rate of speech. These authors reasoned that a simple acoustic speech “decoder”—in which listeners sequentially match the phonemes they hear to inner “templates”—could not operate efficiently on natural speech. Instead, they hypothesized that the decoding of speech depends, at least partly, on the neural substrates responsible for producing speech (Liberman et al., 1967). Later incarnations of this theory specify that “phonetic information is perceived in a biologically distinct system, a ‘module’ specialized to detect the intended [vocal tract] gestures of the speaker that are the basis for phonetic categories” (Liberman & Mattingly, 1985).

These ideas suggest the importance of specialized cognitive systems that have expertise in carrying out functions such as language learning and speech recognition. According to Fodor, human language depends on a dedicated “language system”; one that processes only linguistic stimuli, and does not handle other types of input. Given the extremely fast, seemingly effortless process of language comprehension and production for native speakers, these ideas made sense because they seemed to explain how the brain could so quickly make sense of something so complex as language. If language is processed by an encapsulated system that is relatively isolated from other functions, it should be easy for people to quickly—almost automatically—process and understand language. A more general, flexible system, on the other hand, might be too unwieldy to allow language to be understood so effortlessly. Moreover, the corollary of this in the work by Mattingly and Liberman is that speech is understood as the articulatory gestures that produced the sounds and that locks the spoken language processing system into the requirement of being tuned

for vocal-tract produced sounds. In more recent years, the idea of the language system as a specialized, encapsulated module has been updated and argued for either with imaging and lesion studies, which often focus on seemingly language-specific deficits following particular left-hemisphere lesions (e.g., Grodzinsky, 2000; Maurer, Fromkin, & Cornell, 1993), or with studies that focus on similarities between infant and adult language networks, presumably before much language learning has occurred in the infants (Dehaene-Lambertz et al., 2010; Dehaene-Lambertz, Hertz-Pannier, & Dubois, 2006).

However, the concept of the language system as modular and specialized only for language input has been challenged by evidence that language comprehension relies on powerful and flexible cognitive-general mechanisms. Dick et al. (2001) examined the type of lesion patients that had been previously used to argue in favor of modularity. They found that depending on the precise location of the lesion, these patients can have a variety of additional deficits—in the recognition of environmental sounds, for example (see Bates & Dick, 2002). If language comprehension relies on specialized pathways for the processing of language-only input, such environmental sound deficits should not arise from language area lesions.

Further, Dick, Lee, Nusbaum, & Price (2011) found that perisylvian areas thought to be language-selective were activated in response to violin music in professional violinists, but not in other participants who did not play violin. This evidence suggests that it is expertise with a particular type of meaningful auditory stimuli, rather than something intrinsic to speech itself, that recruits these networks. The reason these areas appear to be language-selective may be simply that most research subjects have extensive experience with spoken language. Work on jazz improvisation has supported these findings; Donnay, Rankin, Lopez-Gonzalez, Jiradejvong, & Limb (2014) found that perisylvian language areas are activated in jazz musicians who are

interactively improvising or “trading fours”. Thus, these findings hold in a live performance setting as well as in a listening paradigm. According to these accounts, the human ability to easily understand language is due to cognitive-general mechanisms that also assist in processing diverse other types of stimuli, and not due to a highly specialized, language-only system.

Also in support of this account, Leech, Holt, Devlin, & Dick, (2009) used a video-game-based task to implicitly train subjects to categorize complex sounds. Importantly, these sounds were computer-generated and did not resemble speech. After training, the sounds were shown via fMRI to activate regions in the posterior superior temporal sulcus (pSTS) traditionally thought to be speech-selective; such activation in response to these sounds was not present before subjects learned to categorize the sounds via extensive experience. This led Leech et al. to hypothesize that it is experience with categorizing meaningful auditory stimuli, rather than anything intrinsic or unique to speech, that recruits the neural systems traditionally associated with speech. This body of work supports the interpretation that “language networks” are recruited to process any complex, meaningful auditory stimuli with which the listeners have listening expertise—not just language. This interpretation in turn supports a cognitive-general, rather than modular specialized, theory of language understanding, as the same resources are recruited to understand both language and non-language stimuli.

Though arguments that speech is unique and modular often take evolutionary perspectives, it is worth noting that there are evolutionary arguments that speech is not as unique and encapsulated as some might theorize. For example, some point out that computational properties of nonhuman primate auditory processing systems share much in common with the neural substrates of human language processing. Rather than suggesting that the neural basis for language is a unique human adaptation, these authors use comparative data to argue that it emerged naturally

from dual dorsal and ventral processing streams in primate audition. Other authors take an evolutionary perspective to argue that human language is simply one instance of vocal learning abilities that have convergently evolved multiple times. For example, work from Erich Jarvis's lab has found homologs to many human "language" areas in vocal learner bird species and supported this argument with gene expression data (Petkov & Jarvis, 2012; Pfenning et al., 2014), while Heffner and Heffner (1984) found that macaques rely on the left superior temporal gyrus (STG; an area important for speech perception in humans) to discriminate conspecific vocalizations. Therefore, there does exist evolutionary work supporting the idea that language relies on auditory processing systems that are much more general-purpose, and evolutionarily older, than language understanding.

1.2 Non-language stimuli and language mechanism extensibility

One way to address the question of the specialization of language mechanisms is to investigate the extent to which they operate similarly on meaningful but non-linguistic information. How would recognition of meaningful, but non-linguistic patterns be affected by linguistic context, in a language understanding task? If language understanding depends on highly specialized mechanisms, making sense of a non-language stimulus embedded in language should be difficult, or at the very least require extra processing to translate it into a language format that the language system can accept.

Potter, Kroll, Yachzel, Carpenter, and Sherman (1986) addressed this question by asking whether printed sentences containing a picture substituted for a noun (called "rebus" sentences) affected the speed and accuracy of plausibility judgments about the sentences. This was done to ascertain whether pictures could directly access the same system of conceptual meaning as words. Potter et al. reasoned that if pictures could directly access this system, then response times for

plausibility judgments should be similar for “rebus” sentences and all-word sentences. This was indeed what they found. These results suggest that words and other meaningful referential stimuli may be understood in a similar way, at least when presented in a printed sentence context.

While the rebus work from Potter et al. suggests that pictures and written words can access a system of general concepts in a similar way, it is possible that reading and spoken language could work very differently. Speech is a much older system evolutionarily, with reading having appeared fairly recently in human history (cf. Dehaene & Cohen, 2011). Evolutionary arguments about language typically concern spoken language and motor/auditory systems (see Non-language stimuli and language mechanism extensibility), so it is important to test if Potter’s findings hold when spoken language, not reading, is used. To accomplish this, environmental sounds (e.g., a dog barking, a car horn honking) can be used. Environmental sounds are auditory patterns that are meaningful, but not “linguistic”: they lack internal phonological segments or higher-order linguistic structure, and are not produced by the human vocal tract. In this way, they are decidedly different from spoken language. They are, however, easily recognized and categorized (Gygi et al., 2007; Warren & Verbrugge, 1984), which means they easily convey meaning. This makes environmental sounds ideal auditory counterparts to the pictures used in Potter’s rebus work.

It follows from Liberman and Mattingly’s work that stimuli that do not have corresponding articulatory gestures should not interact with a putative speech module. In other words, if a listener cannot use their vocal tract to produce what is being heard, the stimulus in question must be understood via a completely different, non-language system. Whereas not all theories of word recognition go so far as to say that word understanding depends on articulatory gestures, many of them depend heavily on breaking words down into constituent units such as phonemes and syllables. For example, the popular TRACE model relies on computations within and between

layers of nodes corresponding to acoustic speech features (the lowest level), phonemes (the middle level), and words (the highest level) (McClelland & Elman, 1986). Shortlist and Shortlist B are computationally very different from TRACE in that they are entirely bottom-up models, but they still rely on a phoneme layer and a lexical layer (Norris, 1994; Norris & McQueen, 2008). Such units (e.g., phonemes) are obviously not present in environmental sounds, which have a much wider variety of spectral and temporal structures than spoken words as they can obviously come from a much wider array of sources.

From the above models, it might seem like recognition of non-language stimuli should be accomplished by entirely different mechanisms than word recognition. The work of Potter et al. suggests that, at least for written language and pictures, both words and non-word pictures can quickly and easily access the same general concepts, which is not surprising. However, their finding that pictures could quickly be understood in the context of an all-word sentence argues against the existence of two separate encapsulated systems, as they found no processing cost for switching between the putative language and non-language pathways. This suggests another alternative: that words and pictures use at least partially the same pathway in order to access concepts. In such a view, both words and pictures could function as patterns that are recognized and immediately access the amodal concept (e.g. the concept of a dog or an apple) to which they are related. If this is true, it makes sense that there would be no processing cost for understanding one in the context of the other, as found by Potter et al. However, as Potter et al. examined only written language, it is still an open question whether spoken language operates this way.

1.3 Context Effects on Word Recognition

1.3.1 Context facilitates word recognition and interpretation

One of the hallmarks of both spoken and written language is the interaction of word recognition and interpretation with meaningful linguistic context (Morris & Harris, 2002; Simpson, Peterson, Casteel, & Burgess, 1989). Information in a proximal context can improve understanding of spoken language. Semantic priming is an example in which words are recognized faster if preceded by a related word rather than an unrelated word (Hutchison et al., 2013). Context strongly affects word understanding in full sentences as well. Gating studies, in which a word is presented incrementally in segments of increasing length, have shown that in a highly constraining, specific context (as opposed to a general, more vague context) people need to hear fewer gates of a word to understand it (e.g., Grosjean, 1980). Additionally, when people are asked to fill in a sentence ending, they supply a word faster for a highly constrained sentence beginning than for a low constraint beginning (Staub, Grant, Astheimer, & Cohen, 2015).

Why are word recognition and understanding influenced by linguistic context? Extant word recognition models incorporate the effects of context information on lexical knowledge to varying degrees (see Dahan & Magnuson, 2006 for a review). Some recognition models suggest that stimulus input (e.g. the acoustic waveform of a spoken word or the visual input of a printed word) is the primary determining factor in the recognition process by matching pattern properties of input to stored psychological representations of words in a mental lexicon (e.g. Norris, 1994; Norris and McQueen, 2008). Given the primacy of stimulus pattern information in recognition, these models hold that top-down knowledge, like contextual or conceptual information, cannot directly affect perception and operate on response selection following pattern recognition. In other words, input feeds forward through a series of transformations until a word is recognized, and it is only at late

stages, when the recognized word is being integrated with its surrounding context and meaning is being assessed, that contextual information exerts an effect. In contrast to such contextually “autonomous” models, interactive recognition models allow for continuous, on-line effects of context on word recognition; top-down information is capable of altering processing at lower levels (e.g. McClelland & Elman, 1986; Mirman et al., 2006). In interactive models, feedback projections from areas representing general conceptual information (potentially including information from linguistic context) can alter processing at lower perceptual levels.

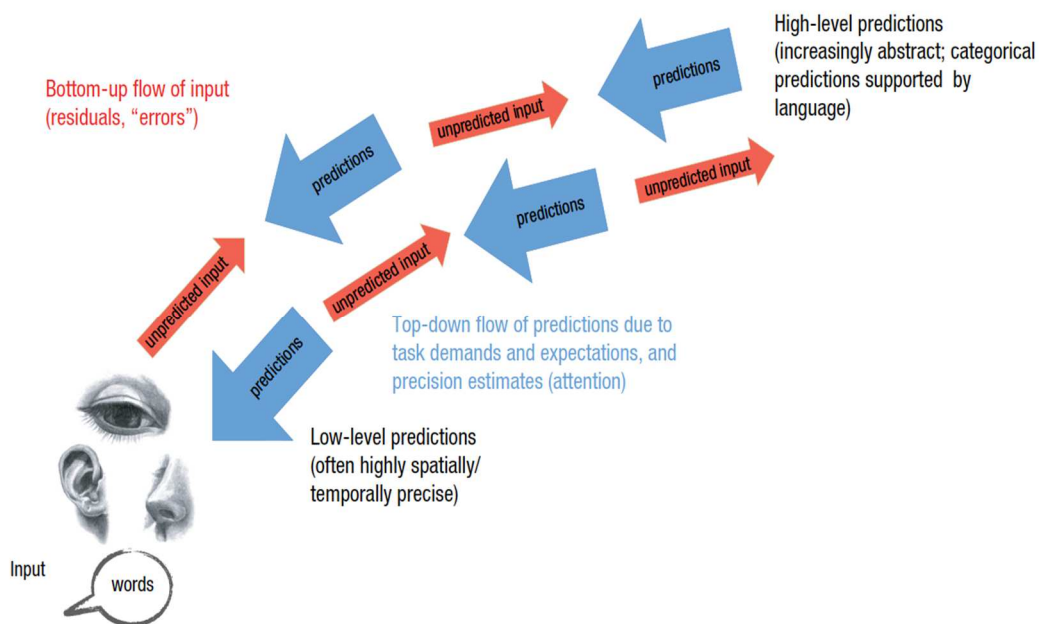


Figure 1.1 A predictive coding framework. Figure 1 from Lupyan & Clark (2015) showing a general framework by which neural predictions (blue) interact with sensory input (orange) in language perception.

More recently, information on predictive coding from the neuroscience literature has been used to reframe our understanding of context effects on word recognition. As can be seen from predictive coding models such as that shown in Figure 1.1, bottom-up sensory input (e.g., the hair cells in the inner ear firing in response to the acoustic input of speech) interacts with neural predictions from cortical regions. While these neural predictions are “top-down” effects, it is

important to realize that many predictive coding theories (like the one shown in Figure 1.1) place importance on neural predictions at many different processing levels. This is in contrast to earlier models (e.g., Norris & McQueen, 2008; McClelland & Elman, 1986) in which top-down information is mainly abstract and conceptual in nature. Lupyan and Clark's (2015) model includes both low-level top-down predictions, which include precise predictions about the sensory input informed by higher-level information, and high-level predictions, which are more abstract and might include conceptual information accessible via semantic associations. In a predictive coding paradigm, the facilitative effects of sentence context on word recognition could be due to the facilitation of predictions at multiple levels. It is obvious that context will aid in the formation of abstract, conceptual predictions as long as the listener is understanding the meaning of the sentence as it unfolds; this is not unlike the earlier models already discussed. However, facilitation could also be a result of the formation of sensory predictions about the upcoming acoustic input as a result of top-down expectations. For example, if the listener is expecting a familiar talker to say the word "dog" at the end of a sentence, under such a predictive coding framework, it is possible that the system is already anticipating the sound of this familiar talker pronouncing the word "dog." In other words, processing will be assisted by neural predictions about input at a low, sensory level. As a side note, it is also possible that low-level predictions could be made as a result of predictive processing within the lower levels themselves (Nusbaum & Schwab, 1986). However, such predictions would necessarily rely on computations (e.g., tracking acoustic regularities) that can be made in lower-level circuitry. It is therefore extremely unlikely that such intrinsic low-level predictions would vary based on the conceptual meaning of the preceding sentence. A final view is that the speech motor system can explain context effects. In certain conditions, for example when spoken word forms are predicted by context, or a non-linguistic motor action is predicted, it

is thought that the speech motor system generates predictions that aid comprehension (Pickering & Garrod, 2007).

1.3.2 Implications of context effects for language-specific vs. general models

The level at which context-related neural predictions happen has implications for theories of language processing. For example, if context facilitates processing mainly via predictions about near-term acoustic input to low-level sensory processes, then predictions might depend on speech- or language-specific properties. Alternatively, the system might generate predictions about future low-level sensory processing based on the speech-motor system. In these cases, an encapsulated model of language understanding would fit well; context should apply to language via neural predictions of these very language-production-specific items. In such a case, information from a fragment of a spoken sentence would generate neural predictions for subsequent words via speech sounds, articulatory gestures, or phonemes that are likely to follow. In addition, in such a system it would be very difficult to derive any benefit of context for processing any stimuli that are not speech (for example, if they are environmental sounds). On the other hand, it is possible that context operates primarily via predictions about higher-level interpretive or conceptual processes, rather than specific sounds or phonemes. In this case, a cognitive-general model of language understanding would fit better. If a spoken sentence sets up expectations, these expectations should affect subsequent meaningful stimuli regardless of whether said stimuli are linguistic or not.

In terms of explaining context effects on word recognition and interpretation, the autonomous and interactive models described above are quite different. However, many such models are entirely reliant on language-specific features (such as phonemes or articulatory gestures), thereby making strong predictions about the recognition and understanding of non-linguistic information. In autonomous models, word recognition must happen before context

effects operate. This means that phonemes or gestures must be recognized as particular words before context can be used to aid in processing the word. Under such a system, it seems virtually impossible that linguistic context could facilitate processing of non-language stimuli, as word recognition is a necessary first step before context information becomes available, at least in accordance with the specific theories that have been described. There is simply no processing path to incorporate other forms of recognition and understanding into the language process. Interactive models, too, do not provide any account for context effects outside of a language-specific system. Even though they allow context to affect perception, most of these models describe highly specific, context-induced feedback modulation of acoustic feature and phoneme perception. For example, the Hebb-Trace model (Mirman et al., 2006) allows for learning and context-related feedback with respect to the recognition of acoustic and articulatory features such as voicing. While these models are quite good at explaining context effects on speech perception specifically, contextual facilitation of non-linguistic stimuli is outside their scope. They are thus difficult to interpret in the context of cognitive-general models of language understanding.

1.3.3 Unanswered questions in context effects on non-language

There is some evidence that systems for understanding language are at least capable of relying on high-level conceptual predictions for facilitation of context. This evidence comes from many studies showing that contextual information from both non-word and lexical sources can affect understanding. One example of this sort of interaction is analog acoustic expression, the phenomenon in which non-linguistic modulations in pitch and speaking rate in speech that are metaphorically descriptive of an event or object affects the listener's understanding of the lexical content of a message (e.g., Shintel et al., 2006; Nygaard et al., 2009). Information from musical underscoring can affect speech understanding in a similar manner (Hedger et al., 2013). The effects

of non-linguistic information on linguistic interpretation are not confined to the auditory modality. Tanenhaus and colleagues have used eye tracking to demonstrate that listeners make rapid on-line use of visual scene context in order to disambiguate spoken verbal instructions (Chambers, Tanenhaus, & Magnuson, 2004; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). In addition, there are many priming studies in which visual or spoken words can facilitate processing of environmental sounds and vice versa (Frey, Aramaki, & Besson, 2014; Orgs, Lange, Dombrowski, & Heil, 2007; Orgs, Lange, Dombrowski, & Heil, 2006; van Petten & Rheinfelder, 1995). Both words and environmental sounds have also been found to prime recognition of pictures (Chen & Spence, 2011; Schneider, Engel, & Debener, 2008). Concepts associated with words can also influence processing in other domains, as when words describing a particular direction of motion (such as the word “approach”) affect visual motion perception (Meteyard, Bahrami, & Vigliocco, 2007). Even when concepts are conveyed in a complex, non-linguistic way (e.g., an auditory scene), they can bias the people’s verbal labels for ambiguous environmental sounds (Ballas & Mullins, 1991). Thus, there is strong evidence that non-linguistic contextual information can facilitate spoken word processing, and vice versa.

Despite the extensive documentation of interactions between non-linguistic and linguistic information, the mechanisms behind these effects remain unclear. One possibility is that participants are covertly naming non-linguistic stimuli in order to guide processing words. This possibility is favored by encapsulated accounts of language processing, as according to this viewpoint, non-linguistic information cannot interact with encapsulated language modules until it is translated into linguistic information. For example, if a listener needs to understand a sentence about a trip to the vet with a dog, the sound of a dog barking may convey this rather than the word “dog”. One way this could happen is for the listener to “translate” the dog bark into the word

“dog” by “naming” or identifying the sound. After this happens, a phonetic representation of the word “dog”, or articulatory gestures associated with the word “dog,” can be accessed, and the word “dog” can be integrated, via language-specific mechanisms, with the preceding sentence. It seems unlikely, however, that this is the case. The results reported by Potter et al. (1986) could not be easily attributed to covert naming, as previous work has demonstrated that picture naming takes too long to be occurring in Potter’s paradigm (cf. Oldfield and Wingfield, 1965). Other work also argues against covert naming as the mechanism for context effects. Studies by Chambers et al. (2004) and Tanenhaus et al. (1995) rely on sufficiently complex visual scenes that covert naming alone would not resolve the ambiguities present. Finally, it is highly unlikely that covert naming could explain analog acoustic expression effects, as listeners would have to translate the metaphoric meaning present in vocal pitch or speech rate information directly into words.

If covert naming is not responsible for the language/non-linguistic context effects that have been previously described, how does this process work? This question has not been fully answered. It is possible that, as suggested by Potter and colleagues, words and non-verbal stimuli such as pictures access a single conceptual system that is not grounded in language. In other words, the same neural representations of semantic information could be accessible via words and other meaningful non-verbal stimuli. This would argue in favor of a general cognitive, rather than language-specific, comprehension system. Work by Zwaan and colleagues describes an effect converse to covert naming, in which words activate “mental pictures” of the objects to which they refer (Zwaan, Stanfield, & Yaxley, 2002), providing support for Potter’s hypothesis that words draw on a general conceptual system that is also used by nonverbal stimuli. In terms of a predictive coding framework, this would mean that predictions are sufficiently amodal (or multimodal) to interact easily with information across different domains of representation. There is also the work

that suggests that motor predictions can facilitate understanding of certain non-linguistic stimuli that are associated with particular motor actions (Pickering & Garrod, 2007); however this does little to explain context effects for non-linguistic stimuli that do not correspond to human motor actions.

1.4 Overview of the studies in this dissertation

In the present dissertation, I asked how understanding of recognizable and meaningful, but non-linguistic, environmental sounds is affected by linguistic context by using spoken sentence frames that are completed as a sentence by either a spoken word or an environmental sound. An account of language understanding that isolates speech processing as a separate system from a broader conceptual system predicts that understanding meaningful non-linguistic information in the context of a preceding sentence frame should be different from understanding matched spoken words. Non-linguistic information should be integrated as post-perceptual problem solving, requiring a covert naming step. As a separate process, this might slow understanding substantially (over 500 ms for covert naming, cf. Oldfield & Wingfield, 1965). However, this account seems unlikely given the results of prior research. As noted already, studies have demonstrated rapid interaction of many types of nonlinguistic information with language processing (see 1.3.3 Unanswered questions in context effects on non-language), and more recent research has suggested that words and meaningful non-linguistic stimuli may have more in common in processing than previously thought given the neural resources involved in understanding both (Cummings et al., 2006; Dick, Krishnan, Leech, & Saygin, 2016; Leech & Saygin, 2011; Saygin, 2003; Saygin et al., 2005). However, there is little research on how environmental sounds are understood, especially in comparison to and in the context of speech sounds, and few studies directly comparing how recognition and understanding of these two classes of sounds is affected by variations in the

amount of contextual constraint.

How does sentential (linguistic) constraint affect processing of environmental sounds roughly matched in meaning to spoken words? Spoken words following semantically constraining sentence frames are recognized with less signal and responded to faster than words following less constraining frames (Grosjean, 1980; Tyler & Wessels, 1983), but whether this “constraint benefit” will occur to the same degree for environmental sounds is an open question. If the constraint through language-specific processes, we would not expect to see similar effects for environmental sounds. But if the semantic constraints of a sentence frame affect speech and meaningful non-speech in a similar manner, that would suggest that a more general cognitive system is at work.

The present experiments can address questions about how sentential constraint may be acting via predictive coding. There is evidence that in particular situations, motor systems generate predictions that aid comprehension (as discussed in Overview of the studies in this dissertation). However, the substitution of environmental sounds for spoken words presents a different kind of situation: The recognition and comprehension of a non-speech sound pattern cannot be aided directly by the speech motor system, because the non-speech sound patterns are not vocally generated. Thus, the question would appear to fall more in the realm of predicting an appropriate non-linguistic motor action based on the linguistic form combined with the environmental sound. However, participants in the current study are not deciding how to act in this situation, but simply recognizing the non-speech sound or understanding the sentence frame-plus-sound. Given that this falls outside the explanatory domain of speech-motor system predictive coding, parallel results for increasing sentence constraint in recognition or comprehension for spoken words and meaningful non-speech sounds would pose a theoretical challenge. Similarly, we can also use this experiment to ask, outside of a speech-motor framework, how predictive coding operates to constrain

processing. The specific acoustics of environmental sounds are likely much harder to predict than the acoustics of the same speaker from the sentence frame saying the final word of the sentence, simply because the range of possible environmental sounds that could fit the intended meaning is larger. Thus, it follows that if neural predictions derived from context are largely at the level of sensory representations, constraining sentence contexts should be more helpful to spoken words than environmental sounds. However, if such predictions are more conceptual in nature, we would expect similar constraint benefits for spoken words and environmental sounds.

Chapter 2: Does spoken sentence context aid recognition of words and environmental sounds equally?

The main question for the first experiment is whether the benefit of sentential constraint for recognition is similar for environmental sounds and spoken words. This question is important to our overarching question because it is one way to ask whether speech is unique (a finding that environmental sounds are unable to benefit from sentential constraint would support this account). On the other hand, if speech is merely another way to convey meaning via acoustic patterns (like environmental sounds), we might expect similar constraint effects on recognition points for environmental sounds and spoken words.

Testing recognition can be accomplished with a gating study. The gating paradigm (e.g. Grosjean, 1980) has been used to measure how much of a spoken word waveform is needed for recognition. In a gating study, a stimulus (usually a spoken word) is presented in small successive increments thereby increasing the length of the stimulus. For example, the first 20 ms of a word is presented first, followed by the first 40, then the first 60, then the first 80 ms of the word. After each partial snippet of the word, the participant is asked to identify the segment as a word. In this way, the experimenter can estimate how much of a word's waveform must be heard in order to support recognition. In general, gating studies show that listeners can correctly identify spoken words even before the end of the word is heard, and that less of the word is needed for recognition in sentence context than in isolation (Grosjean, 1980). In this dissertation the gating paradigm is used in a similar way--to compare the effects of sentence context in facilitating identification for spoken words and non-linguistic (environmental) sounds.

Listeners heard words or environmental sounds in snippets of increasing duration: 20 ms, then 40 ms, then 60 ms of signal, up until the full waveform was heard. For some participants, these stimuli were heard in isolation. For others, they were presented after a spoken sentence stem. The recognition point was defined as the duration of signal where the listener first correctly identified the sound or word.

2.1 Introduction

One of the hallmarks of both spoken and written language is the interaction of word recognition with the meaning of linguistic context (Morris & Harris, 2002; Simpson et al., 1989). A long-known example is semantic priming, in which words are recognized faster when preceded by a related word rather than an unrelated word (Hutchison et al., 2013; Meyer & Schvaneveldt, 1971). Meaningful sentence context affects word recognition as well. Gating studies, in which a spoken word is presented incrementally in small sound segments of increasing length, have shown that in a highly constraining sentence context (as opposed to a vague context), people need to hear less signal to identify a spoken word (Grosjean, 1980; Lorraine Komisarjevsky Tyler & Marslen-Wilson, 1986). Additionally, when people are asked to complete a sentence ending, they supply a word faster for a highly constrained sentence context than for a low constraint context (Staub et al., 2015).

Why is word recognition influenced by linguistic context? Extant word recognition models incorporate the effects of context information on lexical knowledge to varying degrees (see Dahan and Magnuson, 2006 for a review). Some models suggest that bottom-up input (e.g. the acoustic waveform of a spoken word or the visual input of a printed word) is the primary determining factor in the recognition process (e.g. Norris, 1994; Norris and McQueen, 2008). In these models, input is processed in a feed-forward manner through a series of transformations until a word is

recognized, and it is only at late stages, when the recognized word's meaning is being assessed, that it is integrated with and constrained by its surrounding context. Some models draw on evidence from priming studies to argue for a two-stage process in which bottom-up input causes widespread activation of many candidate words that could be consistent with the input, but are not constrained to be consistent with the broader context (for example, the word “bug” primes both “ant” and “spy,” even if the context suggests only the first interpretation; Swinney, 1979). According to such models, context then acts later, in the second stage of the model or “selection phase”, by facilitating the process of narrowing down from the population of activated candidates to the word that best fits the context (Swinney, 1979; Tanenhaus et al., 1995).

In contrast to these “input driven” models, interactive recognition models allow for continuous, on-line effects of context on word recognition. In such models, higher-level information, such as semantic associations, can alter processing at lower levels in a top-down manner via continuous integration (e.g. McClelland and Elman, 1986; Mirman et al., 2006). Shillcock and Bard (1993) were early critics of the modular, two-stage account, arguing that for closed-class words, immediate (as opposed to delayed) context effects support a continuous integration model. Further, eye tracking and fMRI studies have found context effects extremely early in processing, before other models incorporate context effects, and even before the bottom-up input unambiguously identifies a single word (Dahan, Magnuson, & Tanenhaus, 2001; Dahan & Tanenhaus, 2004; Magnuson, Tanenhaus, & Aslin, 2008; Revill, Aslin, Tanenhaus, & Bavelier, 2008). These studies suggest that lexical representations and semantic associations are being accessed simultaneously and integrated with each other continuously.

In recent years, interactive recognition models have been reinterpreted in light of predictive coding. In predictive coding accounts, language comprehension rests on neural predictions, based

on context or prior knowledge, that are continuously compared against input as it is being processed (e.g., Bonhage et al., 2015; DeLong et al., 2005; McRae et al., 2005; Metusalem et al., 2012; Pickering and Garrod, 2007). While some (e.g., Pickering and Garrod, 2007) argue that the speech motor system is integral to predictive coding, this view is by no means universal (see Hickok, 2012). ERP data from (Federmeier & Kutas, 1999) suggests that context allows the prediction of semantic features for upcoming words. However, it is possible that linguistic predictions could instead be happening at the level of sensory (e.g., auditory or visual) representations (cf. Lewis and Bastiaansen, 2015). It is also possible that predictions involve both semantic and sensory information (Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007; Kuperberg & Jaeger, 2016; Lupyan & Clark, 2015; McRae et al., 2005).

The notion that semantic associations influence ongoing and subsequent lexical processing (and vice versa) is supported by a substantial body of work on cross-modal effects. One example of this sort of cross-modal interaction is analog acoustic expression, the phenomenon in which modulations in pitch and speaking rate in speech affects the listener's understanding of the message (e.g., Shintel et al., 2006). Information from musical underscoring can affect speech understanding in a similar manner (Hedger et al., 2013). The effects of non-linguistic information on linguistic interpretation are not confined to the auditory modality. Tanenhaus and colleagues have used eye tracking to demonstrate that listeners make rapid on-line use of visual scene context in order to disambiguate spoken verbal instructions (Chambers et al., 2004; Tanenhaus et al., 1995). Such cross-modal effects on language have also been demonstrated via priming studies, in which visual or spoken words can facilitate processing of environmental sounds and vice versa (Frey et al., 2014; G Orgs et al., 2007; Guido Orgs et al., 2006; van Petten & Rheinfelder, 1995). Both words and environmental sounds have also been found to prime recognition of pictures (Chen

& Spence, 2011; Schneider et al., 2008). Concepts associated with words can also influence processing in other domains, as when words describing a particular direction of motion (such as the word “approach”) affect visual motion perception (Meteyard et al., 2007). Even when concepts are conveyed in a complex, non-linguistic way (e.g., an auditory scene), they can bias the people’s verbal labels for ambiguous environmental sounds (Ballas & Mullins, 1991). Thus, there is strong evidence that such cross-modal interactions occur bidirectionally, such that non-linguistic contextual information can cross-modally facilitate spoken word processing, and verbal context can facilitate processing of non-linguistic stimuli.

Despite the extensive documentation of cross-modal interactions between non-linguistic and linguistic information, the mechanisms behind these effects remain unclear. One possibility is that participants are covertly naming non-linguistic stimuli in order to guide processing words. This possibility is favored by a modular account of language processing, as according to this viewpoint, non-linguistic information cannot interact with encapsulated language modules until it is translated into linguistic information. It seems unlikely, however, that this is the case. Potter, Kroll, Yachzel, Carpenter, and Sherman (1986) asked whether printed sentences containing a picture substituted for a noun affected the speed and accuracy of plausibility judgments about the sentences. They reasoned that if pictures directly access the same system of concepts as words, rather than first being covertly named, then response times for plausibility judgments should be similar for “rebus” sentences (those containing a picture substituting for a word) and all-word sentences. This was indeed what they found. The results could not be easily attributed to covert naming, as previous work has demonstrated that picture naming takes too long to be occurring in Potter’s paradigm (cf. Oldfield and Wingfield, 1965). Other work also suggests against covert naming as the mechanism for context effects. The studies by Chambers et al. (2004) and Tanenhaus

et al. (1995) rely on sufficiently complex visual scenes that covert naming alone would not resolve the ambiguities present. Finally, it is highly unlikely that covert naming could explain analog acoustic expression effects, as listeners would have to translate the metaphoric meaning present in vocal pitch or rate information directly into words.

If covert naming is not responsible for the cross-modal priming effects that have been previously described, how does this process work? It is possible that, as suggested by Potter and colleagues, words and non-verbal stimuli such as pictures access a single conceptual system that is not grounded in language. In other words, the same neural representations of semantic information could be accessible via words and other meaningful non-verbal stimuli. Work by Zwaan and colleagues describes an effect opposite to covert naming, in which words activate “mental pictures” of the objects to which they refer (Zwaan et al., 2002), providing support for Potter’s hypothesis that words draw on a general conceptual system that is also used by nonverbal stimuli. In terms of a predictive coding framework, this would mean that predictions are sufficiently amodal (or multimodal) to interact easily with information from different domains. It is important to note that many models of word recognition are largely concerned with information involving phonemes and lexical representations, and have not been extended to representations that involve general concepts or “mental pictures” (McClelland et al., 2006; Mirman et al., 2006; Norris and McQueen, 2008; Strauss et al., 2007) although it is certainly possible to do so, especially considering the aforementioned studies, which suggest that this non-lexical information is readily and perhaps obligatorily activated by words.

In the present experiment, we asked how recognition of recognizable and meaningful, but non-linguistic, environmental sounds would be affected by linguistic context by using spoken sentence frames that were completed as a sentence by either a spoken word or an environmental

sound. An account of language understanding that isolates speech processing as a separate system from a broader conceptual system predicts that integrating non-linguistic inputs with preceding sentence context should be more difficult than integrating spoken word inputs. Non-linguistic information should be integrated as post-perceptual problem solving, requiring a covert naming step. This might incur heavy processing costs (over 500 ms for covert naming, cf. Oldfield and Wingfield, 1965). Based on prior research, however, it seems unlikely that strictly isolated speech processing would occur. Not only have cross-modal effects involving rapid interaction of many types of non-linguistic information with language been documented, but recent research has suggested that words and meaningful non-linguistic stimuli may have more in common in processing than previously thought given the neural resources involved in understanding both (Cummings et al., 2006; Frederic Dick et al., 2016; Leech & Saygin, 2011; A. P. Saygin, 2003; Ayşe Pinar Saygin et al., 2005). However, there is little research on how environmental sounds are understood, especially in comparison to speech sounds, and few studies directly comparing recognition and understanding of these two classes of sounds under a common contextual constraint.

Using this paradigm, we can measure whether the recognition or understanding of an environmental sound in a sentence frame relies on a reallocation of attention beyond what might be found for re-orienting to a new talker. Recognizing speech when there is a change in the talker in a sequence of utterances increases recognition time by about 40 ms (Nusbaum & Morin, 1992; Heald & Nusbaum, 2014b; Wong et al, 2004). Moreover, a similar recognition cost around 40 ms is found for recognizing musical notes when there is a change in instrument (Van Hedger et al., 2015). Given this recognition cost for shifts in signal source (voice or timbre) we might predict a similar cost or greater for recognition of an environmental sound in the context of speech.

Further, how does sentential (linguistic) constraint affect processing of environmental sounds roughly matched in meaning to spoken words? Spoken words following semantically constraining sentence frames will be recognized with less signal and responded to faster than words following less constraining frames, but whether this “constraint benefit” will occur to the same degree for environmental sounds is an open question. If the constraint operates at a purely linguistic level, we would not expect to see similar effects for environmental sounds.

Our experiments also allow us to address questions about how constraint may be acting via predictive coding. There is evidence that in particular situations, (e.g., when spoken word forms are predicted by context, or a non-linguistic motor action is predicted), the speech motor system generates predictions that aid comprehension (see Pickering & Garrod, 2007, 2013). However, the substitution of environmental sounds for spoken words presents a different kind of situation: the recognition and comprehension of the non-speech sound pattern cannot be aided directly by the speech motor system, because the non-speech sound patterns are not vocally generated. Thus, the question would appear to fall more in the realm of predicting an appropriate non-linguistic motor action based on the linguistic form combined with the environmental sound. However, participants in the current study are not deciding how to act in this situation, but simply recognizing the non-speech sound or understanding the sentence frame-plus-sound. Given that this falls outside the explanatory domain of speech-motor system predictive coding, parallel results for increasing sentence constraint in recognition or comprehension for spoken words and meaningful non-speech sounds would pose a theoretical challenge.

Similarly, we can also use this experiment to ask, outside of a speech-motor framework, how predictive coding operates to constrain processing. The specific acoustics of environmental sounds are likely much harder to predict than the acoustics of the same speaker from the sentence

frame saying the final word of the sentence, simply because the range of possible environmental sounds that could fit the intended meaning is larger. Thus, it follows that if neural predictions derived from context are largely at the level of sensory representations, constraining sentence contexts should be more helpful to spoken words than environmental sounds. However, if such predictions are more conceptual in nature, we would expect similar constraint benefits for spoken words and environmental sounds.

The gating paradigm (e.g. Grosjean, 1980) has been used to measure how much of a spoken word waveform is needed for recognition. In general, gating studies show that listeners can correctly identify spoken words even before the end of the word is heard, and that less of the word is needed for recognition in sentence context than in isolation (Grosjean, 1980). The present study used the gating paradigm in a similar way: to compare the effects of sentence context in facilitating identification for spoken words and non-linguistic (environmental) sounds. The main question for this experiment is whether the benefit of sentential constraint is similar for environmental sounds and spoken words.

2.2 Methods

2.2.1 Participants

There were 131 participants (78 female). Participants were from the University of Chicago community (mean: 19.89 years, range: 18 to 27 years); this includes students, staff, and residents of the surrounding area. Informed consent was obtained in accordance with IRB-approved protocol, and participants were compensated with their choice of one course credit or \$10 per hour of their time. Due to the linguistic and auditory nature of the task, participants were limited to those who reported speaking English as a first language, and who reported having normal hearing.

2.2.2 Stimuli

Stimuli consisted of sentence frames recorded at 44.1 kHz by an adult male native speaker of American English. There were two levels of sentence frame constraint, such that half were relatively constraining of the final word (high cloze probability, median = 0.87, IQR=0.25) and half were less constraining (low cloze probability, median = 0.16, IQR=0.33). We will refer to these two categories as “specific” (high cloze probability) and “general” (low cloze probability). Cloze probability was determined based on sentence completions from 66 Amazon Mechanical Turk participants. The last word of each sentence was the *target*, and was recorded separately from the sentence frame to avoid co-articulation confounds. There were 32 targets.

Each target word was a noun that could be represented by an environmental sound (e.g. “sheep”, and the sound of a sheep bleating, Appendix Table A1). Corresponding environmental sounds were taken from online databases such as soundbible, and if necessary resampled to 44.1 kHz. All sounds were then normalized to the same RMS level as the sentence frames and target words using Matlab. A small survey of lab members was conducted to ensure that the sounds were identifiable when heard in isolation. Mean duration of targets was 0.502 seconds for spoken words and 0.838 seconds for environmental sounds (see Supplement). Eight of the 32 environmental sounds involved repetition (e.g., the sound of a siren involves repeating pitch oscillations; see Supplement). Sentence frames and target sounds were digitally spliced together to create complete sentences. Half the resulting sentences terminated in spoken word targets, and half terminated in categorically matched environmental sounds. Stimuli were presented at 65-70 dB over stereo headphones.

2.2.3 Task

The task was based on Grosjean (1980); participants heard the targets in progressively increasing 20 ms waveform increments, and the task was to identify the target via a freely typed identification response.

There were six groups of participants tested using a 2x3 design crossing target type (between-subjects: sound or word) and context (between-subjects: general, specific, or isolated-- i.e. no sentence context—target). There were between 20 and 23 participants in each group (Table 2.1). Participants heard each target once for a total of 32 trials; the order of these trials was randomized. In the general and specific groups, participants heard the targets after the appropriately constrained sentence frame. In the isolated groups, they heard only the targets.

Table 2.1 Participant *n*'s, Expt 1. Numbers of participants in each group for experiment 1 after the exclusion of the three participants described.

Condition	<i>n</i>
Isolated Sounds	23
General Sounds	22
Specific Sounds	20
Isolated Words	21
General Words	22
Specific Words	20

Targets, either isolated or at the ends of sentence frames (not sentence frames themselves) were presented in successively increasing segment lengths by 20 ms increments until either 1) the entire sound was presented, or 2) the participant's identification responses remained stable for 20 gates in a row. As soon as either 1) or 2) was reached, the participant heard the whole sound, and was asked to identify it one last time.

2.2.4 Data collection

Participants' responses and corresponding gates were collected in Matlab 2014 with Psychtoolbox 3 (Brainard, 1997; Kleiner et al., 2007) and recognition accuracy was scored by hand. For each target, the gate of recognition was defined as the gate of the earliest correct response, to conservatively determine the minimal signal supporting recognition. For each target, recognition points were calculated by converting gates of recognition to seconds and averaging across subjects in the same condition. This resulted in three recognition points for each target: general context, specific context, and isolated target.

2.2.5 Data Analysis

Three participants were excluded: one for reporting not being a native English speaker after the experiment (condition: environmental sounds/specific frames), one for previous participation in the study (condition: spoken words/general frames) and one for not following instructions (condition: environmental sounds/specific frames).

Percent correct responses for each target in each condition were calculated across participants. While we expected that most participants would be close to ceiling for words, we expected some misidentification of environmental sounds, particularly in the isolated target condition. Any target that dropped below 70% correct in any condition across participants was removed from further analysis. Five environmental sounds (baby laughing, creaky door, clock ticking, papers ruffling, and sword being unsheathed) and one word ("horn") were excluded from the final analysis due to this level of poor recognition in the isolated condition. In conditions with sentence context, recognition performance was always well above this level.

After exclusions of low-accuracy targets and problem participants, the mean recognition point for each target in each context was calculated across subjects, and data points outside 2.5

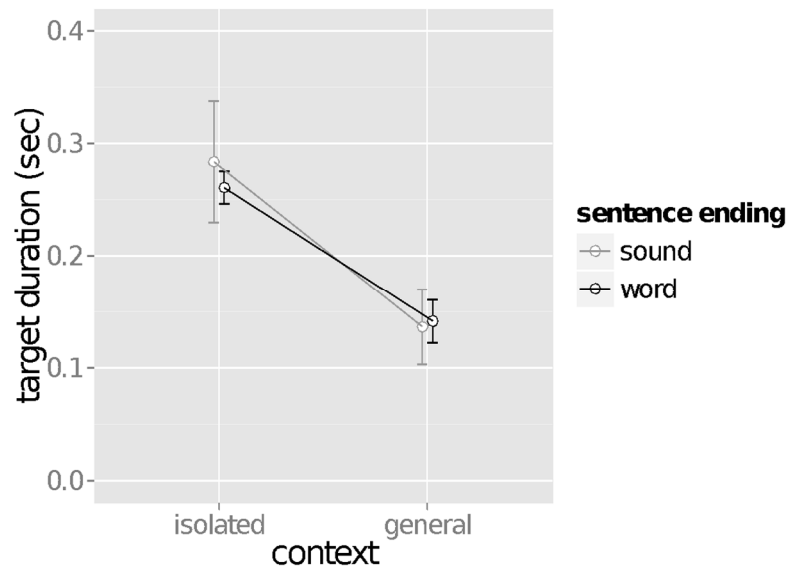
standard deviations from the mean were removed. We excluded recognition points from the specific context condition from further analysis. For a majority of the trials with specific sentence frames, participants answered correctly on the first gate (after hearing only 20 ms of the target), regardless of target type (sound or word). For highly constrained frames, guessing the ending may be too easy when there are no foils or distractors, leading to a ceiling effect. Thus, further examination of the data from the specific sentence frames is not informative.

A Repeated Measures ANOVA was performed on the resulting recognition gate data. The factors modeled included constraint level (isolated versus general frame), and target type (word or environmental sound). Excluding words corresponding to the five poorly recognized environmental sounds, and excluding the environmental sound corresponding to the excluded word “horn” did not substantially change the ANOVA results, so these stimuli are included in further analyses.

2.3 Results

There was strong evidence for a context effect on recognition points for both environmental sounds and words (Figure 2.1), such that for both target types, adding sentence context significantly decreased time to recognition. Participants recognized targets in isolation after hearing an average of 272 ms of waveform; this dropped to 139 ms of waveform for targets occurring after a general sentence frame. Thus context significantly reduces the amount of waveform needed to recognize acoustic targets ($F(1, 26) = 115.1, p < 0.001, d = 4.2$). For spoken words, sentence context reduced the amount of waveform needed for recognition from 261 ms to 141 ms. This reduction of 120 ms by general sentence context compared to isolated words is similar to the effects of general context compared to isolated targets reported in previous gating studies (Cotton & Grosjean, 1984; Grosjean, 1980; Lorraine K. Tyler & Wessels, 1983).

Figure 2.1 Duration of target needed for recognition. Sentence context shortens the amount of target waveform needed for participants to recognize both environmental sounds and words. Error bars represent ± 1 standard error of the mean (SEM).



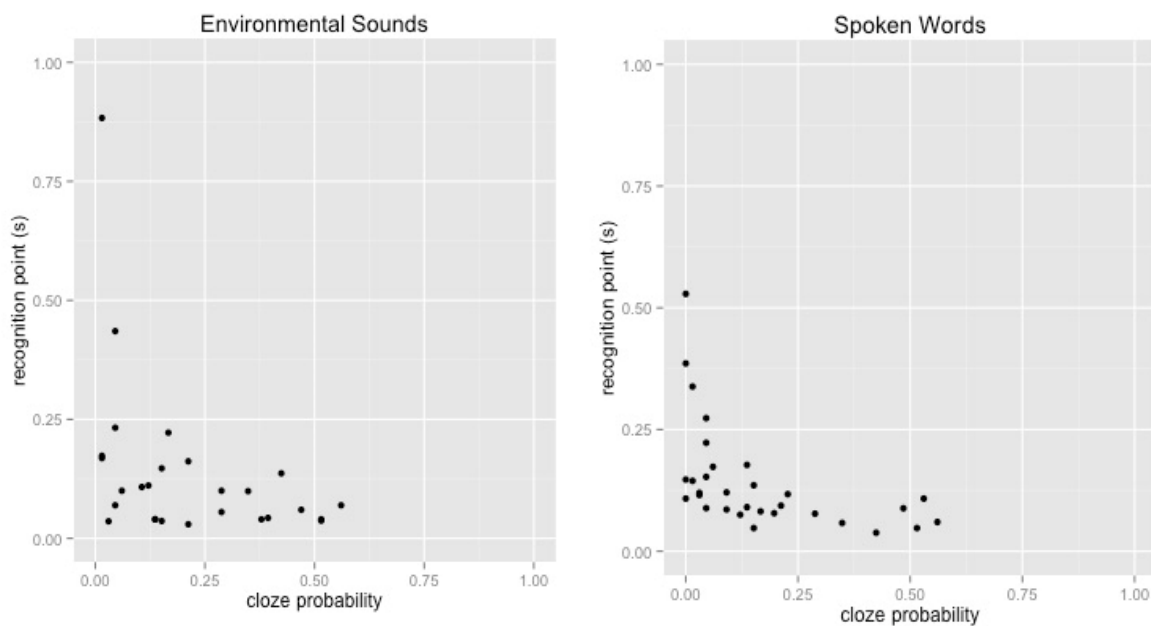
It is also the case that general sentence context reduces the amount of waveform needed for the recognition of identifiable environmental sounds by 148 ms: from 284 ms for isolated sounds to 136 ms for sounds in linguistic context. Clearly, the information in sentence context is informative about the identity of meaningful sounds.

While recognition points for sounds in isolation appear to be slightly later than those for words, and slightly earlier in the presence of a general sentence context (Figure 2.1), there was no significant difference between recognition points for sounds and words ($F(1, 26) = 0.007, p > 0.25, d = 0$). Moreover, there was no significant interaction between context and target type (Figure 2.1, $F(1, 26) = 0.64, p > 0.25, d = 0.31$).

Our general sentence frames represented a range of cloze probabilities. If the reduction in recognition point observed for targets in sentence context is truly due to facilitation involving the conceptual meaning of the sentence (as opposed, for example, to a purely psychophysical effect of any sound preceding the target), there should be an inverse relationship between cloze probability

and recognition point for the sounds and words presented in sentence context. This was indeed the case; there were significant nonlinear inverse relationships as shown by Spearman's rank order correlations (Figure 2.2, Sounds: $\rho[25] = -0.45, p = 0.02$; Words: $\rho[29] = -0.71, p = 7.3\text{e-}6$). The strength of these inverse relationships was not significantly different between sounds and words ($z = 1.45, p = 0.15$).

Figure 2.2 Cloze probability and recognition point. There is an inverse nonlinear relationship between cloze probability of the general sentence frame and recognition point of the target following the sentence frame for both sounds and words.



2.4 Discussion

Sentence context significantly reduces the amount of signal needed for recognition of both spoken words and environmental sounds, even when the sentence frame is general. This reduction happens to the same extent for both sounds and words; there is no evidence for any interaction between context and target type. From a predictive coding perspective, these results suggest that sentence frames support neural predictions that are mainly conceptual rather than sensory, because

context helps word and sound recognition to the same extent, even though sensory predictions should be less precise for environmental sounds. In the general constraint condition, the inverse relationship between recognition point and cloze probability indicates that listeners can use fine-grained meaning constraint information to facilitate processing of both sounds and words to a similar extent. This relationship also indicates that the earlier recognition points to items in general context (as opposed to isolated items) is not due merely to sound preceding the targets, but to the way sentence constraint and meaning interact with the interpretation of the target sound or word.

The gating paradigm is intended to mimic the recognition of a sound as it unfolds in time. However, sounds do not stop in midstream typically, so there is an aspect of unnaturalness to the gating study. Furthermore, listeners in a gating study are not typically pressured to respond as quickly as possible, although Tyler and Wessels (1985) did not find that a speeded naming version of the gating task changed the results. Thus it has been claimed that gating is not qualitatively different from typical speech perception (Cotton and Grosjean, 1984; Tyler and Wessels, 1985, although see Allopenna et al., 1998 for evidence that the gating paradigm may distort typical perception). Given that listeners presented with environmental sounds were not also identifying spoken word targets (between-subjects design), it seems unlikely that the environmental sound recognition was influenced by a specific word recognition strategy. Yet, it is also the case that listeners may use a more cognitive-inferential approach given the nature of the gating task than would be the case in normal speech perception. Furthermore, the gating task is an identification task rather than a meaning comprehension task. To address these concerns, we designed Experiment 2 (Chapter 3) using fluent speech and a comprehension task.

Chapter 3: Can non-linguistic sounds be quickly and easily understood in linguistic context?

3.1 Introduction

The Gating Study tests how recognition of words and environmental sounds is affected by sentence context. This study, in contrast, tests understanding of the meaning of the sounds in sentence context. Subjects were instructed to determine whether sentences were “understandable” or “nonsense” as quickly and accurately as possible. Half of the sentences were created to be understandable (the last word or environmental sound matched fit with the meaning of the sentence) and half were nonsense (the last word or its matched sound was highly implausible in the context of the sentence). Additionally, half of the stimuli ended in a spoken word target and half ended in an environmental sound target. If the similarity of sentence frame context effects for speech and non-speech in the first experiment were governed by a slower problem-solving strategy rather than a more fluent perceptual understanding process, speeded processing of spoken sentences should show a different pattern from the gating task.

3.2 Methods

3.2.1 Participants

Participants were selected from the same population as in Experiment 1 (Chapter 2), but individuals were excluded from participating in both experiments to avoid effects from repeated exposure to the stimuli. There were 31 participants (15 female, mean: 20.93 years, range: 18 to 38 years). Informed consent was obtained and participants were paid at the same rate as in Experiment 1.

3.2.2 Stimuli

The stimuli were taken from the same set as the previous study, but additional “nonsense” sentences were created as distractors. These were constructed by rearranging the ending words of the sentences. The resulting sentences were verified in a short written survey to ensure that they were not easily construed to make sense. For example, the sentence “He closed his winter jacket with the zipper,” which is understandable, might be rearranged with the target “train” to form the nonsense sentence “He closed his winter jacket with the train.” Thus, the “understandable” nature of the sentence depended on the last word of the sentence, which was replaced by a categorically matched environmental sound for half the stimuli. This 2x2x2 design gave rise to eight possible types of sentences: general/specific constraint x word/sound target x understandable/nonsense (all sentences are available in the Supplement).

Stimuli were presented at 65-70 dB over headphones. The experiment was coded in Matlab 2014 with Psychtoolbox 3.

3.2.3 Task

The participants’ task was to decide, as quickly and accurately as possible, whether the sentences were “understandable” or “nonsense”. They responded by pressing one of two labeled keys, the side (right or left) of which was counterbalanced across participants. This task was chosen, as opposed to a sound recognition task, because recognition could require participants to “name” the sound. For this reason, we used an understandable/nonsense judgment rather than a sound recognition task (as in Potter et al., 1986) in order to judge if environmental sounds might also convey meaning without a naming step.

Each session of the experiment consisted of 5 blocks. The first block contained 20 practice spoken sentences, all ending in spoken words, for task familiarization. There was a 2x2 constraint

x meaning design in the practice block, such that general-specific and understandable-nonsense conditions were equally represented. The remaining four experimental blocks consisted either of 32 spoken sentences with targets as environmental sounds (sound blocks) or of 32 sentences with targets as spoken words (word blocks). Half of the participants received blocks in a sound-word-sound-word order, and half the participants in word-sound-word-sound. Within each block, each target occurred exactly once. Targets never appeared in the same context more than once. For example, if “sheep” was heard in a general-frame, understandable context in block 1, it might appear in a general-frame, nonsense context in block 2, and a specific-frame, understandable context in block 3. Within each block, the order of stimuli was randomized. Participants received short breaks between blocks.

3.2.4 Data collection and analysis

Response times (RTs) were recorded in Matlab with Psychtoolbox 3. RTs were defined as the time between the onset of the sentence’s last word/sound (“target onset”) and the participant’s button press. Responses were classified as either 1) correct (the person responded with “understandable” or “nonsense” as appropriate after the onset of the last word or the sound), 2) incorrect (the person assigned the wrong understandability status to the sentence after the onset of the last word) or 3) guessing (the person responded before the onset of the last word, yielding a negative RT).

Three subjects were excluded, one due to low English proficiency and two for failure to follow instructions. Of these two, one was excluded for performance below 80% correct in the general/understandable/sounds condition, and one was excluded for excessive guessing, evidenced by negative RTs in 66% (general/nonsense/sounds) and 80% (specific/nonsense/sounds) of trials. Only understandable sentences with correct responses (i.e., understandable sentences to which

participants correctly responded “understandable”) were included in further analysis of RTs; nonsense sentences were treated as distractors and their RTs were not analyzed further. Incorrect trials and trials where participants guessed (negative RTs indicate responses before the target onset) were also excluded from further analysis.

A Repeated Measures ANOVA was performed on RT data. The factors modeled included constraint level of frame (general or specific), and target type (word or environmental sound), both within-subjects. As ANOVAs are not well-suited to modeling categorical data, we used a logit mixed model approach for the percent correct data (Jaeger, 2008). Using the *lme4* package in R (D. Bates, Maechler, Bolker, & Walker, 2015), we compared model A (main effects of constraint and target type, with random intercepts for subjects and sentence endings) with model B (model A plus constraint * target type interaction). There was no evidence that including the interaction term improved the performance of the model ($\chi^2(1) = 0.028, p > 0.8$); therefore we performed further analyses with model A.

3.3 Results

Accuracy was high in all conditions, close to 90% or higher (Figure 3.1). There was significantly higher accuracy in specific (compared to general) context conditions ($p = 0.009$, Table 3.1); and significantly higher accuracy in word (compared to environmental sound) conditions ($p = 0.002$, Table 3.1). There was no evidence for an interaction between these terms.

Figure 3.1 Percent correct for RT task. Percent correct responses for meaningful vs. nonsense judgments about sentences ending in either words or sounds. Accuracy was high ($\geq 90\%$) for all conditions but higher for words than sounds. Mean percent correct $\pm$ SEM error bars shown.

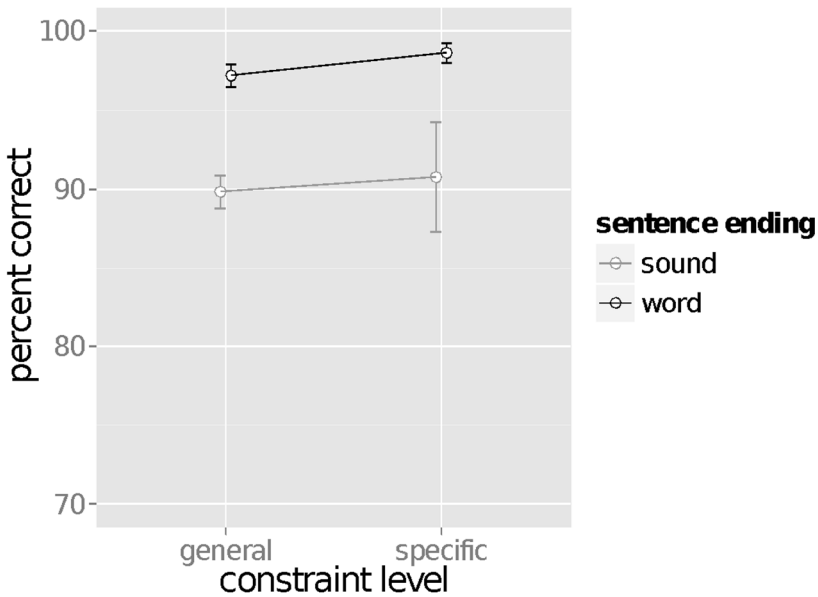
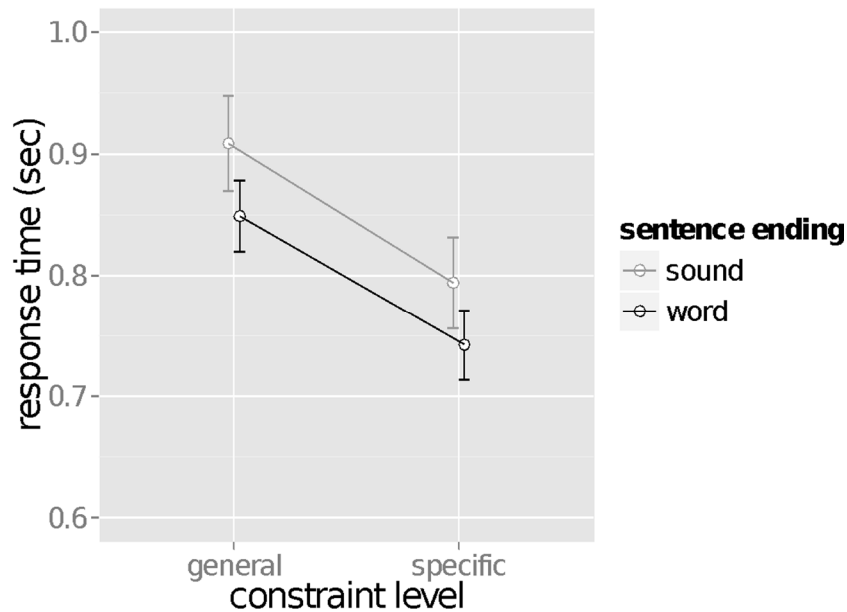


Table 3.1 Coefficients for factors in a logit mixed model for percent correct data.

Factor	Coeff	SE	Z	p
intercept	4.26	0.41	10.36	2E-16
target type	-1.45	0.47	-3.10	0.002
constraint	0.66	0.25	2.62	0.009

For response times, there was a main effect of sentence constraint, such that meaningfulness judgments for specific (i.e. high constraint) sentences were faster than general (i.e. low constraint) sentences (Figure 3.2, mean RTs 768 vs. 879 ms, $F(1, 27) = 52.34$, $p < 0.001$, $d = 2.79$). The main effect of target type (i.e. spoken word versus non-speech sound) approached significance ($F(1, 27) = 3.67$, $p = 0.066$, $d = 0.74$), but the interaction of target type with constraint level was not significant ($p > 0.25$).

Figure 3.2 RTs for meaningful vs. nonsense judgment. Response time measured from target (i.e. last item in sentence) onset shows that sentence constraint speeds meaningfulness responses similarly for stimuli ending in spoken words and environmental sounds. Mean RT $\pm$ SEM error bars shown.



3.4 Discussion: Implications of Experiments 1 and 2

Listeners have little difficulty understanding spoken sentences that end in “sound effects” that substitute for spoken words. Though accuracy for all-word sentences was higher than for sentences ending in non-speech environmental sounds, overall accuracy was quite high in both conditions at 90% correct or above, suggesting that even such an unusual, unfamiliar task as understanding environmental sounds in sentence context is not much more difficult than the everyday task of understanding a normal sentence. How can we explain this effect? It is unlikely that listeners use a covert naming strategy—naming in other studies takes several hundred milliseconds, and we find no delays of this magnitude. In Experiment 2 (Chapter 3), RTs were on average approximately 60 ms slower for environmental sounds, and this difference did not reach significance. It is difficult to justify any model of language understanding that depends on a dedicated speech processor given that non-speech can be understood in spoken sentences and

derive the same contextual benefit as spoken words. In light of the current results, a purely lexical model (e.g., the distributed Cohort model, Gaskell and Marslen-Wilson, 1997) cannot explain the results without substantial and fundamental changes to the underlying assumptions. Of course, environmental sounds, like spoken words, could be treated as meaning-associated patterns in the same system albeit with a different kind of feature base and pattern processing system. Tyler et al. (1996) suggested that the top tier of the nodes in the TRACE model (i.e., words) could be linked to an even higher layer, containing nodes that represent concepts or semantic information; context effects could occur by activity in this layer feeding back onto the lower levels. If there is a separate network of neural representations of environmental sounds that is also one layer below the conceptual layer, then activity from sentence context in the conceptual layer could feed back onto the sound recognition layer, affecting recognition.

The present experiments demonstrate that sentential context can provide equal benefit for both spoken words and non-speech sounds in recognition of the targets as well as in understanding a whole sentence, as if the non-speech sound is a natural part of the sentence. From a predictive coding perspective, this suggests that even in situations where it is not possible to make neural predictions via motor systems (as our environmental sounds do not have clear speech or other motor representations), some other type of predictions can constrain processing (cf. Hickok, 2012). In other words, these results suggest that neural predictions can occur at a conceptual level in constraining recognition and comprehension. The acoustics of environmental sounds are more variable and less predictable than the acoustics of words from the same speaker that the participants heard in the sentence frame. Therefore, it follows that accurate sensory predictions for environmental sounds must be more difficult to form. Moreover, if predictive coding is highly statistically dependent as suggested by Kuperberg and Jaeger (2016), the statistical rarity of

environmental sounds serving to complete sentence frames would pose a substantial challenge. If participants need to rely heavily on precisely tuned sensory predictions from context to constrain processing, we would not have found such similar context effects for both environmental sounds and words. Of course, this is not to say that sensory predictions do not contribute to language understanding; merely that conceptual predictions can be sufficient to constrain processing to much the same extent. Thus, a processing framework like that proposed by Lupyan and Clark (2015) could apply, with the caveat that the importance of the “low-level predictions” is variable based on the sensory quality or statistical properties of the stimuli. Future work could compare neural responses to environmental sounds and spoken words in sentence context. Lewis and Bastiaansen (2015) theorize that low and high gamma-range oscillations represent propagation of top-down predictions and computation of prediction errors, respectively. The relative importance of these processes for recognizing and understanding environmental sounds versus spoken words in sentence context can be tested with our stimuli set and EEG time-frequency analyses.

While in many respects environmental sounds and words behaved similarly in our experiments, the 60 ms processing cost for environmental sounds was an important difference. It is unlikely that this difference is an artifact of the properties of the stimuli sets used, because there was no significant difference in amount of waveform needed for recognition of meaningful sounds and spoken words in Experiment 1 (Chapter 2). If the small RT difference observed for meaningfulness decisions for normal and rebus sentences was due to longer average recognition time for the environmental sounds than the words, we might expect systematic differences in recognition points to show up in the gating study, but none were apparent (i.e. no main effect of target type or context-target type interaction).

One explanation for this processing cost is that in order to switch from interpreting the experiment's male speaker to an environmental sound, listeners may have to shift attention in some sense. For example, when an object appears in an unexpected location or changes form unexpectedly, observers engage in shifting attention (Yantis and Serences, 2003). This is similar to the ~40 ms processing cost that is incurred when talkers change (Nusbaum and Morin, 1992). Interestingly, music work suggests that a processing cost of about this size is not unique to language, but could reflect attention reallocation in other types of auditory processing as well. Van Hedger et al. (2015) found a processing cost near 40 ms for both switching between timbres and between octaves in a paradigm where absolute pitch possessors were asked to respond to certain target notes. Perhaps the additional 20 ms delay observed in our paradigm reflects switching attention to a farther-away modality, as both the talker change and timbre/octave change experiments took place in the same modality. Regardless, the need for attention reallocation did not impair participants' ability to make use of context information to speed processing. This suggests that signal source changes (e.g., timbre, voice, speech/non-speech) can slow recognition but still engage the same processes that are also needed for sentence understanding (Heald and Nusbaum, 2014a).

Another potential source for the 60 ms environmental sound processing cost is the relative unfamiliarity of the sounds, particularly in sentence context. It may be the case that, on top of reallocation of attention, a processing lag is imposed by virtue of the environmental sounds being less frequently heard than spoken words, and almost never heard as meaningful items in a spoken sentence as they were in our experiment. This interpretation is supported by the large body of work on frequency effects in spoken word recognition (e.g., Dahan et al., 2001; Luce & Pisoni, 1998), as well as by Van Hedger et al. (2015), who found that responses were significantly slower to

octaves with which the participants had less musical experience based on the instrument that they played. An effect of reduced familiarity can also be explained as a predictive coding disadvantage, as more commonly encountered stimuli are more likely to have stronger neural representations that can be activated as predictions. Perhaps because in this paradigm, it is easier to form sensory-level predictions for spoken words, the words can benefit from stronger low-level predictions that speed processing. It is interesting to note that this cost is only 60 ms, which implies that strong conceptual-level predictions are able to accommodate substantial variance in the input to the system—even when that input is no longer linguistic. In any case, it follows that if at least part of the sound-word difference is due to differential experience, training should be able to narrow this gap. Future experiments might train participants on a subset of environmental sounds by pairing them with pictures, and then compare participants' speed of integrating trained sounds versus words with sentence context in order to address this question. ERP analyses could also reveal differences in the time courses of sound and word processing that are too fine-grained to be picked up by behavioral studies such as the ones reported here: this will be explored in the next chapter.

It is worth noting that many extant models of speech recognition do not explicitly account for how such seamless interactions between verbal and nonverbal stimuli might be happening in terms of a mechanism. From a perspective concerning how these models are used to explain language understanding, our results are interesting because they suggest some updates to these models. A number of different studies (Shintel & Nusbaum, 2007; Zwaan & Pecher, 2012; Zwaan et al., 2002) have demonstrated that language, specifically an intact clause referring to an object, can facilitate understanding of that object. In these cases, an understandable linguistic form (a word, a clause, or a sentence) refers to a non-linguistic object and speeds processing for that object. The present experiments go one step further, because a non-linguistic object is not referred

to by a complete sentence, but is directly incorporated into the sentence as if it were itself a linguistic form. In other words, the sentences in our Experiment 2 do not function as complete ideas without incorporating the meanings of the environmental sounds. From the position of a general cognitive processing system in which linguistic forms hold no special or privileged status either by virtue of their high degree of statistical association or by virtue of specialized mechanisms subserving their processing, the present results are predicted and expected. Even without a claim of modularity however, connectionist models—e.g., Shortlist B, updates of TRACE, interactive Hebbian models—that are used to model speech recognition would not fluently treat a non-speech sound as substitutable for recognition or understanding purposes (McClelland et al., 2006; Mirman et al., 2006; Norris & McQueen, 2008; Strauss et al., 2007). Learning models of speech (Jurafsky & Martin, 2000; Kuperberg & Jaeger, 2016; McMurray, Aslin, & Toscano, 2009) operate on associative statistical principles, and non-speech sounds do not occur in these contexts. While a general cognitive processing perspective can easily account for the rapid shift of attention to an unlikely sound if that sound's interpretation fits with the contextual meaning of the antecedent frame, most language models do not take this into account.

The present results demonstrate that there is no evidence for a differential effect of context for meaningful non-speech sounds relative to matched spoken words. Whether the meaning is derived from a vocal-tract produced utterance, or from environmental generators, context appears to similarly limit recognition and understanding. These results highlight the importance of conceptual meaning in context effects, as these two types of signals are vastly different in their acoustic properties, sources, and statistical occurrences, but are well matched in terms of conceptual meaning.

In summary, our results have strong implications for language processing theories. They add to the considerable body of evidence arguing against modular, encapsulated language processing by demonstrating that understanding words and environmental sounds in the same sentence requires an attention switch akin to switching between two talkers, rather than a deductive or covert naming strategy. Moreover, our results suggest that, if predictive coding is responsible for the facilitative effects of constraint on language processing, it is likely that predictions involving general conceptual representations (as opposed to low-level sensory predictions), are largely sufficient to drive constraint effects. Finally, our results suggest that models of speech comprehension that largely rely on lexical attributes should be modified to include a larger contribution from general cognitive processes that take conceptual meaning into account.

Chapter 4: Do the neural responses to speech and non-speech reveal important processing differences related to understanding them in linguistic context?

The purpose of this study is to examine neural responses to environmental sounds in sentence context. Whereas the first two studies examine behavior (recognition points and response times), it is unclear what neural processes are responsible for the similarities and differences found. In this study, participants listened passively to the same sentences from the response time study. Again, half of the sentences were understandable and half were nonsense (randomized within blocks) and half ended in words versus environmental sounds (blocked). The participants were instructed simply to listen quietly, and think about whether the sentences they heard made sense or not. The period of time immediately following the onset of the sentence-final items was analyzed by averaging together the responses for meaningful, nonsense, speech, and non-speech stimuli to obtain the event-related potentials (ERPs). Using a nonparametric randomization procedure, we compared the scalp topography maps of the ERPs in the different conditions to test whether responses were significantly different.

The main ERP examined in this experiment was the N400. The N400 is a negative-going ERP occurring anywhere from 200 – 600 ms (with an average of 400 ms) after the onset of a stimulus (Kutas & Hillyard, 1980). The N400 is well known to be congruency-sensitive; that is, it is larger in amplitude in response to words that do not fit with the preceding context (Kutas & Hillyard, 1980; Kutas & Federmeier, 2011). It has been demonstrated to have widely distributed sources including the superior and middle temporal gyri, parahippocampal areas, hippocampus, temporoparietal junction, and dorsolateral prefrontal cortex (Kwon et al., 2005; Simos, Basile, &

Papanicolaou, 1997). It therefore seems likely that as proposed by Kutas & Federmeier (2011), the N400 reflects the activity of a “multimodal semantic system.” Its larger amplitude in response to incongruent stimuli may reflect greater integration difficulty (e.g., Hagoort, Baggio, & Willems, 2009) or an error correction process (e.g., Rabovsky & McRae, 2014). Whatever the exact mechanism, given that it appears that the N400 reflects some sort of process of integrating a new item with preceding context, we can therefore ask if both environmental sounds and words elicit context-sensitive N400s in sentence context. It is possible that the neural mechanisms responsible for understanding the conceptual meaning of language and integrating words with preceding context cannot be shared with environmental sounds. In such a case, we might not expect to see reliable N400s to environmental sounds in sentence context at all. Alternatively, perhaps environmental sounds are too different from spoken words to be integrated immediately with them in the context of a fluid, spoken sentence. In such a case, sounds might always be processed as mismatches, and N400s to them might occur regardless of whether or not the sound is congruent with the preceding context. Finally, if similar, context-sensitive N400s are found for environmental sounds and words, this suggests that similar processes underlying the integration of these items with preceding context are at play.

4.1 Conceptual Introduction

The question of whether speech understanding is mediated by a specialized neural system (e.g., Grodzinsky, 2000; Liberman & Mattingly, 1985) or more general neural mechanisms (Christiansen, Allen, & Seidenberg, 1998; Dick et al., 2001; Kleinschmidt & Jaeger, 2015; Leech et al., 2009) is a longstanding theoretical issue. This debate has often focused on the characteristics of language that set it apart from other kinds of information (e.g., Chomsky, 1986; Fodor, 1983), but more generally, it addresses age-old questions about the balance between specialization and

modularity on the one hand, and distributed processes and domain-general mechanisms on the other hand.

Environmental sounds (ES) are auditory patterns that are meaningful, but not “linguistic”: they lack internal phonological segments or higher-order linguistic structure, and in most cases are not produced by the human vocal tract. They are, however, easily recognized and categorized (Gygi et al., 2007; Warren & Verbrugge, 1984), and can be combined with each other or with words in order to form meaningful concepts (Ballas & Mullins, 1991). If speech perception is carried out by a separate dedicated processing mechanism, any similarity between understanding ES and spoken words would be due to chance and should not be systematic, whereas if perception and comprehension of spoken sentences is mediated by general auditory and cognitive processing, there should be substantial overlap between these processes.

An important characteristic of human language is the facilitative effect of context; it is well known that constraining contexts speed processes such as word recognition and sentence completion (Morris & Harris, 2002; Staub et al., 2015). Therefore, one way to address whether there are similarities between the processing of nonlinguistic stimuli and words is to ask whether we can understand a sentence that substitutes an ES for a spoken word. In doing so, we are asking whether the processes for understanding an item in light of its preceding context differ substantially between these types of stimuli. Readers easily understand “rebus” sentences in which a picture replaces a printed word (Potter et al., 1986), but it is possible that speech perception, as a more basic system than reading in human development (cf. Dehaene, 2011), might operate as a separate modular system (cf. Fodor, 1983). We have previously compared behavioral measures of perception of spoken sentences that end in either a word or an ES. In a gating paradigm (see Grosjean, 1980) constraining sentence frames (e.g., “he bought diapers for his ____”) reduced the

duration of signal needed for recognition compared to general frames (e.g., “his back hurt from holding the ____”) similarly for words and ES. Further, response times for congruency judgments were similar for sentences ending in words and ES (Uddin, Heald, Van Hedger, Klos, & Nusbaum, 2018). While non-speech stimuli can be understood, even substituted for words, in a spoken linguistic frame, similar patterns of behavior do not unequivocally indicate similar underlying neural processing (cf. Reuter-Lorenz, 2002). Thus it is important to assess whether neural responses differ between ES and words when they are understood in spoken sentence contexts.

Of course, it is important to note that neural responses to ES and words might differ for reasons unrelated to their interaction with context. There are substantial acoustic differences between environmental and speech sounds (e.g., Lewicki, 2002). ES can be derived from a wide variety of sources, and can range from man-made sounds such as machinery, to nature-related sounds such as water rushing or an animal vocalizing. Therefore, ES have a much wider variety of sources and acoustic features than words pronounced by a single speaker (Lewicki, 2002). Moreover, fMRI studies show that ES recruit different cortical areas than speech (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; J. W. Lewis, 2005; Vouloumanos, Kiehl, Werker, & Liddle, 2001), although more recent work suggests that cortical representations of environmental sounds and speech overlap substantially (Dick et al., 2001; Leech et al., 2009; Leech & Saygin, 2011). Therefore, while we expect differences in neural responses based on acoustic pattern and stimulus frequency differences between ES and spoken words, the important question is whether there are differences that reflect fundamentally different processes for understanding these stimuli in sentence context.

In processing an utterance’s meaning, the N400 is a negative-going ERP in human EEG that can arise from a mismatch of a word with preceding context (Kutas & Hillyard, 1980a).

Semantically incongruous words elicit larger negativities approximately 400 milliseconds after word onset. Kutas and Hillyard postulated that this comes from neural processes related to sentence meaning repair, or reprocessing of the unexpected word in contextual integration. The N400's sensitivity to expectation violations can be used to investigate neural responses to ES in spoken sentence context.

Stimuli do not have to be linguistic to elicit an N400; they can also be pictures, environmental sounds, or other meaningful items (Kutas & Federmeier, 2011). While it is known that environmental sound probes presented after written word, picture, or spoken word primes elicit more negative N400s when probes and primes are incongruent (Cummings et al., 2006; Guido Orgs et al., 2006; van Petten & Rheinfelder, 1995), this has never been tested in the context of a fluent speech sentence frame. This is an important distinction, because in previous experiments, primes and probes are single words or ES separated by a period of silence. Such isolated words or sounds stand alone as concepts. The present experiment, in contrast, presents either words or ES as the final item in a continuously presented sentence frame which is related in its meaning to a concluding final noun, if ending in speech. Before the final item is presented, it is not apparent whether it will be congruent or incongruent with the meaning of the antecedent sentence frame. Therefore, given a spoken sentence fragment, listeners continuously incorporate the meaning of each word with previous words in order to understand the sentence context before the final item arrives, at which point this final item is understood in this context. If the final item is a spoken word, listeners will certainly recognize and understand this word in the linguistic context established by the sentence frame. The question is what happens when the final item is not a word but is an environmental sound. Due to the continuous nature of the sentence stimuli in this paradigm, if there are processing costs, i.e. slower processing or obligatory extra processing

steps, related to ES being more difficult to understand than words in context, such costs should be more apparent than in a paradigm where isolated primes and probes are presented several hundred milliseconds apart from each other. This is because 1) extra processing, or delays in processing, could manifest in the time between the presentation of primes and probes, and 2) a spoken sentence requires continuous attention and interpretation which will likely tax mechanisms for understanding beyond a prime/probe pair.

Though N400s to non-linguistic stimuli are routinely found using prime/probe type designs, it is possible that understanding ES in a fluent spoken sentence might draw on different neural mechanisms. While our behavioral work suggests against this possibility (Uddin et al., 2018), if this is true we might not expect any reliable N400 effects for ES in our experiment. On the other hand, ES in such a context may always be processed as incongruent because they are so different from speech. In this case, we would expect N400s to all ES regardless of the relationship between the meaning of the sound and the meaning of the preceding sentence frame.

If ES and spoken words are assigned to meaning, and that meaning is combined with preceding context in a similar fashion, we might expect a generally similar pattern of N400 results—higher-amplitude for sentence-frame incongruent final items—for both ES and word targets. However, it is also possible that ES are recognized and understood in these sentence frames similarly to words, but are more difficult to process in this context. While measured response times do not support this, as they are not substantially slower for ES (Uddin et al., 2018), increased processing cost could yield a delayed N400 without slowing response times (e.g., Ardal, Donald, Meuter, Muldrew, & Luce, 1990).

Finally, there is ample evidence that the constraint level of a sentence affects understanding of words in that sentence (e.g., Staub et al., 2015). These constraint effects extend to the N400: for

words congruent with the preceding context, the N400 is larger in amplitude for low-constraint sentences (Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007; Kutas & Hillyard, 1984). However, for words incongruent with or highly unexpected in the preceding context, the N400 does not appear to depend on constraint level (Federmeier et al., 2007; Kutas & Federmeier, 2011). While obvious factors such as congruence with context may lead to similar neural responses for ES and words, it is possible that constraint information is a finer nuance of sentence processing that will only affect words. There is at least some precedence for this idea: Hendrickson, Walenski, Friend, & Love (2015) examined N400s to ES or spoken words presented after a picture. The ES or words were either matches (i.e., congruent with the preceding picture), near violations (incongruent but conceptually related), or far violations (incongruent and conceptually distant). For words, voltage in the 300-400 ms post-onset time window was graded based on degree of congruence with the picture, but for ES, near violations and matches were statistically indistinguishable. It is possible that in a fluent sentence paradigm, where ES must be rapidly understood in an unfolding spoken sentence, such differences between words and ES could be magnified. In our study, an interaction of sentence ending type (ES or word) with constraint would suggest that constraint differentially affects the interaction of words and ES with the context supplied by a full spoken sentence.

4.2 Methods

4.2.1 Participants

Participants were 23 (8 female, 13 male, 1 agender, 1 genderfluid) adults from the University of Chicago and surrounding community. Their mean age was 22.1 years (SD: 3.7, range: 18-29). Fifteen were right-handed and eight were left-handed. Participants completed questionnaires to ensure that they knew English to native proficiency, and that they were not taking

medications that could interfere with cognitive or neurological function (Appendix 4). Participants chose between 3 course credits ($n = 11$) or \$30 cash ($n = 12$) for their participation. A power analysis of our behavioral data (Uddin et al., 2018) suggested that we needed a sample size of 16 participants to achieve a power level of 0.95; therefore, for this experiment, we rounded up to a goal of at least 20 participants.

4.2.2 Stimuli

The stimuli were the same as those used in Experiment 2 (3.2.2 Stimuli), but the block structure of the experiment was different; we presented twice as many blocks by allowing sentences to repeat. This was done to have enough artifact-free trials to obtain clean ERPs. Sentences were blocked by target type, such that there were four blocks of sentences ending in sounds, and four blocks of sentences ending in words. Block types alternated across the experiment, and the type of starting block (i.e. sound or word) was counterbalanced across subjects. Within each block, match and mismatch sentences were pseudo-randomly presented such that half were matches and half were mismatches. The sentences within each block were similarly divided and randomized between general and specific. The design was balanced such that each word and sound appeared in match and mismatch conditions an equal number of times. Stimuli were experienced at 65-70 dB over insert earphones (3M E-A-RTone Gold) and were presented using Matlab 2015 (MathWorks, Inc., Natick, MA) with Psychtoolbox 3 (Brainard, 1997; Kleiner et al., 2007). The Matlab code used for stimulus randomization and presentation is available on Open Science Framework (<https://osf.io/asw48/>).

4.2.3 Testing procedure

The participants were informed about the EEG procedure, and head circumference was measured. Electrodes were applied, and participants were seated at a desk in front of a computer

monitor and keyboard for the rest of the experiment. Participants were instructed to listen to the sentences and think about whether they made sense. They were instructed to keep eye blinks and other movements confined to the silent periods between the stimuli. To encourage participants to pay attention, they were tested on recognition of the target words or sounds four times per block. Specifically, they heard a random sentence target item (either an isolated sound or word, depending on the type of stimuli in the current block) and were asked, “Have you heard this item? If yes, was it in a meaningful or nonsense context?” In this case, “meaningful” refers to congruent/match and “nonsense” refers to incongruent/mismatch. They responded via button press with two buttons marked “yes” and “no” on the keyboard. After the experiment, the position of electrodes on participants’ heads were imaged in an 11-camera geodesic dome (Geodesic Photogrammetry System, EGI, Eugene, OR) to determine the precise spatial location of all 128 electrodes (Russell, Eriksen, Poolman, Luu, & Tucker, 2005). One participant was unable to have the photos taken due to difficulties with mobility.

4.2.4 EEG Setup

Saline Hydrocel Geodesic Sensor Nets with 128 electrodes (EGI, Eugene, OR) were used for the EEG recordings. After the net was applied, impedance was minimized (to 50 k Ω or less) by repositioning electrodes, or if necessary rewetting electrode sponges. Recordings were sampled at 1000 Hz and amplified with a 128-channel high-input impedance amplifier (400 M Ω , Net Amps™, Electrical Geodesics Inc., Eugene, OR). The software used for EEG data collection was Netstation 5 (Electrical Geodesics Inc., Eugene, OR).

4.2.5 Data preprocessing

Preprocessing was done in BESA 6.0 (Scherg et al., 2002). EEG recordings were filtered with 0.1-30 Hz bandpass (Tanner, Morgan-Short, & Luck, 2015), and a 60 Hz notch filter was

applied to remove electrical noise. The recordings were then segmented based on trial type; trials were marked from 100 ms before to 900 ms after the onset of the sentence-terminal target sound or word. These trial segments were then examined for recording artifacts including eye blinks and movements; trials with eye blinks or other contaminating signals were removed, and exceptionally noisy channels were interpolated. The waveforms were baseline corrected using the 100 ms before target onset. Participants with 50% or more artifact-contaminated trials in any one condition were removed from further analysis. This procedure resulted in removal of one participant who lost over half the trials in the specific/mismatch/sounds condition.

Electrode coordinates from individuals' net placement images were used to assign individual sensor locations for each participant. For one participant who could not sit in the geodesic dome due to mobility difficulties, an average coordinate file provided by EGI was used (Electrical Geodesics Inc., Eugene, OR).

4.2.6 Analyses

4.2.6.1 Topographic Analyses

We used BESA 6.0 to generate participant-level averaged waveforms and [mismatch – match] difference waves. We also used BESA to create ascii files of time-varying voltage at every electrode; these were used for topographic analysis in RAGU (Randomization Graphical User interface, Koenig, Kottlow, Stein, & Melie-García, 2011). Averaged waveforms and difference waves for each participant are available on Open Science Framework (<https://osf.io/asw48/>).

RAGU is an unbiased method for testing for statistically significant main effects or interactions between experimental factors using a scalp topographic map randomization procedure (a detailed description of this procedure is provided in the Supplement). RAGU has the advantage of using all 128 electrodes, i.e. the entire scalp topography, rather than requiring individual

electrodes to be chosen for statistical testing. It relies on randomizations using only the collected dataset, and makes no assumptions about data distributions. We performed two analyses using 5,000 randomizations of the data in RAGU.

The first analysis was intended to address our questions about differences between understanding words and ES in preceding sentence context. In this analysis, we analyzed [mismatch – match] difference topographies (as in, for example, Frishkoff & Tucker, 2001). This is because raw voltage between responses to ES and spoken words might be different for many reasons unrelated to our manipulations in the present study, as discussed in the Introduction. Therefore, in this analysis, we examined the difference topographies for main effects of sentence constraint (specific vs. general) and target type (word vs. ES). This analysis identified time windows where there were significant main effects and interactions; it also output scalp topographies for the different conditions at each time point.

The second analysis included only factors of target type (word vs. ES) and congruency (match vs. mismatch), as it pooled together the two constraint levels. This analysis was conducted to 1) identify the time window in the vicinity of the N400 where there is a significant match vs. mismatch difference in topography, so that we could export data from this time window for N400 latency analysis, and 2) uncover potentially important differences between the responses to ES and words at other time points. Even though responses to ES versus spoken words might differ for many possible reasons unrelated to context effects as already discussed, we ran this exploratory analysis to see if any of these differences were of note. Note that in neither analysis did we pool together ES and words; sentence ending type was always a main factor being investigated as both a main effect and an interaction in our models.

Because our randomization analyses involve 5,000 randomizations of the topographical maps at each time point, there are multiple statistical comparisons across time. To avoid false positives, we implemented a threshold of 40 ms for significance windows identified in our analyses (e.g., Guthrie & Buchwald, 1991). Time windows shorter than 40 ms showing significant main effects or interactions were not considered for further analysis.

4.2.6.2 Regions of Interest (ROIs)

In order to represent most of the topography of the scalp in our analysis without arbitrarily choosing just a few electrodes, data were pooled into nine ROIs in a fashion similar to Potts & Tucker (2001), who used four adjacent electrodes in each ROI. Our ROIs and the component electrodes of each are listed in Table 4.1. The pooled ROI data were used for two purposes: 1) to represent voltage traces in figures, and 2) for statistical analysis of latency data described in the next section.

Table 4.1 Electrodes included in ROIs. Numbers correspond to electrode numbers in the EGI Hydrocel 128-electrode Geodesic Sensor Net. A spatial layout of this net is available on this study's Open Science Framework page <https://osf.io/asw48/>.

ROI	Electrodes	# electrodes
Anterior left	26, 27, 32, 33	4
Anterior midline	4, 11, 16, 19	4
Anterior right	1, 2, 122, 123	4
Center left	40, 45, 46, 50	4
Center midline	7, 55, 107, Cz	4
Center right	101, 102, 108, 109	4
Posterior left	58, 59, 64, 65	4
Posterior midline	71, 72, 75, 76	4
Posterior right	90, 91, 95, 96	4

4.2.6.3 Latency Analysis

As peak latency differences do not reliably reflect timing differences (Hansen & Hillyard, 1984; Luck, 1998), we used the time point dividing the area under the [mismatch – match]

difference curve into two equal halves to estimate latency of the N400. To make sure we were looking at the N400, we limited this analysis to the N400 time window identified in our second randomization analysis (significant main effects of match vs. mismatch; 309 – 512 ms post target onset). For each subject, and for sounds and words separately, we pooled electrodes into the nine ROIs described above. These data were entered into a repeated measures ANOVA in R using the “ez” package (Lawrence, 2013). The dependent variable was latency in milliseconds post-target-onset; within-subject factors were ROI (9 levels, Table 4.1) and condition (speech vs. non-speech).

4.3 Results

First, we assessed whether the present paradigm produces an N400. In the typical N400 time window 200-600 ms post-target-onset (e.g., Kutas & Federmeier, 2011), we found significant differences between match and mismatch conditions. Specifically, between 309 and 512 ms after target onset, we observed significant topographic ERP map dissimilarities between match and mismatch sentence endings ($p < 0.05$, Figure 4.1, Table 4.4) without considering target type (i.e., word vs. ES). The observed generalized dissimilarity between match and mismatch topographies in this time period exceeded the generalized dissimilarity obtained in at least 95% of the randomizations (Figure 4.1b). The mean observed mismatch vs. match dissimilarity in this time window was 8.13; the mean dissimilarity expected due to random chance was 4.66 (95% CI: 2.91 – 7.37). While there were other windows exhibiting significant main effects of congruency between 1) 254 and 286 ms and 2) between 612 and 639 ms, these windows did not pass our duration threshold (≥ 40 ms) for further examination (Figure 4.1c, Table 4.2).

Figure 4.1 Match vs. mismatch N400. a) Scalp topographies for match and mismatch endings at 360, 400, and 430 ms post ending onset. Blue indicates negative potential; red indicates positive potential. **b)** Time-varying generalized dissimilarity between raw match and mismatch topographies. To give a sense of the meaning of this effect size, the mean and 95% CI for the generalized dissimilarity expected due to random chance (estimated from randomizing the data) is also represented.

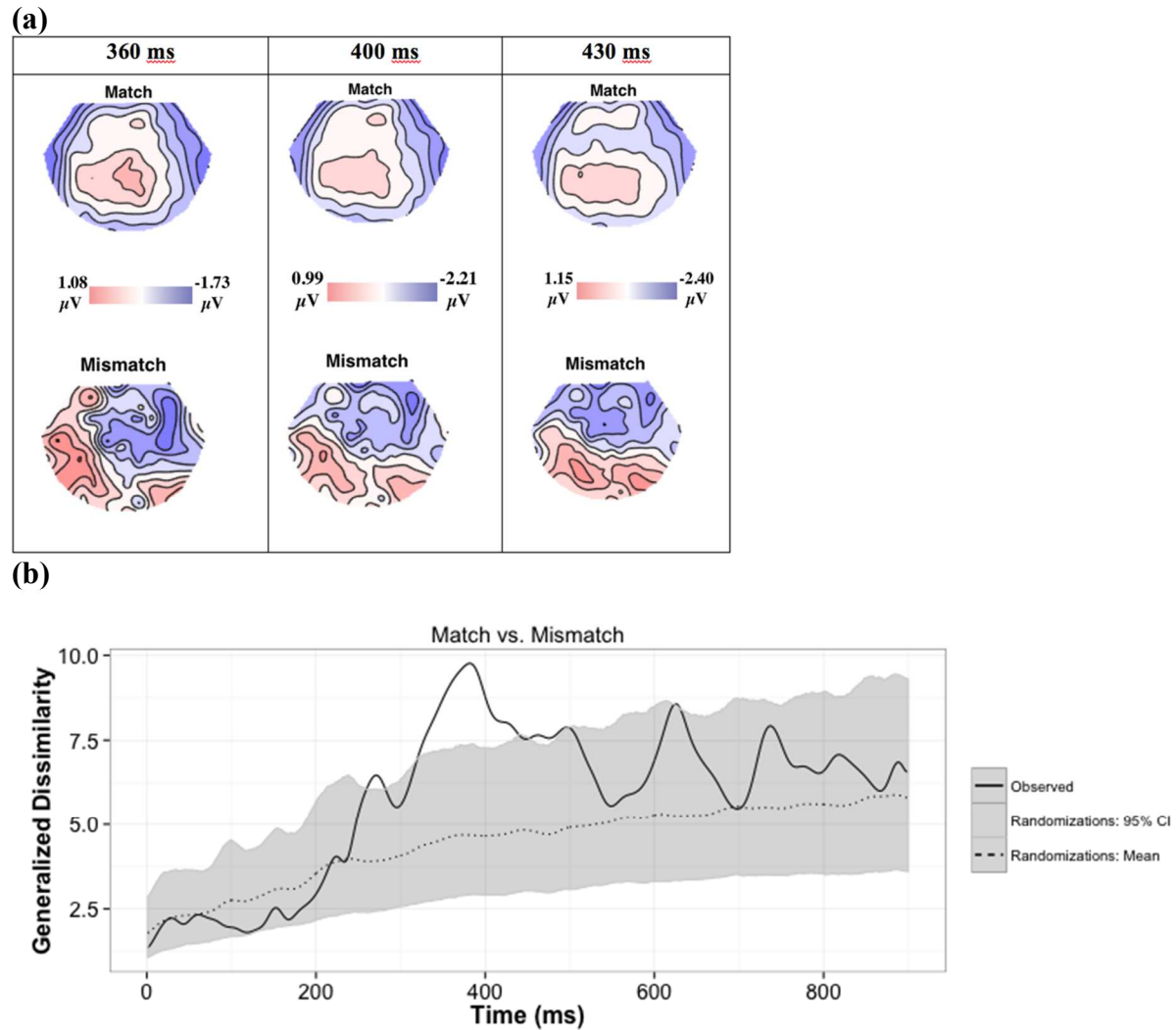


Figure 4.1 Match vs. mismatch N400, *cont.* **c)** Time-varying p-value, i.e., proportion of randomizations leading to a larger effect size than observed. We can see that the largest window showing a reliable main effect of congruency is from approximately 300 – 500 ms post target onset. **d)** Voltage traces for pooled match and mismatch sentence endings in the nine examined ROIs. Note that negative is up and time = 0 ms corresponds to ending onset.

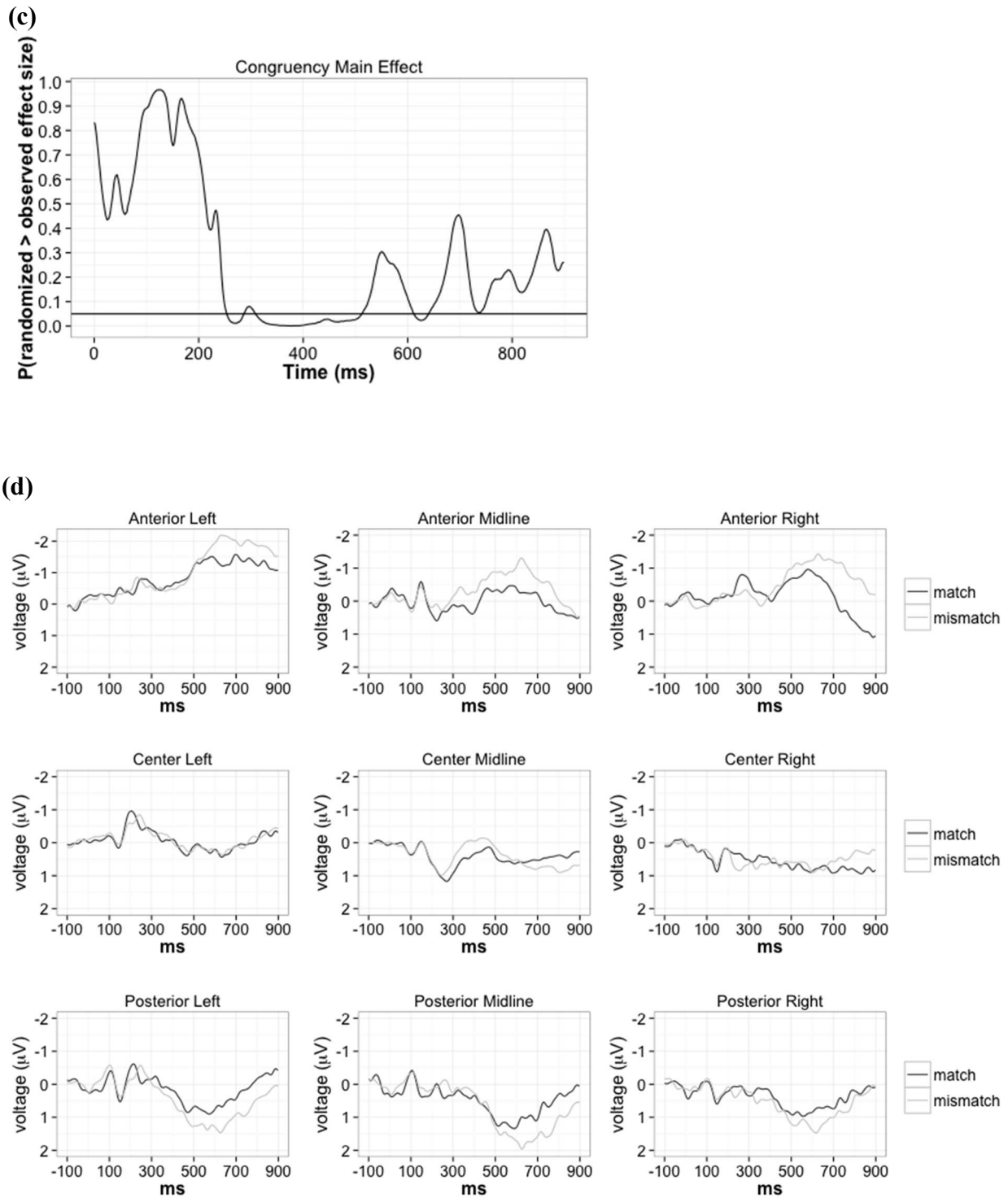


Table 4.2 Raw topography TANOVA windows. Significance windows identified by the randomization analysis (TANOVA) of raw topographies. There were no significant interactions of congruency (match/mismatch) and target type (ES/word).

Factor	Window Start	Window End	Window Length	>= 40 ms?
Congruency main	254	286	32	FALSE
Congruency main	309	512	203	TRUE
<i>Congruency main</i>	612	639	27	FALSE
Target Type main	77	118	41	TRUE
Target Type main	132	729	597	TRUE
Target Type main	883	899	16	FALSE
Congruency * Targ Type	NULL			

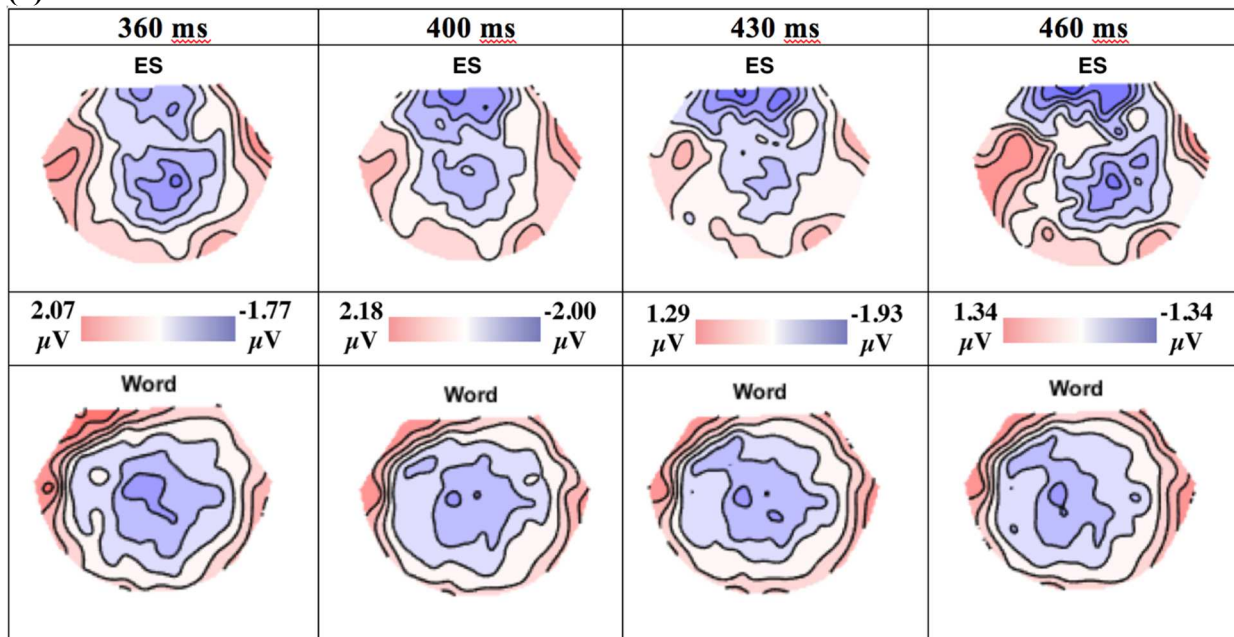
The scalp topographies in the 309 - 512 ms time window indicate a stronger frontocentral, slightly right-lateralized negativity when the target is not congruent with the preceding sentence (Figure 4.1a, 4.1d). By 430 ms, responses to match endings show a similar, albeit weaker, frontocentral negativity (Figure 4.1a, 4.1d). This topography is similar to previous N400 studies with language presented in the auditory modality (Kutas & Federmeier, 2011; Kutas & Van Petten, 1994). It is important to note that while there was a significant main effect of congruency (in which a larger N400 occurs for both spoken word and ES mismatches as opposed to matches), there were no interactions between target type and congruency in this analysis ($p > 0.25$, Table 4.4, Table 4.2). Moreover, our analysis of [mismatch – match] differences (which will be discussed in more detail later) showed no effect of ending type ($p > 0.25$, Table 4.4, Table 4.3). Taken together, these results indicate that there was no evidence for substantially different congruency-related N400 activity for ES and words.

Once we confirmed that our paradigm elicited N400 activity related to incongruency, we could ask our main questions about differences between speech and ES N400s. Most importantly, we wanted to assess if the N400's sensitivity to incongruency was similar for ES and words in the

context of a fluent spoken sentence. Our [mismatch – match] difference topography analysis revealed that both ES and words had greater frontocentral negativities in mismatch conditions (Figure 4.2a). In fact, there was no statistically significant difference between the [mismatch – match] topographies of words and ES in the vicinity of the N400 (Figure 4.2b, c). The mean observed dissimilarity between words and ES [mismatch – match] topographies in the N400 time window identified above (309 - 512 ms) was 9.27. The mean observed dissimilarity that could be expected due to chance was 8.24 (95% CI: 5.44 - 12.58). Thus, the congruency sensitivity of the N400 was not statistically distinguishable between words and ES ($p>0.25$; Table 4.4). There were three very short windows after 600 ms where a significant main effect of target type was found (Figure 4.2c, Table 4.3), however as the longest of these was 16 ms, none passed our 40 ms threshold for further consideration.

Figure 4.2 ES and word N400 difference topographies. **a)** Scalp topographies for [mismatch – match] difference topographies for ES and words at 360, 400, 430, and 460 ms post target onset. Blue indicates negative potential; red indicates positive potential. **b)** Time-varying generalized dissimilarity between ES and word [mismatch – match] difference topographies. To give a sense of the meaning of this effect size, the mean and 95% CI for the generalized dissimilarity expected due to random chance (estimated from randomizing the data) is also represented.

(a)



(b)

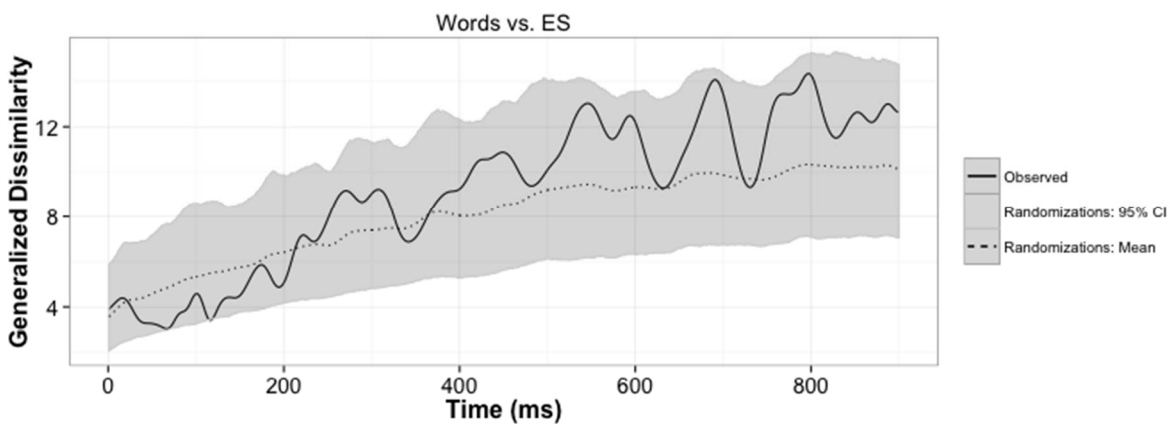


Figure 4.2 ES and word N400 difference topographies, *cont.* **c)** Time-varying p-value, i.e., proportion of randomizations leading to a larger effect size than observed. We can see that there are no points in the vicinity of the N400 where the difference between word and ES topographies is statistically significant. **d)** Voltage traces for ES and word [mismatch – match] difference waves in the nine examined ROIs. Note that negative is up and time = 0 ms corresponds to ending onset.

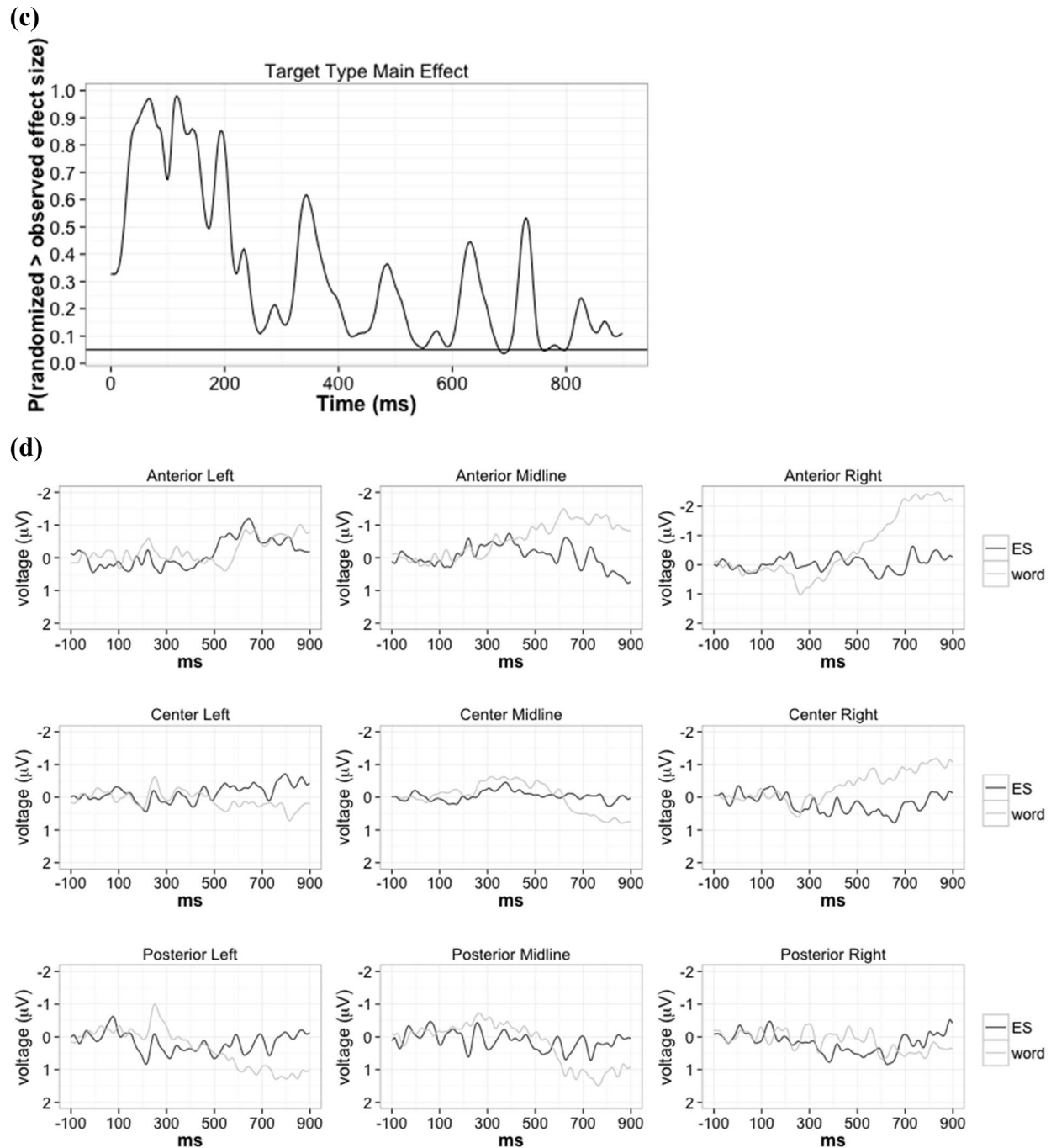


Table 4.3 Difference topography TANOVA windows. Significance windows identified by the randomization analysis (TANOVA) of [mismatch – match] difference topographies. There were no significant interactions of constraint (general/specific) and target type (ES/word).

Factor	Window Start	Window End	Window Length	>= 40 ms?
Constraint main	259	287	28	FALSE
Constraint main	359	421	62	TRUE
Constraint main	445	476	31	FALSE
Constraint main	505	524	19	FALSE
Target Type main	683	699	16	FALSE
Target Type main	759	766	7	FALSE
Target Type main	795	799	4	FALSE
Constraint * Targ Type	NULL			

Table 4.4 Summary of statistics for randomization analyses of topographical maps.

Window	analysis	factor	observed generalized dissim. (GD)	expected GD under null hyp.	95% CI of GD under null hyp.	p
309 - 512 ms	raw voltage, congruency x target type	congruency (match vs. mismatch)	8.13	4.66	2.91 - 7.37	*0.015
309 - 512 ms	raw voltage, congruency x target type	congruency * target type interaction	9.09	9.33	5.82 - 14.77	0.49
309 - 512 ms	mismatch - match difference, constraint x target type	target type (words vs. ES)	9.27	8.24	5.44 - 12.58	0.28
359 - 421 ms	mismatch - match difference, constraint x target type	constraint (general vs. specific)	12.60	8.09	5.30 - 12.53	*0.025
359 - 421 ms	mismatch - match difference, constraint x target type	constraint * target type interaction	12.92	16.22	10.65 - 25.05	0.27
612 - 639 ms	raw voltage, congruency x target type	congruency (match vs. mismatch)	8.28	5.26	3.33 - 8.51	*0.032

Table 4.5 Summary of statistics for repeated measures ANOVA examining N400 latency.

analysis	factor	latency difference	F	Cohen's d	95% CI	p
latency repeated measures ANOVA	target type	5.26 ms (ES earlier)	2.51	0.56	-1.25 - 11.78	0.13
latency repeated measures ANOVA	ROI	-	0.92	-	-	>0.25
latency repeated measures ANOVA	target type * ROI	-	0.99	-	-	>0.25

If understanding ES in a spoken sentence frame is more difficult than doing so for spoken words, the ES N400 could be delayed. We found no evidence that the N400 to ES was delayed; a 2 x 9 (target type [word, ES] x ROI [AL, AR, AM, CL, CR, CM, PL, PR, PM]) repeated measures ANOVA of [mismatch – match] difference wave latencies showed that there was no main effect of target type on latency [$F(1,21)=2.51, p=0.13$, diff=5.26 ms, $d=0.56$, 95% CI (-1.25 - 11.78 ms)]. The mean latency for ES was 406.56 ms [401.51 - 411.61] and for words was 411.82 ms [407.62 - 416.03]. Therefore, even though the difference in latency trends towards significance, the trend is for N400s to ES to be earlier, not later, than N400s to words. There was also no evidence of a main effect of ROI [$F(8,168)=0.92, p>0.25$] or an interaction of target type and ROI [$F(8,168)=0.99, p>0.25$], indicating that N400 latencies were not statistically distinguishable between different regions of the scalp (Table 4.5).

Another important question we asked was whether sentence constraint affected responses to ES and words similarly. In order to address this question, we first asked if previous literature was replicated by finding constraint main effects on the N400. As outlined in the Introduction, based on previous literature we expect to find larger N400s for low constraint sentences when the ending is congruent, and roughly equal N400s for low and high constraint sentences when the

ending is incongruent. By this logic, the [mismatch – match] difference wave for low constraint sentences should be smaller, as the larger match N400 would cancel out the large mismatch N400. For high constraint sentences, we should see a stronger N400 negativity in the [mismatch – match] difference wave, as the weak match N400 would do little to cancel out the mismatch N400. Our randomization analysis of the difference topographies showed a significant main effect of constraint between 359 and 421 ms after target onset ($p < 0.05$, Figure 4.3b-c, Table 4.4, Table 4.3). In this window, the mean observed dissimilarity between general and specific [mismatch – match] topographies was 12.60. The mean observed dissimilarity that could be expected due to chance was 8.09 (95% CI: 5.30 - 12.53).

Figure 4.3 General vs. specific difference topographies. a) Scalp topographies for [mismatch – match] difference topographies for endings following low constraint (general) vs. high constraint (specific) sentences at 275, 400, and 460 ms post target onset. Blue indicates negative potential; red indicates positive potential. We can appreciate that general difference topographies are characterized by negativities that are shifted anteriorly relative to specific difference topographies.

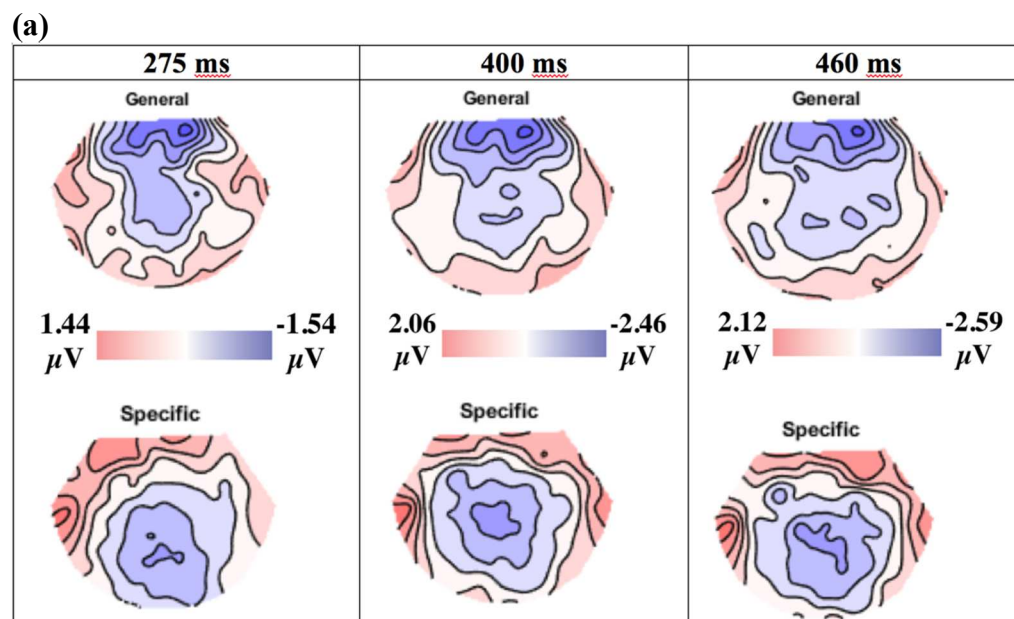
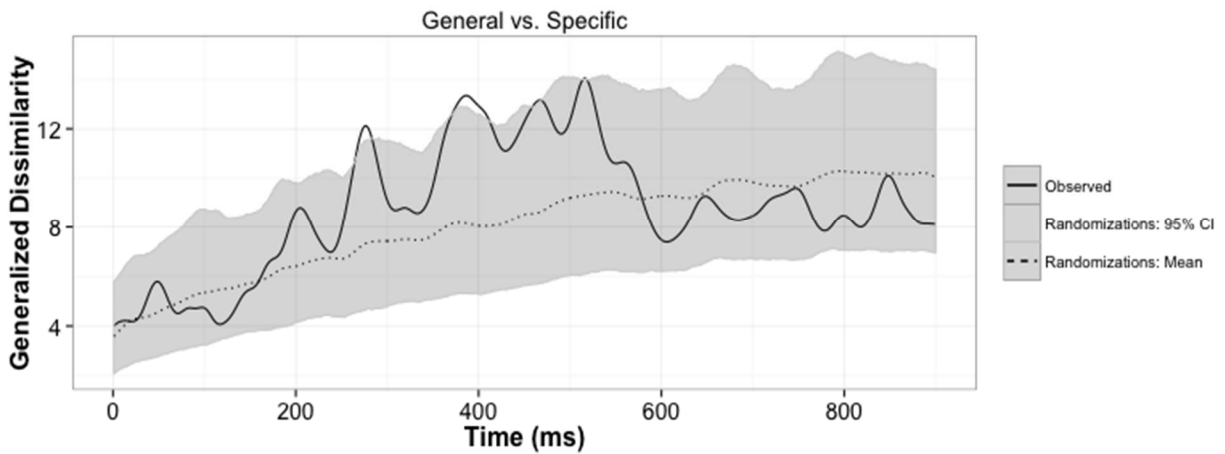


Figure 4.3 General vs. specific difference topographies, *cont.* b) Time-varying generalized dissimilarity between general and specific [mismatch – match] difference topographies. To give a sense of the meaning of this effect size, the mean and 95% CI for the generalized dissimilarity expected due to random chance (estimated from randomizing the data) is also represented. **c)** Time-varying p-value, i.e., proportion of randomizations leading to a larger effect size than observed. We can see that there is a main effect of constraint centered around 400 ms.

(b)



(c)

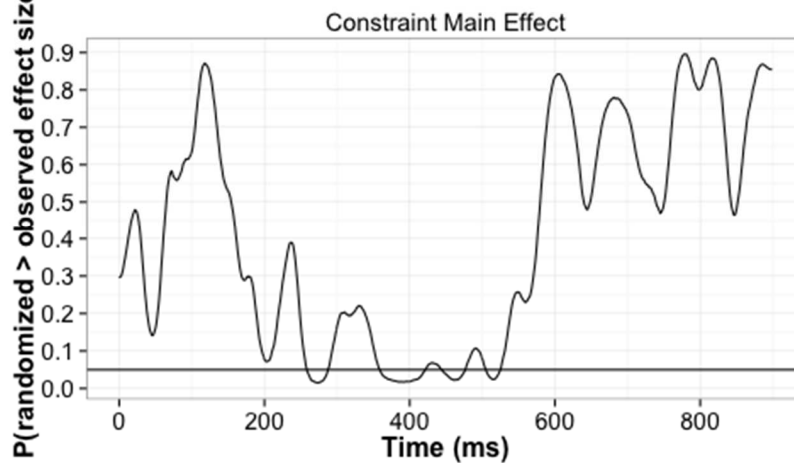
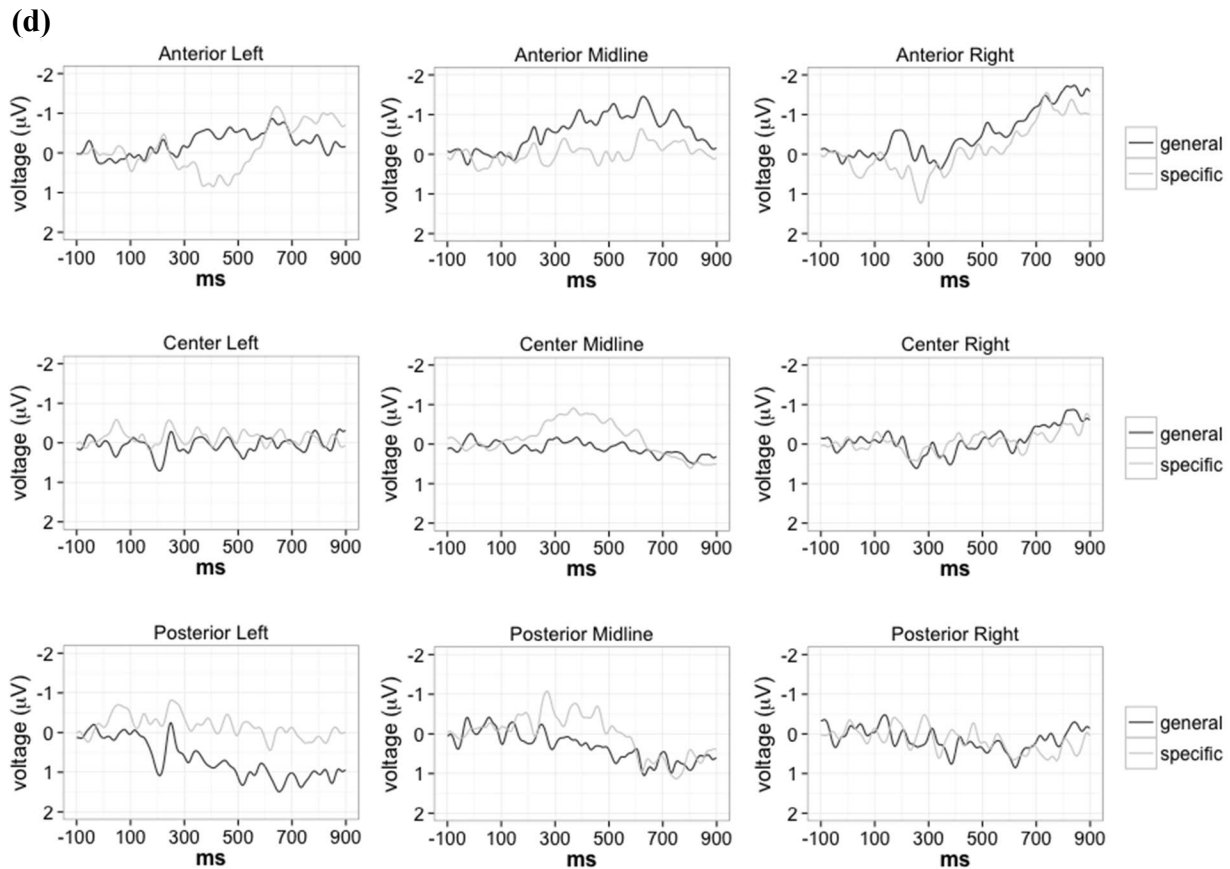


Figure 4.3 General vs. specific difference topographies, *cont.* d) Voltage traces for general and specific [mismatch – match] difference waves in the nine examined ROIs. Note that negative is up and time = 0 ms corresponds to ending onset.



There were also main effects of constraint beginning at 259 and again at 445 ms, which did not reach our 40 ms threshold (Table 4.3); however, these windows exhibited the same topographic difference between general and specific as our longest window from 359 – 421 ms. Namely, [mismatch – match] topographies to endings after low constraint sentences exhibit a negativity that is shifted frontally relative to high constraint (Figure 4.3a). If we focus on central midline and posterior regions, we can see that the N400 [mismatch – match] wave appears weaker in general conditions, as previous literature would predict (Figure 4.3d). It is possible that previous studies have reported a weaker N400 in low constraint conditions because they focused on central regions rather than the entire scalp topography.

Once we demonstrated that constraint affects the N400, we asked whether it had similar effects on neural responses to ES and words in sentence context. If constraint affects ES and words differently, we would expect to see an interaction between target type and constraint in our randomization analysis. There was no evidence of such an interaction (Figure 4.4b-c); the mean generalized dissimilarity between 359 and 421 ms associated with the interaction was 12.92, whereas the dissimilarity expected due to chance in this time period was 16.22 ($p > 0.25$, 95% CI: 10.65 – 25.06, Table 4.4). The topographies separated out for the four conditions (general/ES, specific/ES, general/word, and specific/word) all show the same pattern of a frontal shift for general conditions (Figure 4.4a). Thus, constraint does not appear to differentially affect ES and words in the context of a spoken sentence.

Figure 4.4 N400 for general vs. specific, words vs. ES. a) Scalp topographies for [mismatch – match] difference topographies for ES or word endings following low constraint (general) vs. high constraint (specific) sentences at 400 ms post target onset. Blue indicates negative potential; red indicates positive potential.

(a)

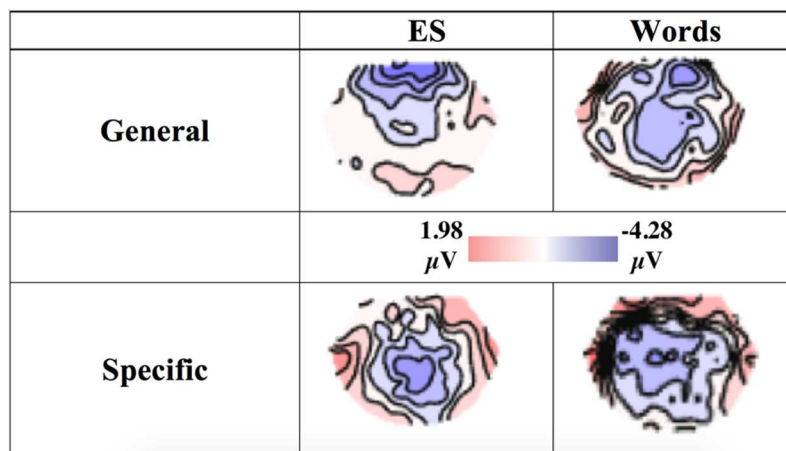
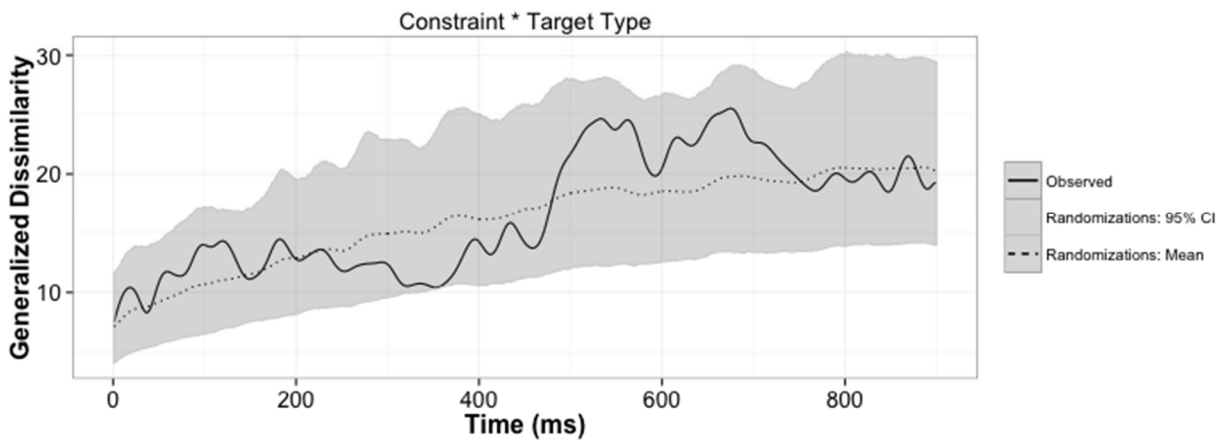
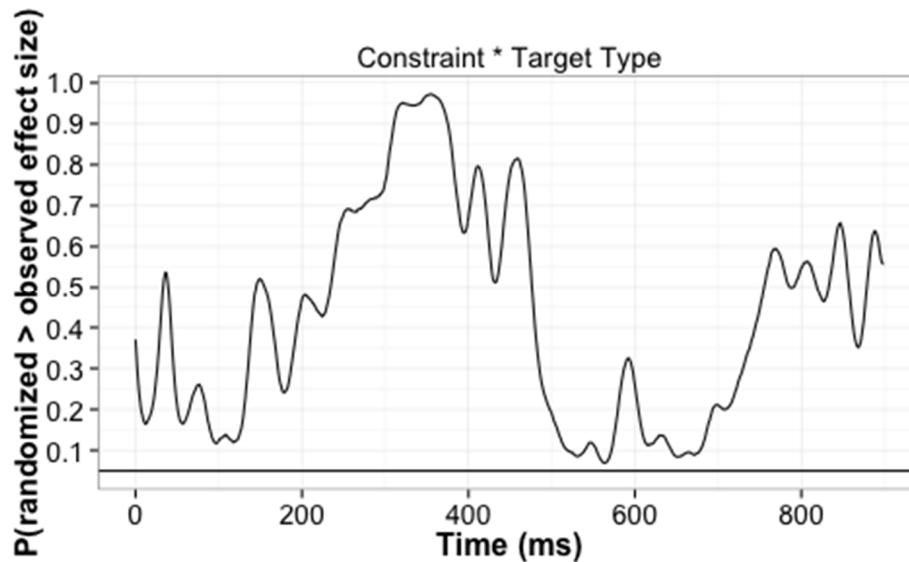


Figure 4.4 N400 for general vs. specific x words vs. ES, *cont.* **b)** Time-varying generalized dissimilarity associated with the interaction between constraint (general/specific) and target type (ES/word). To give a sense of the meaning of this effect size, the mean and 95% CI for the generalized dissimilarity expected due to random chance (estimated from randomizing the data) is also represented. **c)** Time-varying p-value, i.e., proportion of randomizations leading to a larger interaction between constraint and target type than observed. Note that the p value always remains above the 0.05 threshold in this case, particularly in the vicinity of the constraint main effect, which is from 359 – 421 ms.

(b)



(c)



As noted in the Introduction, there are reasons to expect a difference in the neural processing of speech and non-speech sounds due to acoustic differences, frequency of experience, and differences in cortical activation patterns. The N1-P2 is a pair of event-related potentials (ERPs) that marks acoustic stimulus change (e.g., Hillyard & Picton, 1978); when it occurs after a change in an ongoing sound, it is called an acoustic change complex (ACC) (Kim, 2015). Our second randomization analysis revealed an ACC in response to ES following a spoken sentence, as shown clearly both from scalp topographical maps, and from voltage traces in ROIs (Figure 4.5a-b). The presence of this ACC was statistically supported by significant main effects of Target Type on topographies in a long window (132 - 729 ms) encompassing the ACC time frame in our second randomization analysis (Table 4.2). Though there were differences between raw voltage for ES and words in places other than the ACC, these were not a focus of the current study because (as already discussed in the Introduction) they could be due to many factors beyond the current study manipulations. Therefore, they will not be discussed further.

Figure 4.5 ACC topographies for ES. a) Scalp topographies for words vs. ES at 150, 200, and 250 ms post target onset, i.e. during the ACC. Blue indicates negative potential; red indicates positive potential.

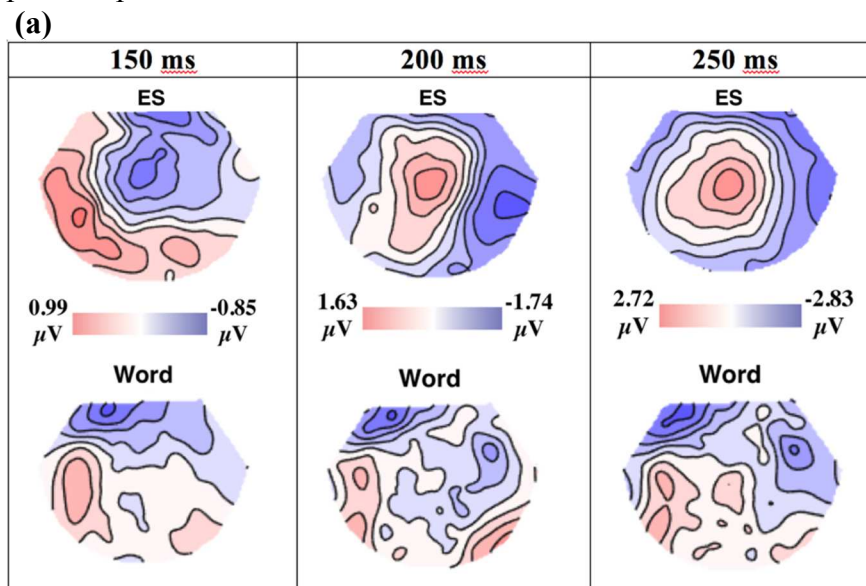
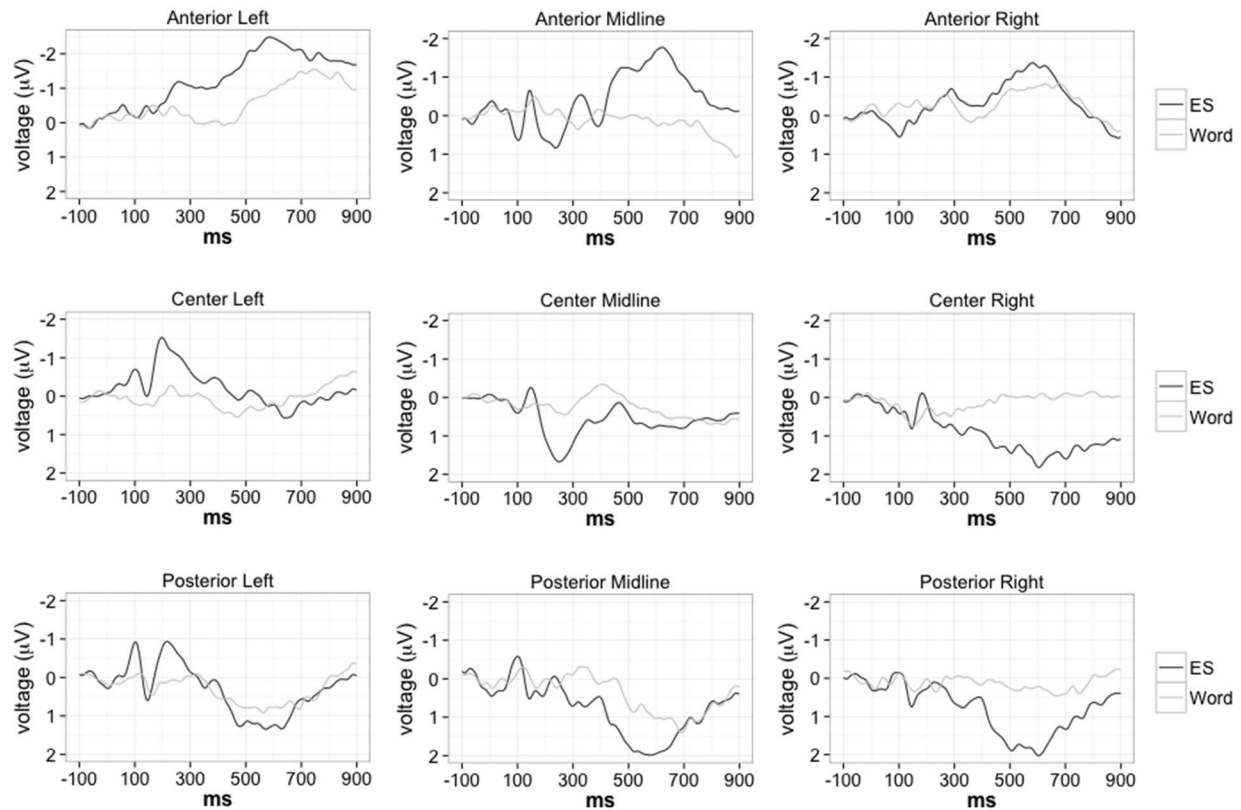


Figure 4.5 ACC topographies for ES, *cont.* b) Voltage traces for ES and words in the nine examined ROIs. Note that negative is up and time = 0 ms corresponds to ending onset.

(b)



Finally, though there was no statistically significant interaction between congruency and target type, there is an apparent morphological difference between the shape of the N400 for ES compared to words (Figure 4.6a-b, particularly anterior and center midline ROIs). Namely, it appears that the N400 to ES consists of two peaks instead of one. When we break this down by condition (Figure 4.6a), it can be seen that the first peak appears to be more sensitive to congruency than the second one. This effect does not reach significance, but because of similarities with previous literature we will address this morphological difference in the discussion.

Figure 4.6 ES and words match vs. mismatch voltage traces. a) Voltage traces for ES, showing match and mismatch separately, in the nine examined ROIs. Note that negative is up and time = 0 ms corresponds to ending onset.

(a)

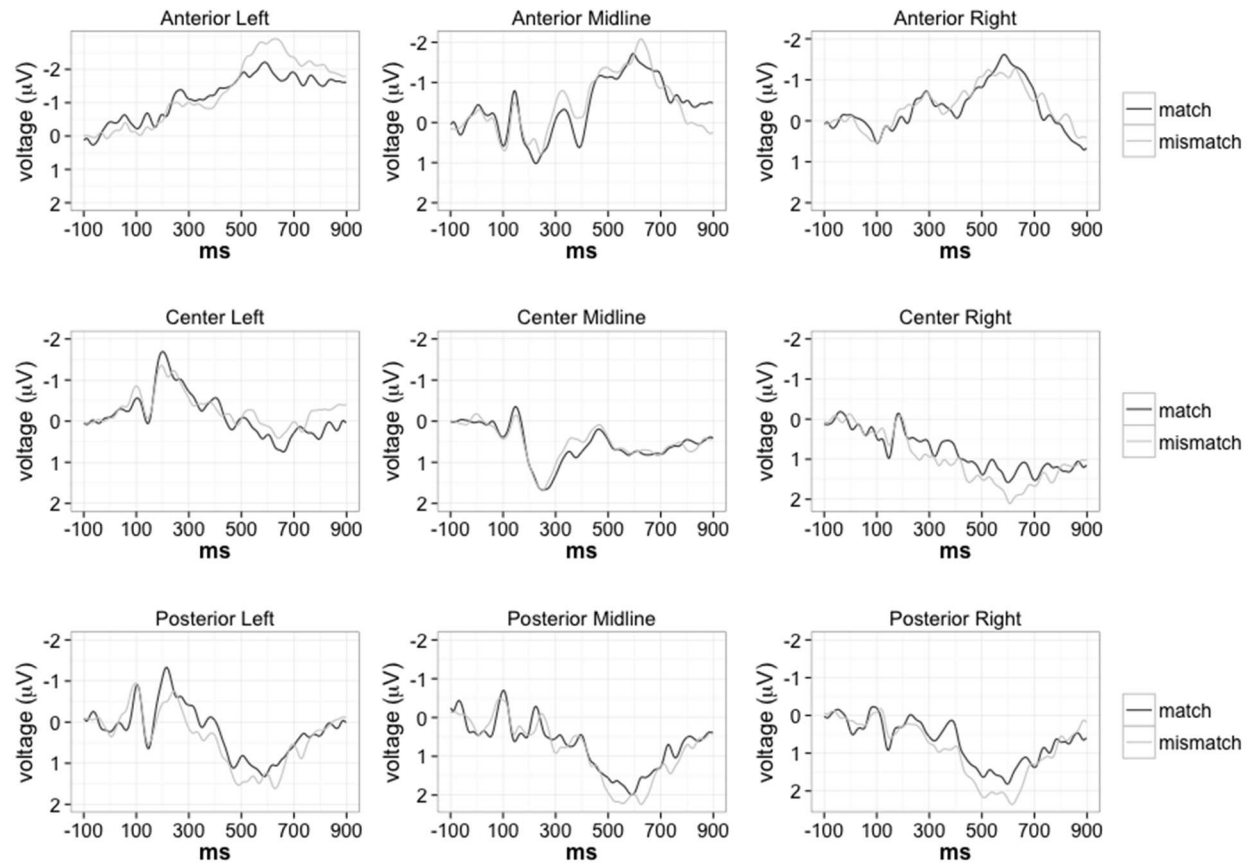
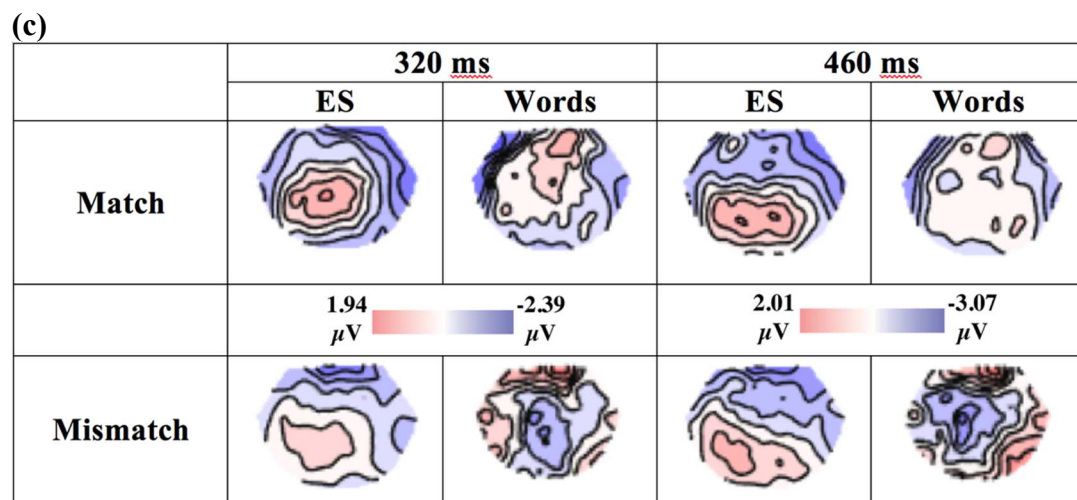
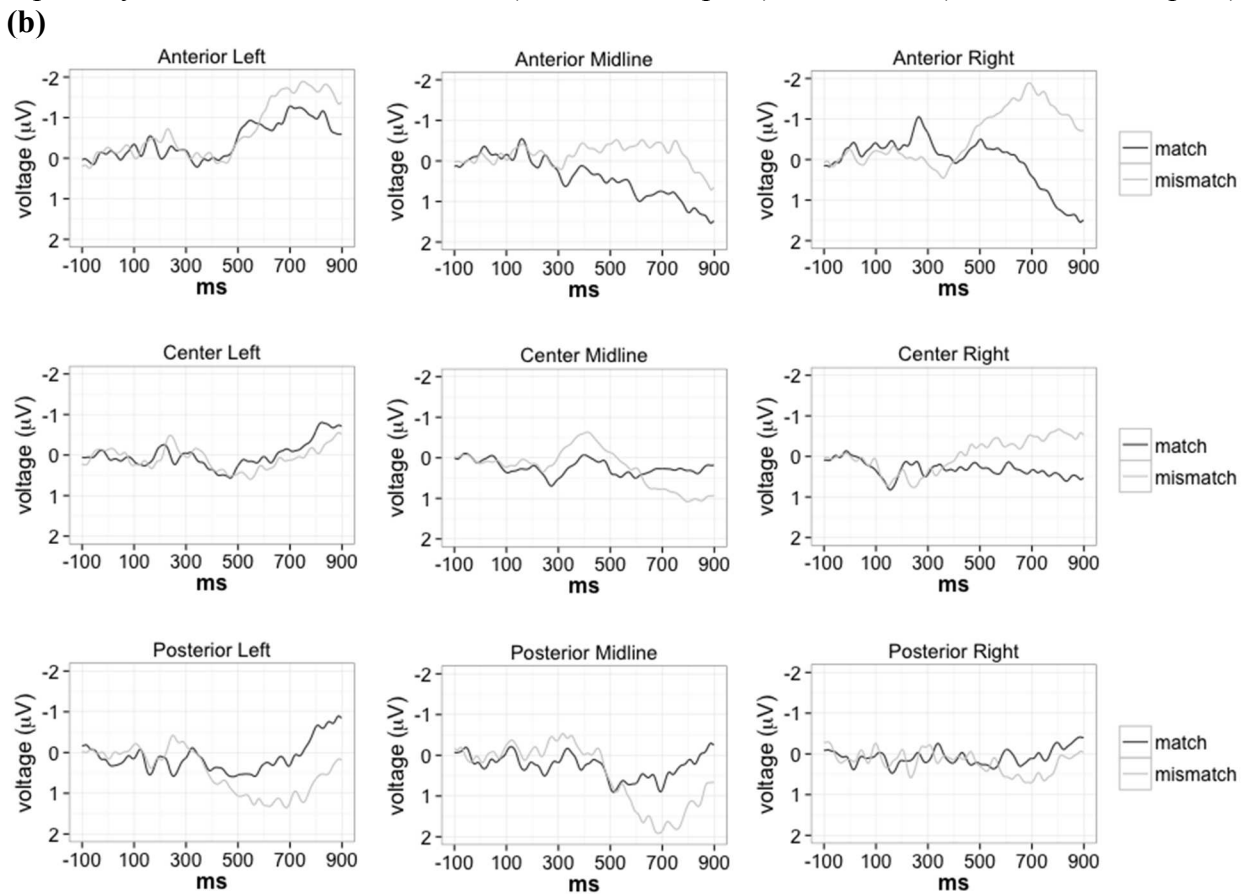


Figure 4.6 ES and words match vs. mismatch voltage traces, *cont.* b) Voltage traces for words, showing match and mismatch separately, in the nine examined ROIs. Note that negative is up and time = 0 ms corresponds to ending onset. **c)** Match and mismatch topographies shown separately for ES and words at 320 ms (first ES N400 peak) and 460 ms (second ES N400 peak).



Finally, though the focus of this experiment was the N400, it is worth noting that a brief main effect of congruency was found in the P600 time window when raw topographies (not difference topographies) were examined. This effect was short-lived (27 ms, Table 4.2, Figure 4.1 b-c), weak (Table 4.4, Figure 4.1b), and involved stronger posterior positivities for incongruent sentence endings (Figure 4.1d, posterior and center midline panels; Figure 4.7). This effect did not interact with target type.

Figure 4.7 P600 topographies for match vs. mismatch. Raw match and mismatch topographies at 625 ms post-ending onset.



4.4 Discussion

The main question in the present study is whether environmental sounds and spoken words produce similar patterns of brain electrical responses when they are being understood in the same fluently spoken sentence context. Regardless of target type (word/ES), a significantly more negative frontocentral N400 occurred for mismatch conditions 309 - 512 ms after target onset (Figure 4.1). This finding replicates previous research showing stronger N400s to incongruent stimuli. The start of the significance window at 309 ms is characteristic of auditory N400s to fluent speech, which happen earlier than visual N400s (Kutas & Federmeier, 2011).

One crucial question was whether sentences ending in ES produce an N400 effect; one with properties similar to those for speech. The answer to this question is clear: [mismatch – match] difference topographies were not statistically different between ES and words. In both cases, difference topographies showed central negativities characteristic of the expected stronger N400 to mismatch stimuli (Figure 4.2). This indicates that congruency with the preceding context

affects the N400 to ES and words in the same way: for both, N400 responses are more negative in response to mismatches. To our knowledge, this is the first demonstration that ES can give rise to an N400 effect in a fluent speech context, and suggests some level of processing similarity with speech.

Incidentally, other studies have sometimes demonstrated lateral asymmetry effects in the difference topographies, such that the N400 congruency effect is more right-lateralized for ES (e.g., van Petten & Rheinfelder, 1995). We did not find such an effect, which would have shown up as a target type effect in the difference topography randomization analysis. Previous work shows that N400 lateralization can differ based on handedness (Fagard, Sirri, & Rămă, 2014; Kutas & Hillyard, 1980b). Therefore, a possible reason we failed to find lateral asymmetries between ES and words is the high proportion of left-handed participants in our study (over a third, whereas other studies often use all right-handed participants).

A further question we asked was whether environmental sounds produce N400s regardless of congruency with the sentence frame, simply by virtue of dissimilarity from speech. Our difference topography analysis also answers this question. If N400s to ES were always large, regardless of congruency with context, the [mismatch – match] difference for ES would approach zero, and a main effect of target type would be observed in our difference topography randomization analysis. Clearly this is not the case, as there is no main effect of target type on the [mismatch – match] differences. A second possibility was that ES are more difficult to understand than words in a sentence context, leading to a delayed N400. This hypothesis was also rejected; N400s to ES occurred 5.26 ms earlier than N400s to words. This agrees with previous findings that N400s are in fact slightly earlier in response to environmental sounds than to words (Cummings et al., 2006; Guido Orgs et al., 2006), although it suggests that perhaps when the

participant is focused on understanding the meaning of a full sentence, such differences are somewhat mitigated, as our 5.26 ms difference was quite small, and not statistically significant.

Finally, we asked whether the constraint level of the sentence would affect the N400 activity of ES and words differently. We did find main effects of constraint on scalp topography of [mismatch – match] differences. In particular, endings after general (low constraint) sentences elicited more frontally biased negativities, while endings after specific (high constraint) sentences elicited more central negativities. Importantly for our question, these main effects of constraint did not interact with target type; therefore, constraint appears to affect the N400 responses to ES and spoken words in a similar way. This finding goes against the prediction that constraint information might be too fine-grained or nuanced to affect ES to the same degree as words (cf. Hendrickson et al., 2015), although because we did not systematically vary near vs. far violations, the capability of this paradigm to thoroughly test this idea is limited. However, it is true that even when ES are being understood in a fluently spoken sentence—an artificial and unusual task—neural responses are affected by constraint the same way as spoken words. This suggests a surprising degree of seamless integration between these different stimulus types. Future experiments might more systematically vary the within-category nature of the semantic violations in order to test constraint effects on ES and spoken words in a more fine-grained way.

Interestingly, as outlined in the results section, previous literature predicts a weaker [mismatch – match] difference negativity for low constraint sentences, and a stronger negativity for high constraint ones. We found that this appears to be the case if we look at central and posterior regions of the scalp (Figure 4.3d). However, looking at the entire scalp topography allows us to see that this appears to be a consequence of the negativity shifting frontally for low constraint conditions (Figure 4.3a). Much of the previous literature on the effects of constraint on the N400

uses fewer electrodes and/or focuses on central scalp locations (e.g., DeLong, Urbach, & Kutas, 2005; Federmeier et al., 2007). Therefore, it is possible that future work using higher density electrode arrays and topographical analyses could uncover interesting patterns involving cloze probability and/or constraint effects on the entire scalp topography.

We also noticed an apparent morphological difference between the N400 to ES and spoken words at central and frontal midline sites. In these regions, the N400 to ES appeared to consist of two peaks instead of one. Alternatively, it can be thought of as having a positive deflection in the middle. Though this morphological difference did not lead to any statistically significant effects, we found it noteworthy because such two-peaked N400s in response to ES have been reported before (Cummings et al., 2006; Hendrickson et al., 2015; Orgs et al., 2006). Hendrickson et al. theorize that this deflection could be P3b activity. The P3b is a centroparietal positivity with a latency of 300-450 ms, and is often observed to be larger in response to stimuli that are relatively more rare than others (Comerchero & Polich, 1999). Though it is typically associated with active participation in a task, it can be elicited by passive listening (Bennington & Polich, 1999). In some sense, ES are relatively more rare than words in our study due to the fact that the spoken sentences preceding every target consist entirely of words. However, because participants were instructed to pay attention to whether the sentence made sense or not, the ending targets were the closest to task-relevant in our passive listening paradigm. Therefore, in a task-relevant sense, spoken words and ES occurred in equal proportions, making the P3b explanation unlikely. A more likely explanation is that favored by Orgs, Lange, Dombrowski, & Heil (2007) who found a remarkably similar double-peaked N400 to ES at central and frontal sites. Similarly to our voltage traces, the first of the two peaks appeared to be more sensitive to congruency than the second. Orgs et al. explained this as two separate N400 subprocesses. Though the two peaks in our study might be a

length effect stemming from the longer ES than word stimuli (0.838 vs. 0.502 s, $p=0.008$), this seems unlikely, as two-peaked N400s to ES have been found in cases where ES and spoken words were similar lengths (Hendrickson et al., 2015) and in cases where ES were all trimmed to 300 ms (Orgs et al., 2007). Moreover, word length has seldom been found to affect N400 latency, and has never been found to affect the number of peaks (Hauk & Pulvermüller, 2004). Further research is necessary to replicate and characterize this two-peaked N400, to assess whether it is characteristic of N400s to ES, and to uncover the functional significance of the two peaks.

Finally, we will discuss the implications of the brief, weak P600 effect observed for both ES and words. In human EEG, the P600 is a late parietal positivity, peaking approximately 600 ms after stimulus onset, but often stretching over quite a long period of time. For years, the P600 was described as a syntactic counterpart to the well-known presumably semantic N400: just as the N400 has been said to reflect semantic processing effort, the P600 was thought to indicate syntactic processing effort (Bornkessel-Schlesewsky & Schlesewsky, 2008; Gouvea, Phillips, Kazanina, & Poeppel, 2010; Kaan, Harris, Gibson, & Holcomb, 2000; Osterhout & Holcomb, 1992). Based on recent evidence, however, the significance of the P600 is likely considerably more complicated.

One observation that argues against a syntactic-only interpretation of the P600 is that certain types of semantic violations tend to elicit P600s rather than N400s (see Kuperberg, 2007 for a review). Such violations include semantically related words that are arranged in a way that does not make sense, sometimes called “semantic reversal anomalies (as opposed to the semantically unrelated words that typically elicit N400s) (Bornkessel-Schlesewsky & Schlesewsky, 2008). As a concrete example of this semantic reversal P600 effect, Kim & Osterhout (2005) found that unexpected, semantically improbable verbs in sentences (e.g., “The meal was devouring...”) elicited P600s rather than N400s. A narrow syntactic-only interpretation

of the P600 would also suggest that one should not find P600s in response to stimuli lacking linguistic syntax, such as musical sequences. However, anomalous musical chords in a sequence can indeed elicit a P600, calling the syntactic nature of the P600 into question, and for some scholars, raising interesting questions about whether “syntax” should be defined more broadly (Patel, Gibson, Ratner, Besson, & Holcomb, 1998).

Though the significance of the P600 appears to extend beyond syntax to semantics, there is substantial evidence arguing that the P600 is not a simple marker of semantic or conceptual incongruity. For example, no P600 is found in response to isolated words or environmental sounds when presented after a priming picture (Hendrickson et al., 2015). In this sense, the P600 is unlike the N400, which consistently appears in response to probes after primes, even if they are presented in different modalities (Hendrickson et al., 2015; Guido Orgs et al., 2006). Therefore, taken together, the literature suggests that while the N400 might result from processes of conceptual match or mismatch, the P600 is found only when the stimulus in question is placed in the context of a more complex, multi-part frame such as a sentence or a musical sequence. To complicate the picture, syntactically confusing “garden-path” sentences, while technically and semantically correct, may also elicit P600s (Osterhout & Holcomb, 1992), as can ironic rather than literal statements (Regel, Meyer, & Gunter, 2014), or instances of subject-verb disagreement (Coulson, King, & Kutas, 1998; Kaan et al., 2000). Because of results like these, it has been suggested that the P600 does not reflect repair processes, but rather reflects integration effort (Kaan et al., 2000). If one assumes that integrating pairs of isolated stimuli requires less effort than integrating an entire sentence as it unfolds, this interpretation would be consistent with findings that no P600 is elicited by picture primes followed by word probes (Hendrickson et al., 2015).

In light of this information about the P600, we can see that our results agree with previous literature, which predicts that P600s will be stronger to incongruent rather than congruent sentence endings (Figure 4.7). Moreover, the P600 does not interact with target type, supporting the interpretation that this effect is occurring for both ES and words. If the “integration difficulty” account of the P600 is correct, this means that ES and spoken words suffer from a similar increase in integration difficulty when they are incongruent with the preceding sentence frame. If ES in sentence context were always characterized by higher integration difficulty regardless of their meaning, we would expect to see a meaning-by-target type interaction in which there is always a strong P600 for ES, but a marked congruency-related difference in P600 for spoken words. Instead, we find that the P600 is stronger for incongruent stimuli whether they are ES or words. This supports the N400 results which suggest similar mechanisms for understanding ES and words in the context of a sentence frame.

Another reason this result is interesting is that it sheds light on the debate about the nature of the P600 itself. If the P600 is mainly a syntactic phenomenon, the discovery of a main effect of congruency on the the P600 for both ES and words would be surprising, as ES do not have syntactic properties in any obvious way. Given our results, it is more likely that the P600 reflects integration difficulty or a repair process. However, if the P600 does indeed reflect syntactic processes, these results speak to the “what is syntax” debate by forcing us to conclude that environmental sounds can indeed have syntactic properties if they are placed in a linguistic context.

4.4.1 Conclusions

To our knowledge, this is the first demonstration of an N400 in response to meaningful non-speech in the context of fluent speech. This study suggests that even when embedded in fluent speech, environmental sounds can benefit from interpretation and context mechanisms typically

used for understanding spoken language, along a similar timescale as spoken words. Our results suggest that neural mechanisms for integrating the meanings of words with context are flexible, and can adapt to accommodate environmental sounds, or at least such mechanisms are sufficiently general to accommodate processing of non-linguistic but meaningful acoustic patterns. Not only does congruency with context affect ES and words similarly in the context of a fluent spoken sentence—sentence constraint does as well. This work provides crucial evidence for the flexibility and adaptability of mechanisms, like linguistic ones, that at first glance appear to be quite specialized.

Chapter 5: Are differences in the ERP's observed for speech and non-speech in speech context due purely to acoustic differences?

The purpose of this study is to clarify to what extent the differences between environmental sounds and word ERPs found in the original ERP study are due to acoustic differences indicating source change. While we did find context-sensitive N400s in response to both sentence-terminal environmental sounds and spoken words, there were differences in ERPs to the two types of stimuli. It is possible these differences are due to obligatory processing differences between environmental sounds and spoken words. This account would support an interpretation that language systems are relatively specialized. However, it is also possible that these differences are simply due to the fact that the source of the sound has changed. This interpretation would suggest that the differences in ERPs are not due to the non-linguistic nature of the sounds, but merely to the fact that there is a change in source. In order to distinguish between these possibilities, we replicated the above experiment, except that instead of hearing an environmental sound at the end of half the sentences, participants heard a woman (instead of a man) saying the last word of the sentence. ERPs to a changing talker that resemble the ERPs to environmental sounds embedded in sentences would suggest that the differences found in the previous experiment are due simply to a change in source, rather than to a difference between linguistic and non-linguistic meaningful stimuli.

5.1 Conceptual Introduction

Though the primary purpose of this experiment was to better contextualize differences between ERPs to ES and words as described above, the experiment can also inform questions about talker normalization. Talker normalization is the process by which listeners resolve acoustic-

phonetic ambiguities that are introduced when a talker changes (Nusbaum & Morin, 1992). A great deal of variability in the acoustic speech signal is introduced when a talker changes, yet people are still able to quickly and easily understand multiple talkers (Heald & Nusbaum, 2014; Nusbaum & Magnuson, 1997). Though listeners are not usually aware of the fact that they must adapt to the characteristics of the new talker, there is extensive evidence that this process requires attention and working memory (WM). For example, listeners recognize spoken target syllables more slowly when the talker changes under high WM load conditions; yet when the talker does not change, the high WM load does not have the same effect (Nusbaum & Morin, 1992). These results strongly suggest that adapting to a new talker draws on WM resources. Response time data indicate that these results also hold for vowels and whole words; participants are slower to categorize vowels when the talker is changing (Kaganovich, Francis, & Melara, 2006), and slower to classify words as beginning with a “b” or “p” when the number of talkers is higher (Mullennix & Pisoni, 1990). Participants are also slower to identify target words in a list when the talker is changing (Wong, Nusbaum, & Small, 2004).

There has been limited neuroimaging work on talker normalization. Wong et al. (2004) used fMRI to examine activity related to changes in talker. In this study, participants listened to lists of individual spoken words, and their task was to recognize a target word in the list. Each participant listened to lists in two conditions: in the “blocked” condition, the words were all spoken by the same talker. In the “mixed” condition, the words were spoken by four different talkers. In addition to replicating previous behavioral effects showing that participants are slower to respond when the talker is changing, the fMRI paradigm allowed Wong et al. to identify brain areas that were differentially active in mixed and blocked conditions. They found two areas that responded significantly differently to mixed and blocked talkers; these were middle/superior temporal areas

([56 -28 1] and [56 -28 2]), and the superior parietal lobule ([33 -67 45] and [33 -63 48]). The authors reasoned that the difference in temporal area activation might correspond to increased processing difficulty in the mixed talker condition, and the difference in parietal lobule activation might correspond to a refocusing of spectral and spatial attention that takes place when talkers change.

While experiments like those by Wong et al. are important for understanding the neural substrates of talker normalization, they are limited by the poor time resolution of fMRI compared to techniques such as EEG. It has been known for a long time that attention can affect the N1 and P2 ERPs; in general attended stimuli evoke larger N1 and P2 potentials (Picton & Hillyard, 1974). To the extent that talker normalization involves attentional mechanisms, we might reasonably expect changing talkers to affect the N1 and P2. In fact, existing ERP talker normalization work has suggested that talker normalization affects the N1 (i.e., 100 ms after stimulus presentation; Kaganovich et al., 2006; Zhang, Peng, & Wang, 2013). These findings support ideas that talker normalization involves attentional reallocation to new features of the acoustic speech signal (Heald & Nusbaum, 2014; Nusbaum & Magnuson, 1997; Nusbaum & Schwab, 1986). However, to our knowledge, fMRI and EEG have not been combined (nor source localization performed using high-density EEG data) to investigate the correspondence between these N1 effects and the brain areas identified via fMRI in, for example, Wong et al. (2004). Moreover, descriptions of talker change effects on potentials before and after the N1 (e.g., the P2) are rare.

Much of the literature on talker normalization examines vowels (e.g., Kaganovich et al., 2006), syllables (e.g., Morin & Nusbaum, 1989), or words (e.g., Mullennix & Pisoni, 1990) without incorporating these smaller units into the context of a full sentence. Though the present experiment was designed as a control to the environmental sound vs. word experiment described

in chapter 4, it is nevertheless an interesting framework in which to ask whether the previously observed neural signatures of talker normalization will still be present when the stimuli are fluent, whole sentences rather than lists of isolated words or vowels. Moreover, the task in this experiment differs from tasks in other talker normalization work in important ways. In the current experiment, participants focus on the meaning of the sentence as a whole, without being asked to perform any task such as recognizing a target word. It is an open question whether neural signatures of talker normalization will still be present in whole, fluent sentences where participants are not required to make behavioral responses.

If talker normalization does indeed occur in whole, fluent sentences without a behavioral task, we would expect to see N1 time frame differences between “same” and “different” talker conditions (following Kaganovich et al., 2006; Zhang et al., 2013). Because the refocusing of attention to enable processing of a new talker likely occurs quite early in the processing stream, we might also see effects earlier than the N1. We do not know to what extent later potentials will be influenced by early differences, and thus our analysis of later time windows will be exploratory. Finally, if talker normalization in our current paradigm is mediated by the same mechanisms proposed by Wong et al. (2004), we expect that talker normalization will depend on WM, and that differences in ERPs will be characterized by topographies consistent with activity changes in either middle/superior temporal areas, the superior parietal lobule, or both.

5.2 Methods

5.2.1 Participants

Participants were twenty-two (12 female, 10 male) adults from the University of Chicago and surrounding community. Their mean age was 19.59 years (SD: 1.0, range: 18-21). Twenty-one were right-handed and one was left-handed. Participants completed questionnaires to ensure

that they knew English to native proficiency, and that they were not taking medications that could interfere with cognitive or neurological function (Appendix 4). Participants received \$30 cash for their participation. We set our target number of participants based on a goal of matching the previous experiment as closely as possible.

5.2.2 Stimuli

The stimuli were the same as those used in Experiment 3 (4.2.2 Stimuli), except that instead of sentences ending in environmental sounds, sentences ending in a different talker were presented. Whereas the speaker for the sentence stems was an adult male speaker of Midwestern English, the speaker for the ending words in half the blocks was an adult female speaker of Midwestern English. As in Experiment 3, sentences were blocked by target type, such that there were four blocks of sentences ending in the male talker, and four blocks of sentences ending in the female talker. Block types alternated across the experiment, and the type of starting block (i.e. male or female talker) was counterbalanced across subjects. Subjects were told repeatedly that a different talker would be saying some of the words, and warned before blocks in which the last word was said by the female talker. Stimuli were experienced at 65-70 dB over insert earphones (3M E-A-RTone Gold) and were presented using Matlab 2015 (MathWorks, Inc., Natick, MA) with Psychtoolbox 3 (Brainard, 1997; Kleiner et al., 2007).

5.2.3 Testing procedure

The testing procedure was exactly the same as for Experiment 3 (4.2.3 Testing procedure).

5.2.4 EEG Setup

The EEG setup was exactly the same as for Experiment 3 (4.2.4 EEG Setup).

5.2.5 Data preprocessing

Preprocessing was done according to the same procedure as in Experiment 3 (4.2.5 Data preprocessing). Participants with 50% or more artifact-contaminated trials in any one condition were removed from further analysis. This procedure resulted in removal of one participant who lost over half the trials in all conditions.

Electrode coordinates from individuals' net placement images were used to assign individual sensor locations for each participant.

5.2.6 Analyses

5.2.6.1 Topographic Analyses

As in Experiment 3, we used BESA 6.0 to generate participant-level averaged waveforms and [mismatch – match] difference waves. We also used BESA to create ascii files of time-varying voltage at every electrode; these were used for topographic analysis in RAGU (Randomization Graphical User interface, Koenig, Kottlow, Stein, & Melie-García, 2011).

As in Experiment 3, we performed two analyses using 5,000 randomizations of the data in RAGU. The first analysis was intended to address the question of whether changing the talker affects the way the word is understood in sentence context. In this analysis, we analyzed [mismatch – match] difference topographies, similarly to the analysis in Experiment 3. We examined the difference topographies for main effects of sentence constraint (specific vs. general) and talker (same vs. different). This analysis identified time windows where there were significant main effects and interactions; it also output scalp topographies for the different conditions at each time point. Also mirroring Experiment 3, the second analysis, also involving 5,000 randomizations, included only factors of talker (same vs. different) and congruency (match vs. mismatch), as it pooled together the two constraint levels.

Finally, we ran one experimental analysis to examine whether working memory (WM) interacted with ERPs, as based on previous research, we expected WM to affect switching between perception of different talkers. This analysis was identical to the second analysis described above, except that it included WM (as measured by d' for a 3-back task) as a continuous covariate in the model.

5.2.6.2 Regions of Interest (ROIs)

In order to represent most of the topography of the scalp in our analysis without arbitrarily choosing just a few electrodes, data were pooled into nine ROIs in a fashion similar to Potts & Tucker (2001), who used four adjacent electrodes in each ROI. These ROIs were the same as in Experiment 3, and can be seen in Table 4.1. The pooled ROI data were used for two purposes: 1) to represent voltage traces in figures, and 2) for statistical analysis of latency data described in the next section.

5.2.6.3 Latency Analysis

As in Experiment 3, we used the time point dividing the area under the [mismatch – match] difference curve into two equal halves to estimate latency of the N400. The current experiment identified a much shorter N400 time window than in Experiment 3. Therefore, to make sure we were looking at a large enough N400 time window, we used the same time window as in Experiment 3 (309 - 512 ms post target onset), as those data showed that we need not expect the N400 to fall outside this window for the sentences and endings used. For each subject, and for sounds and words separately, we pooled electrodes into the nine ROIs described above. These data were entered into a repeated measures ANOVA in R using the “ez” package (Lawrence, 2013). The dependent variable was latency in milliseconds post-target-onset; within-subject factors were ROI (9 levels, Table 4.1) and condition (speech vs. non-speech).

5.2.6.4 Microstate Analysis

As an exploratory analysis to address questions about talker normalization mechanisms, we fit microstates to the data using RAGU. The optimal number of microstates to explain the data was assessed by fitting three (minimum) up to ten (maximum) distinct microstates to the data; three microstates best explained the data as including more microstates did not explain more of the variance. The scalp topographies across the time course of the ERPs were then categorized into three microstates using an Atomize and Agglomerate Hierarchical Clustering (AAHC) algorithm. This classifier has an advantage over k-means clustering in that it prioritizes topographies with higher global field potential as stable microstates, whereas k-means clustering treats all topographies equally (Murray, Brunet, & Michel, 2008). A side effect of this property is that k-means clustering can erase stable topographies of short duration, subsuming them into longer-lived microstates. Because we were particularly interested in early differences between same and different talker conditions that might indicate early attentional processes, we reasoned that some of our microstates of interest might be fairly short-lived (as opposed to longer, slower potentials that occur later in the ERP) and thus chose AAHC over k-means.

5.2.6.5 Source Analysis

In order to relate our findings to the brain areas identified in the talker normalization work by Wong et al. (2004), we performed an exploratory analysis involving source analysis using BESA 6.0. For source analysis, the lowpass filter was disabled as it can lead to distortion in the process of transforming topographies into sources (BESA 6.0). The time periods for source analysis were chosen as follows: 1) before or during the ACC/N1-P2 time window i.e. before 300 ms post-target onset, 2) identified by microstate or topographical randomization analyses to have main effects of talker, main effects of WM, or an interaction between these. Largely overlapping

time windows were combined into single windows, e.g. a window from 124 – 147 ms was combined with a window from 136 – 179 ms to make a larger window from 124 – 179 ms (Table 5.6). This was done in order to reduce the number of comparisons in further statistical testing.

In each of these time windows, we fit three models to data on a participant level: a “parietal model” including source dipoles at the parietal locations identified by Wong et al. (2004) to be involved in talker normalization ([33 -67 45] and [-33 -67 45]), a “temporal model” including [56 -28 1] and [-56 -28 1] (also identified by Wong et al.) and a “both” model including all four coordinates as sources. The subject-level averages used for source localization were “different talker” and “same talker”, pooled across different constraint levels and congruency status. The models produce a residual variance (RV) i.e., the amount of variance in the scalp voltage maps that is left unexplained by the sources included in the model. Thus, a smaller RV indicates a better fit.

5.3 Results

5.3.1 Experiment 3 Follow-up Analyses

We expected this paradigm to produce a congruency-sensitive N400; as in Experiment 3, our randomization analysis using raw topographies identified a window with a main effect of congruency in the vicinity of the traditional N400 ($p < 0.05$, Table 5.1, Figure 5.1c, Table 5.3). However, unlike Experiment 3, this window was quite short (only 10 ms, from 433 to 443 ms post target onset), and the effect size was rather small (Figure 5.1b, Table 5.3). Despite the short duration of this window, the topographies inside it do indicate larger central negativities for mismatch endings (Figure 5.1a and d), as expected.

We also expected a congruency-sensitive P600 similar to the short-lived one we found in Experiment 3. Confirming this expectation, the analysis of raw topographies also identified main

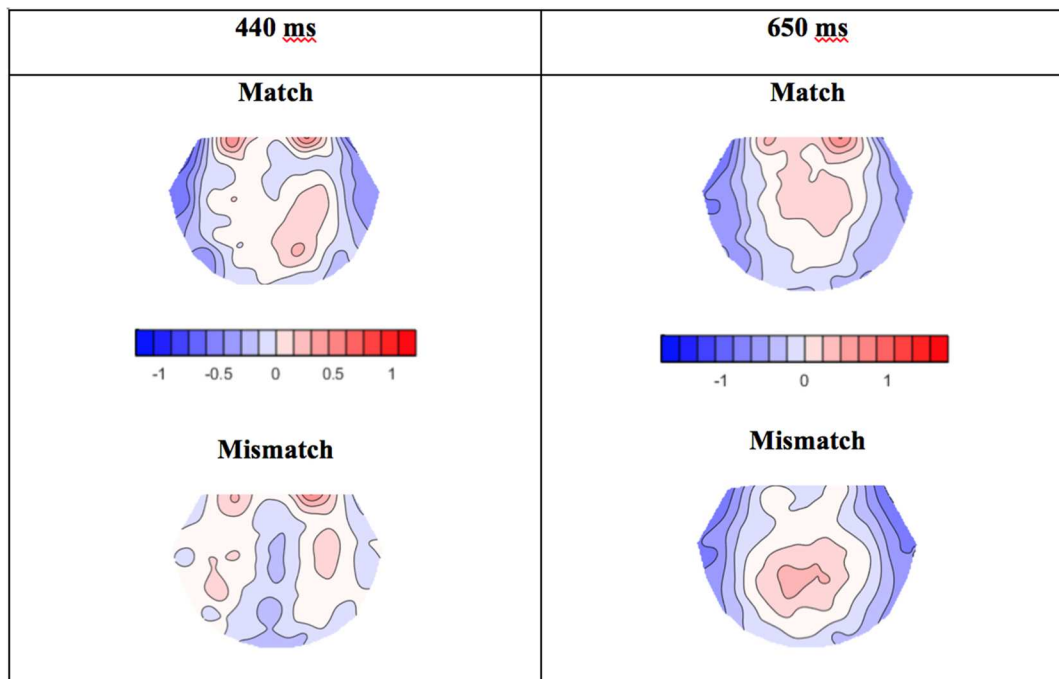
effects of congruency in the vicinity of the P600 ($p < 0.05$, Table 5.3). Whereas in Experiment 3, these P600 effects were short-lived, they were longer and more robust in the present experiment (Table 5.1, Table 5.3, Figure 5.1b and c). As in experiment 3, these P600 effects involved stronger posterior positivities in mismatch conditions (Figure 5.1a and d).

Table 5.1 Raw topography TANOVA windows. Significance windows identified by the randomization analysis (TANOVA) of raw topographies. There were no significant interactions of congruency (match/mismatch) and talker (same/different). There were, however, large differences in topography based on talker, as well as a short N400 and longer P600 congruency main effects. Windows that do not pass the 40 ms duration threshold, but that are located where we had an *a priori* expectation of a main effect, are italicized.

Factor	Window Start	Window End	Length	≥ 40 ms
Congruency main	18	25	7	FALSE
<i>Congruency main</i>	<i>433</i>	<i>443</i>	<i>10</i>	<i>FALSE</i>
Congruency main	589	621	32	FALSE
Congruency main	628	675	47	TRUE
Congruency main	686	818	132	TRUE
<i>Talker main</i>	<i>124</i>	<i>147</i>	<i>23</i>	<i>FALSE</i>
Talker main	235	278	43	TRUE
Talker main	314	394	80	TRUE
Talker main	437	498	61	TRUE
Talker main	520	640	120	TRUE
Talker main	713	729	16	FALSE
Congruency * Talker	NULL			

Figure 5.1 N400 and P600 match vs. mismatch topographies. **a)** Scalp topographies for match and mismatch endings at 440 and 650 ms post ending onset. Blue indicates negative potential; red indicates positive potential. **b)** Time-varying generalized dissimilarity between raw match and mismatch topographies. To give a sense of the meaning of this effect size, the mean and 95% CI for the generalized dissimilarity expected due to random chance (estimated from randomizing the data) is also represented.

a)



b)

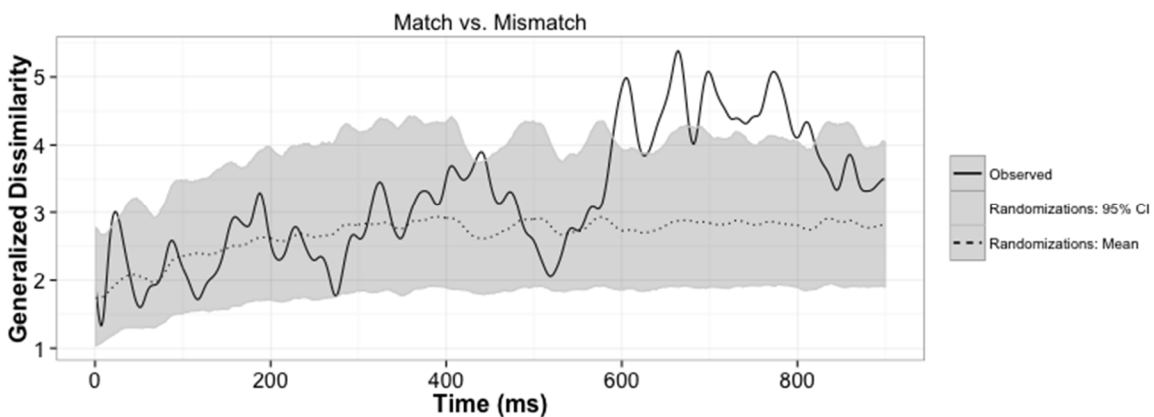
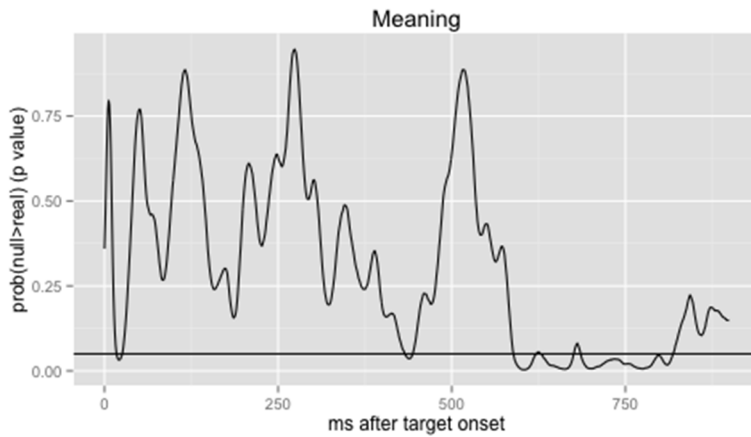
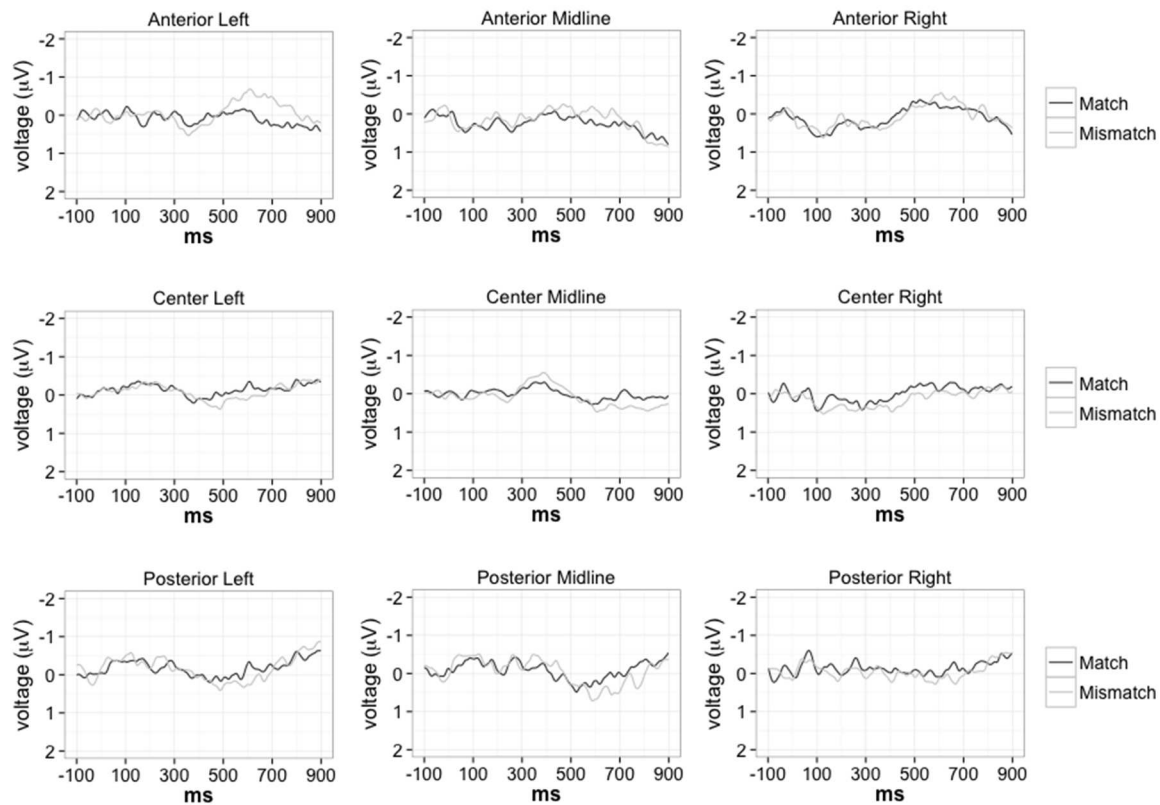


Figure 5.1 N400 and P600 match vs. mismatch topographies, *cont.* **c)** Time-varying p-value, i.e., proportion of randomizations leading to a larger effect size than observed. We can see a brief main effect of congruency near 400 ms (the N400) as well as longer-lasting post-600 ms effects (P600). **d)** Voltage traces for pooled match and mismatch sentence endings in the nine examined ROIs. Note that negative is up and time = 0 ms corresponds to ending onset.



d)



We also expected to find constraint effects like those in Experiment 3, characterized by a frontal shift of the N400 under lower-constraint conditions. Unlike Experiment 3, in which this constraint effect showed up as a main effect of constraint in the N400 time window in the analysis of [mismatch – match] difference topographies, we found no long-lived effects of constraint in the present experiment ($p > 0.25$, Table 5.2, Table 5.3). In fact, the observed generalized dissimilarity (GD) between general and specific conditions in the time window where the effect was observed in Experiment 3 (359 – 421 ms) was less than the mean expected GD under the null hypothesis in this time window (5.23 as opposed to 5.27, Table 5.3). Whereas in Experiment 3, difference topographies showed a pronounced frontal shift to the N400 for low-constraint sentences, this shift was not statistically supported in the present experiment, although it is possible to convince oneself that there is a hint of this frontal shift for the “general” condition (Figure 5.2). Potential reasons for the absence of constraint effects will be discussed in 5.4 Discussion. Similar to Experiment 3, the difference topography analysis did not reveal any main effects of talker long enough to be convincingly distinct from noise, or interactions between constraint and talker (Table 5.2, Figure 5.3).

Table 5.2 Difference topography TANOVA windows. Significance windows identified by the randomization analysis (TANOVA) of [mismatch – match] difference topographies. There were no significant interactions of constraint (general/specific) and talker (same/different).

Factor	Window Start	Window End	Length	≥ 40 ms
Constraint main	21	45	24	FALSE
Constraint main	536	548	12	FALSE
Constraint main	766	781	15	FALSE
Talker main	488	493	5	FALSE
Constraint * Talker	NULL			

Figure 5.2 General vs. specific N400 difference topographies. Scalp topographies for [mismatch – match] difference topographies for endings following low constraint (general) vs. high constraint (specific) sentences at 400 ms post target onset. Blue indicates negative potential; red indicates positive potential. Note that these topographies are not statistically different from each other.

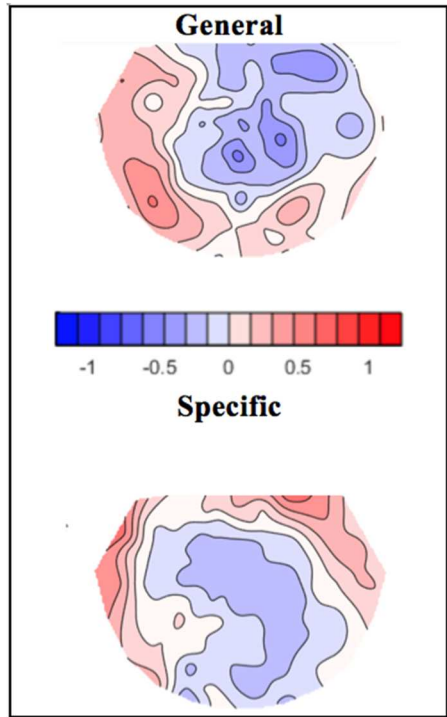


Figure 5.3 Time course of GD for match vs. mismatch. Time course of generalized dissimilarity for [mismatch – match] difference topographies indicates that there were no substantial **a)** main effects of talker, or **b)** interactions between talker and constraint, on this difference topography anywhere in the time course of the ERP.

a)

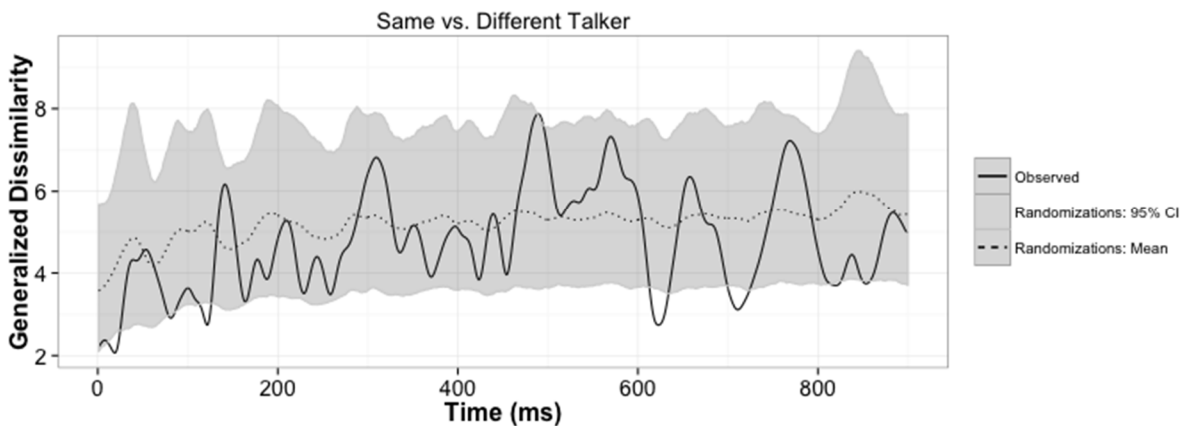


Figure 5.3 Time course of GD for match vs. mismatch, *cont.*

b)

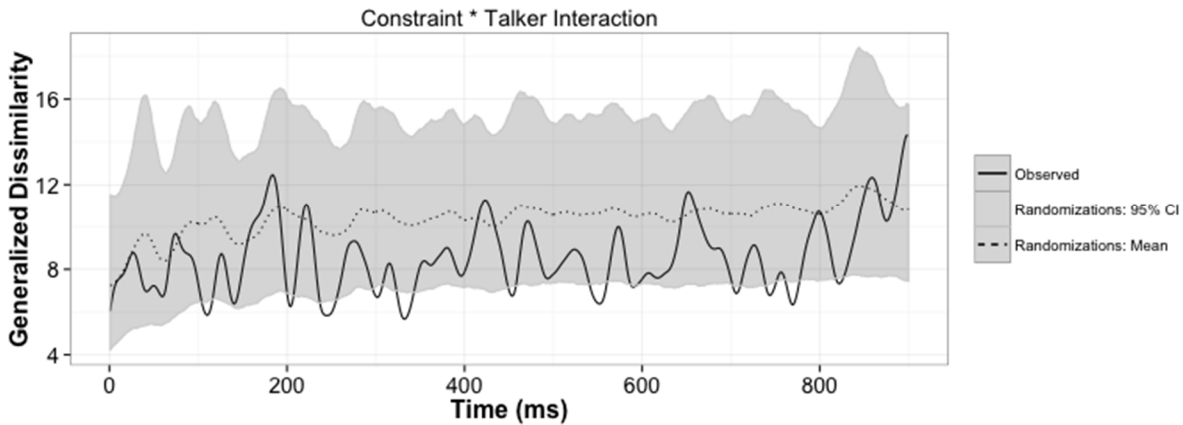


Table 5.3 Summary of statistics for randomization analyses of topographical maps.

window	analysis	factor	observed gen. dissim. (GD)	expected GD under null hyp.	95% CI of GD under null hyp.	p
433 - 443 ms	raw voltage, congruency x target type	congruency (match vs. mismatch)	3.85	2.62	1.80 – 3.75	*0.040
433 - 443 ms	raw voltage, congruency x target type	congruency * target type interaction	5.55	5.24	3.60 – 7.48	0.33
628 - 675 ms	raw voltage, congruency x target type	congruency (match vs. mismatch)	4.64	2.80	1.89 – 4.07	*0.019
124 - 147 ms	raw voltage, congruency x target type	Talker (same vs. different)	3.91	2.41	1.56 – 3.62	*0.027
359 - 421 ms	mismatch - match difference, constraint x talker	constraint (general vs. specific)	5.23	5.27	3.62 – 7.59	0.47

If perceiving a different talker at the end of a sentence introduces delays that persist into later processing stages, we might find a delayed N400 in the different talker condition. Given that

no significant difference in N400 latency was found for speech and ES in experiment 3, however, we thought latency differences would be unlikely. In fact, no latency differences were found; a 2 x 9 (target type [same/different talker] x ROI [AL, AR, AM, CL, CR, CM, PL, PR, PM]) repeated measures ANOVA of [mismatch – match] difference wave latencies showed that there was no main effect of target type on latency [$F(1,20)=4.51E-5$, $p>0.25$, $\text{diff}=0.03$ ms, $d=0.04$, 95% CI (-8.08 - 8.13 ms)]. The mean latency for different talker was 411.42 ms [406.16 - 416.68] and for words was 411.39 ms [404.41 to 418.37]. Incidentally this latter latency is extremely close to what we found for these stimuli in Experiment 3, which corroborates our previous data. There was also no evidence of a main effect of ROI [$F(8,160)=1.49$, $p=0.16$] or an interaction of target type and ROI [$F(8,160)=0.38$, $p>0.25$], indicating that N400 latencies were not statistically distinguishable between different regions of the scalp (Table 5.4).

Table 5.4 Summary of statistics for repeated measures ANOVA examining N400 latency.

analysis	factor	latency difference	F	Cohen's d	95% CI	p
latency repeated measures ANOVA	talker	0.03 ms	4.51E-5	0.04	-8.08 – 8.13	>0.25
latency repeated measures ANOVA	ROI	-	1.49	-	-	0.16
latency repeated measures ANOVA	target type * ROI	-	0.38	-	-	>0.25

Finally, as in experiment 3, there were widespread differences in the raw topographies between the different ending types (Table 5.1, Figure 5.4). As in experiment 3, windows with significant differences included time windows characteristic of the acoustic change complex (ACC): 124 – 147 ms (N1 window) and 235 - 278 ms (P2 window) (Table 5.1). However, when the topographies are examined, it is clear that these differences do not include the ACC's canonical

central negativity followed by central positivity (Figure 5.4 a). Instead, topographies at both time points are characterized by a left-lateralized, frontocentral positivity. In the same talker condition in the N1 window, this positivity is encroached upon by a posterior central negativity that is absent in the different talker condition. In the P2 window, the positivity is more frontal in the different talker condition than in the same talker condition (Figure 5.4a). It is also worth noting that the double-peaked N400 that appeared in Experiment 3 for the ES condition is absent from the waveforms for both the same and different talker conditions (5.3.4d).

Figure 5.4 N1 and P2 window topographies across talker. a) Scalp topographies for sentence endings said by the same vs. a different talker at 140 and 250 ms post ending onset. Blue indicates negative potential; red indicates positive potential.

a)

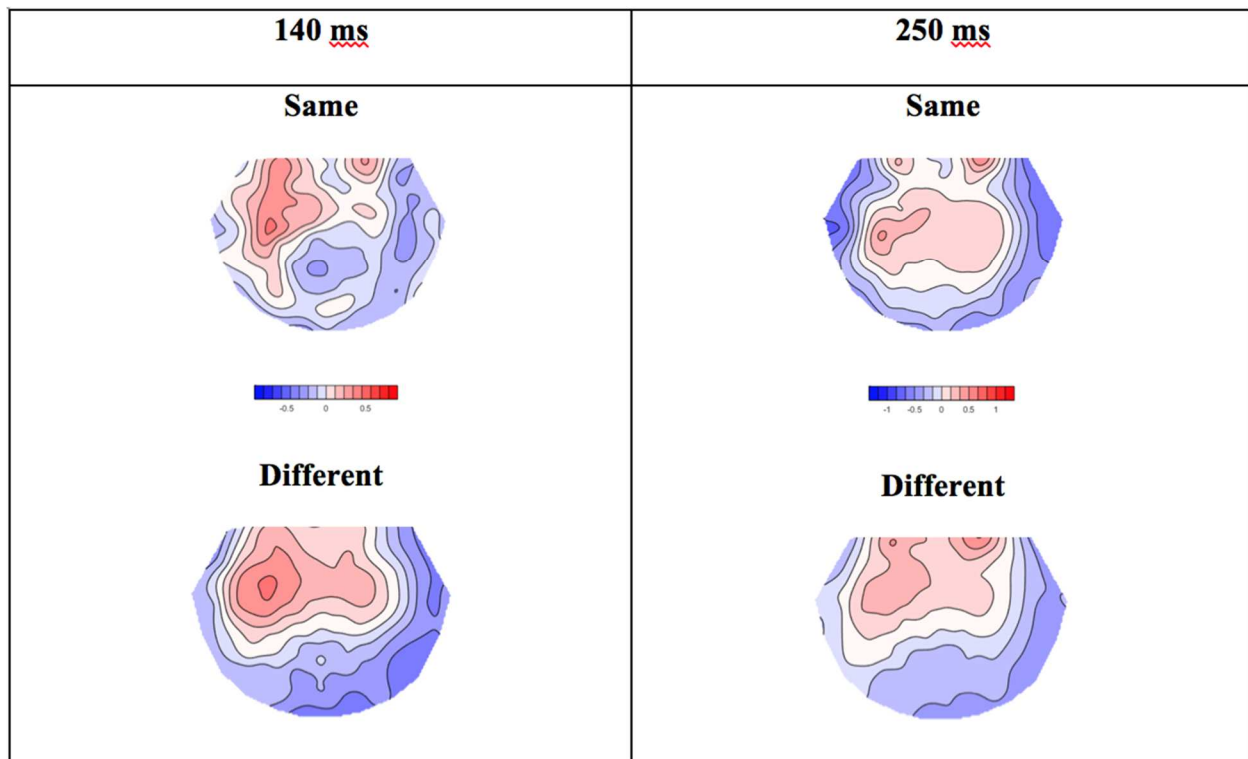
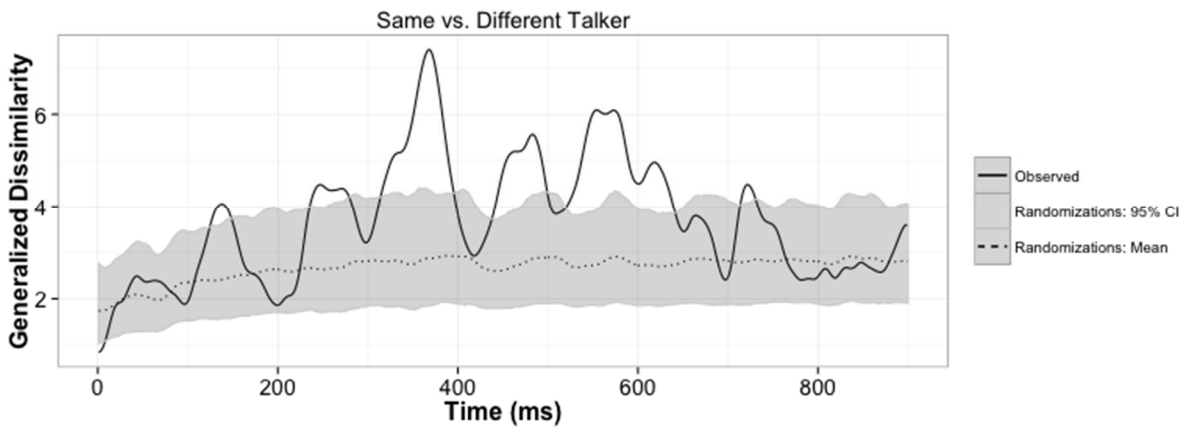


Figure 5.4 N1 and P2 window topographies across talker, *cont.* **b)** Time-varying generalized dissimilarity between raw same and different talker topographies. To give a sense of the meaning of this effect size, the mean and 95% CI for the generalized dissimilarity expected due to random chance (estimated from randomizing the data) is also represented. **c)** Time-varying p-value, i.e., proportion of randomizations leading to a larger effect size than observed. We can see widespread main effects of talker across the entire ERP.

b)



c)

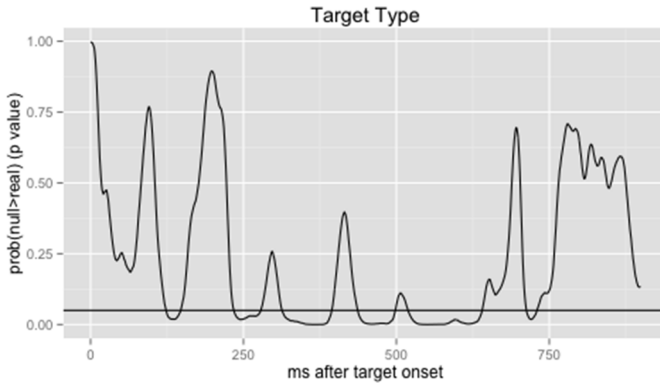
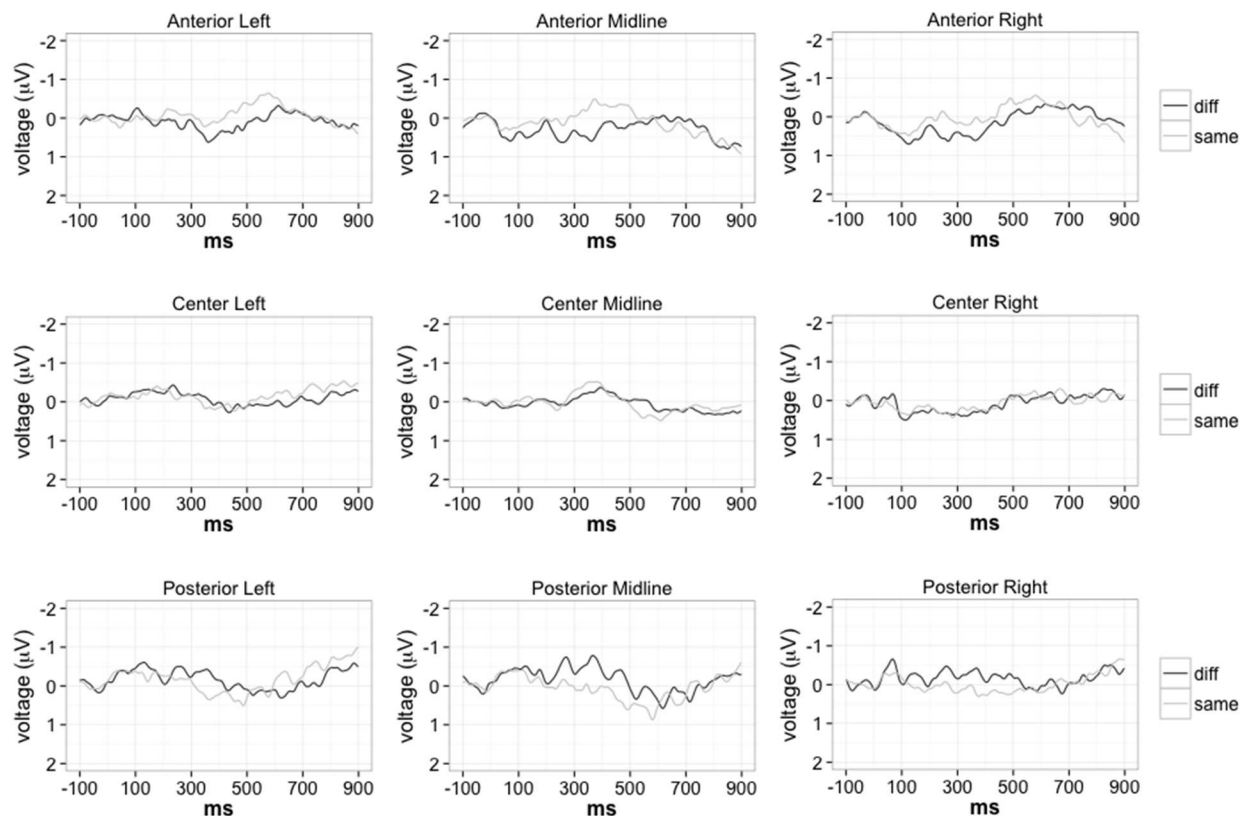


Figure 5.4 N1 and P2 window topographies across talker, *cont.* d) Voltage traces for pooled same and different talker sentence endings in the nine examined ROIs. Note that negative is up and time = 0 ms corresponds to ending onset.

d)



5.3.2 Exploratory analyses—talker normalization

If, as suggested by a large body of literature, adapting to a new talker involves early WM-mediated attentional processes, we might expect that differences between ERPs to the same vs. a different talker might manifest as differences in the early portion of the ERPs as the listener quickly adapts to the new talker. We tested this idea in several ways. First, we performed an additional randomization analysis that included WM as measured by d' on a 3-back task. As expected, this analysis identified the same windows with main effects of talker as already described (Table 5.1; 124 – 147 and 235 – 278 ms). While this analysis did not identify any time windows in which WM

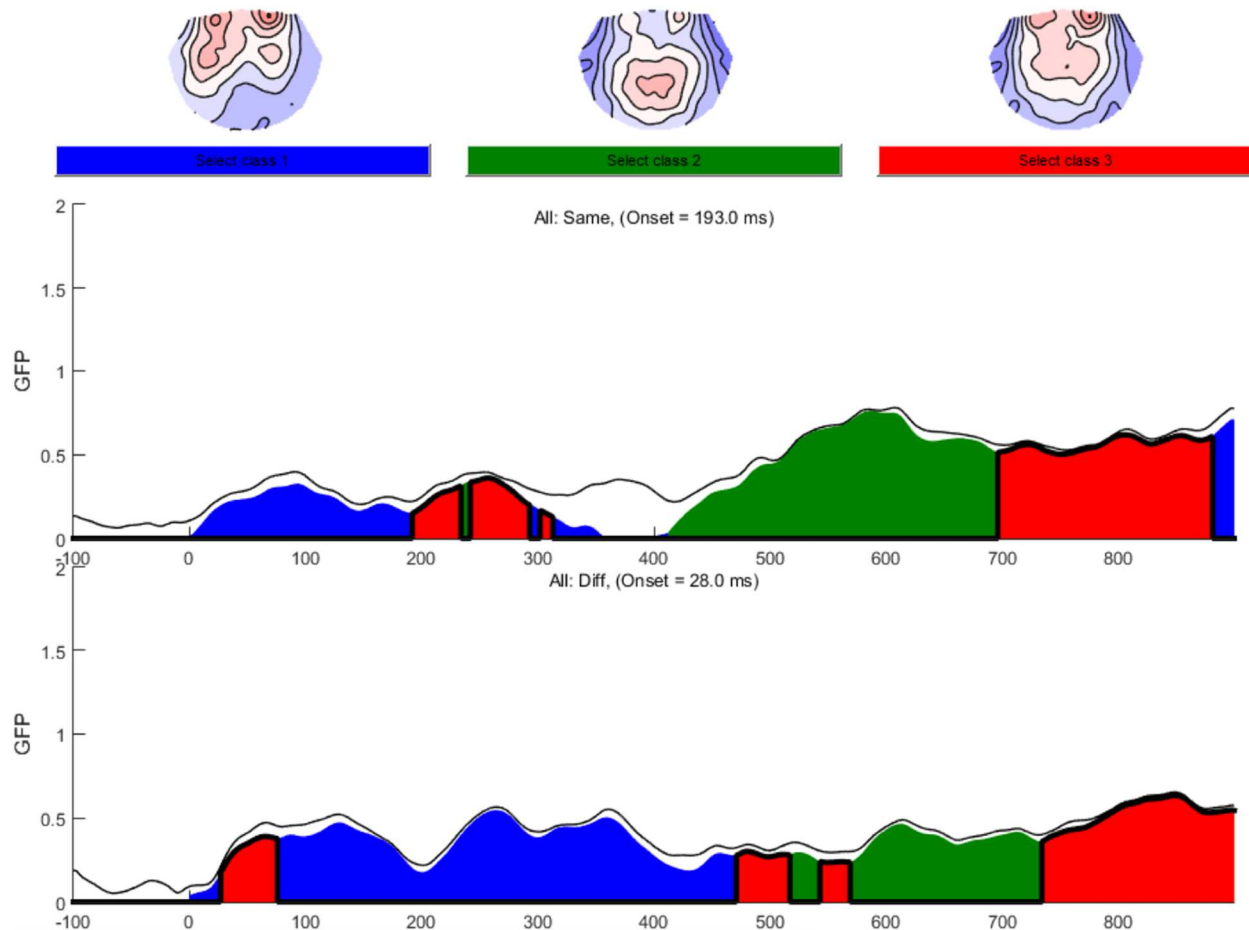
interacted with talker, it did identify several windows, two of which fell into the early ACC-and-prior time window, where there were main effects of WM (Table 5.5).

Table 5.5 TANOVA including WM. Results of a randomization analysis including working memory (d' on a 3-back task) as a covariate.

Factor	Window Start	Window End	Length	>=40 ms
d' main	3	29	26	FALSE
d' main	136	179	43	TRUE
d' main	220	258	38	FALSE
d' main	300	323	23	FALSE
d' main	690	700	10	FALSE
d' * Talker	NULL			

Another way we investigated for talker-related differences in the early portion of the ERPs was by performing a microstate analysis. This analysis identified one particular such difference in the first 150 milliseconds post-target onset. When the ending word was said by the same talker as the rest of the sentence, this region of the ERP was dominated by microstate class 1. However, when the talker changed on the last word, the ERP between 27 and 75 ms was dominated by microstate class 3 (Figure 5.5).

Figure 5.5 Microstates. Microstate analysis of responses to words said at the end of a sentence by the same vs. a different talker.



In order to further investigate this difference in microstates, as well as the time windows identified as having main effects of talker and WM, we used source analysis in BESA 6.0 to estimate the sources in the brain that led to this scalp topography. If the temporal and parietal coordinates identified by Wong et al. (2004) as involved in talker normalization are more active for the different talker condition, we would expect to see systematically lower residual variances or RVs (i.e. better fits) for our models in the different talker condition when compared to same talker. In fact, for every single time point examined, the average RV was lower for the different talker condition, suggesting that the sources from Wong et al. better explained the different talker data than the same talker data (Table 5.7). To statistically test these differences, we compared the

RV for each model, at each time point, between same and different talker using one-sided, paired Wilcoxon rank-sum tests. As this led to twelve comparisons, a Bonferroni-corrected threshold of $p < 4.2\text{E-}3$ was used.

Table 5.6 Time windows examined in source analysis. Numbers are in milliseconds after target onset.

Start	Stop
3	29
27	75
124	179
220	278

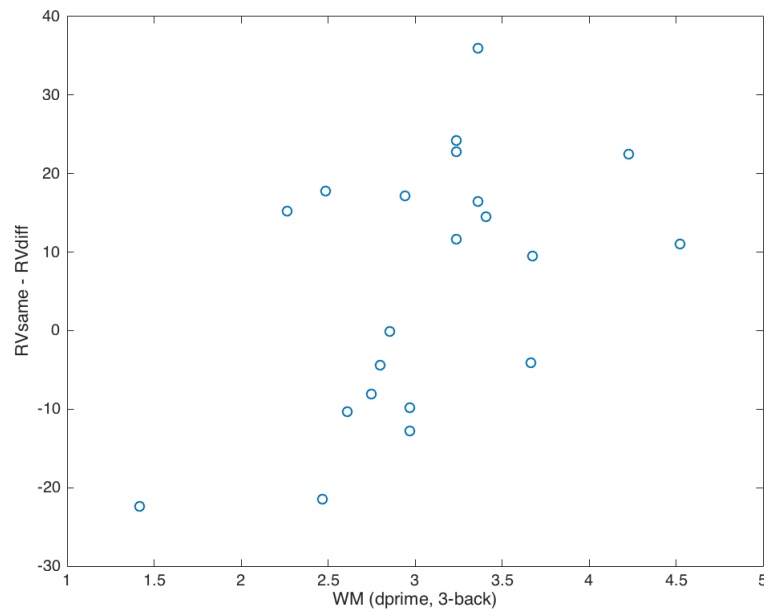
Table 5.7 Residual variance for source analysis models. Average RV for different vs. same talker in different time windows, with parietal, temporal, and “both” models tested separately. Bold italics denotes the window where the RV for different and same talker was significantly different after controlling for multiple comparisons. *Note that a lower RV indicates more variance explained by the model and thus indicates a better fit.*

Time win	Model	Different	95% CI	Same	95% CI
3-29 ms	parietal	49.77	42.32 - 57.23	59.64	51.33 - 67.95
	temporal	42.88	36.17 - 49.60	49.32	41.36 - 57.29
	both	33.27	26.61 - 39.92	40.41	33.15 - 47.67
27-75 ms	parietal	50.02	44.18 - 55.85	53.70	47.40 - 60.01
	temporal	39.89	33.35 - 46.44	45.32	38.04 - 52.61
	both	30.93	24.92 - 36.95	34.62	29.17 - 40.06
124-179 ms	parietal	50.86	45.35 - 56.36	56.83	51.63 - 62.03
	temporal	40.36	34.97 - 45.74	47.00	40.29 - 53.72
	both	33.32	28.50 - 38.13	37.67	31.39 - 43.95
220-278 ms	parietal	55.35	48.67 - 62.03	62.29	56.41 - 68.17
	<i>temporal</i>	<i>42.84</i>	<i>36.69 - 48.98</i>	<i>51.24</i>	<i>45.48 - 57.01</i>
	both	35.08	28.23 - 41.93	42.17	35.81 - 48.52

In most of the time windows, the difference between the fit of the models in the different and same talker conditions was not significantly different ($V \leq 174$, $p \geq 0.021$). However, in the 220 – 278 ms

time window, the model including temporal sources was a significantly better fit for the different talker condition ($V=191$, $p=3.6\text{E-}3$, $d=0.6$). In order to further explore what might have been happening in the other time windows, we performed one further analysis. Namely, we calculated correlations between WM as measured by d' on the 3-back task, and the fit of the source models to the scalp topography elicited by a changing talker. We reasoned that even though the pre-P2 windows did not show any significant differences between the fit of the models to different vs. same talker topographies, perhaps there was some relationship between the fit of the models and WM. In order to capture possible correlations related to adapting to a changing talker, we correlated WM with the difference between model fit in the same and different talker conditions, that is: $RV_{same} - RV_{diff}$. While in most of the time windows (including P2 windows), correlations were not significant ($r<0.4$), in the N1 time window (124 – 179 ms), there was a significant correlation between WM and $RV_{same} - RV_{diff}$ for the parietal model (Figure 5.6, $r=0.51$, $p=0.018$), which likely led to the marginally significant correlation between WM and $RV_{same} - RV_{diff}$ for the model including both temporal and parietal sources ($r=0.41$, $p=0.06$). The $RV_{same} - RV_{diff}$ of the temporal-only model was not significantly correlated with WM in this time window ($r=0.34$, $p=0.13$). This indicates that as WM increases, so does the gap between neural responses to same and different talker, at least when measured by how well the parietal source model explains that data. Keep in mind that a high $RV_{same} - RV_{diff}$ indicates that the parietal model is better at explaining the different-talker topography than the same-talker topography, while an $RV_{same} - RV_{diff}$ of zero indicates that the parietal model is equally good at explaining the topography in the same and different talker conditions, and a negative $RV_{same} - RV_{diff}$ indicates that the parietal model is better at explaining the same-talker topography.

Figure 5.6 Correlation between source analysis model fit and WM. For the N1 time window, correlation between WM as measured by d' on a 3-back task, and the difference in the goodness of fit of a parietal source model to scalp topography data between same and different talkers ($RV_{same} - RV_{diff}$).



5.3.3 Do WM-related mechanisms mediating talker normalization also mediate the understanding of ES in sentence context?

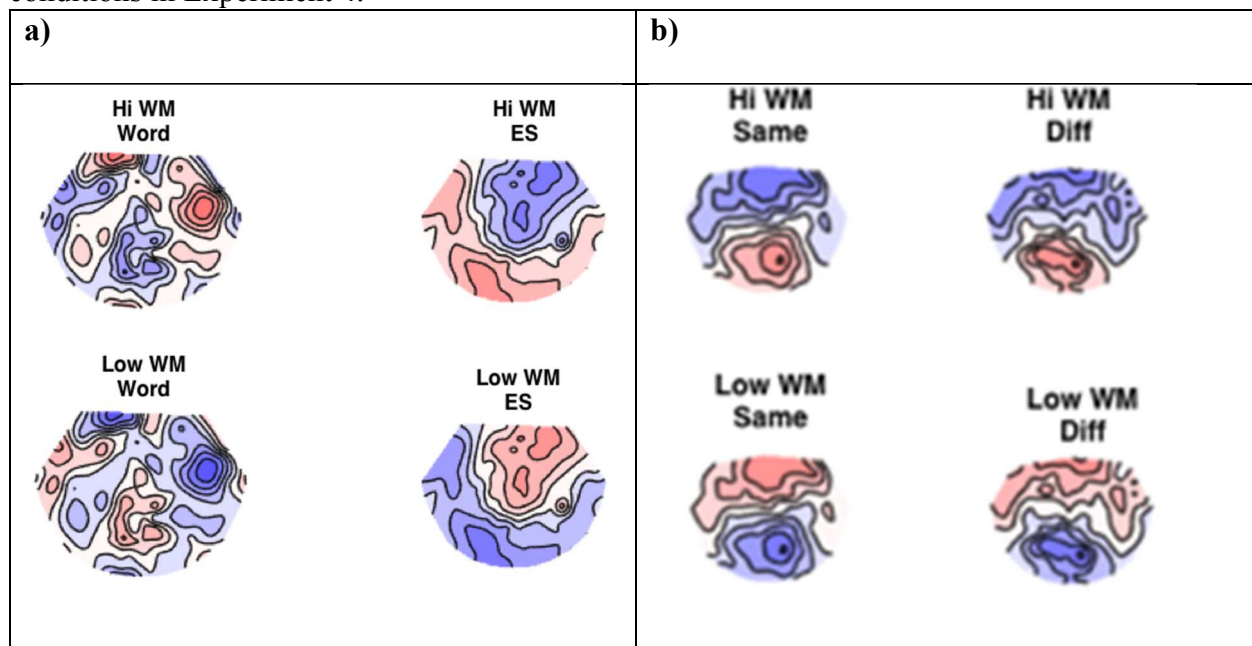
As described in the previous section, the WM and source analyses for Experiment 4 revealed significant main effects of WM, as well as implicating the sources described by Wong et al. (2004). If language capabilities are specialized, these mechanisms might be recruited only to deal with changes in speech, such as a change in talker. However, if the basis of language is more cognitive-general, we might expect that these mechanisms could be recruited to cope with a transition from language to non-language. Therefore, we asked, “is WM recruited to process ES in sentence context in a way similar to its recruitment for spoken words when the talker changes?” In order to address this question, we repeated the analysis described in the previous section (Section 5.2.2 Exploratory Analyses), this time using the data collected in Experiment 3, that is,

the responses to words vs. ES in sentence context. First, we ran a TANOVA including WM as a continuous covariate, with additional factors of congruency (match vs. mismatch) and target type (word vs. ES). While unlike Experiment 4, this analysis did not identify any windows with main effects of WM, it did reveal three windows, also in the vicinity of the N1 and P2, where there was an interaction between target type and WM (Table 5.8).

Table 5.8 TANOVA including WM for Expt 3. Results of a randomization analysis including working memory (d' on a 3-back task) as a covariate for the ES and word data.

Factor	Window Start	Window End	Length
$d' * \text{target type}$	122	137	15
$d' * \text{target type}$	180	219	39
$d' * \text{target type}$	273	293	20

Figure 5.7 Scalp topographies dependent on WM at **a)** 200 ms post-target onset for Experiment 3 (ES vs. words), and **b)** at 250 ms post-target onset for Experiment 4 (same vs. different talker). Note that the topographies for ES in experiment 3 are similar to the topographies for both conditions in Experiment 4.



Notably, the scalp topographies in these windows for ES only (not spoken words) corresponded with the scalp topographies seen for both same and different talker in Experiment 4

(Figure 5.7); specifically, in these cases, the higher the participant's WM, the closer their scalp topography corresponded to a frontal-negative, posterior-positive topography as shown in Figure 5.7. Conversely, the lower their WM, the more closely their scalp topography corresponded to a frontal-positive, posterior-negative pattern.

In order to make sense of these topographical differences, we performed source analysis with temporal and parietal models separately on responses to ES and words in the time windows in which interactions between ES and target type were observed, i.e. 122-137 ms, 180 – 219 ms, and 273 – 293 ms post-target onset. Except for the difference in the source data and time windows, this analysis was identical to that performed in 5.2.6.5 *Source Analysis*. Similar to Experiment 4, we used paired one-sided Wilcoxon rank-sum tests and a Bonferroni-corrected threshold of $p \leq 5.6E-3$ for nine tests. We found that the model including both parietal and temporal sources explained the ES data significantly better than the word data in both the 122 – 137 and 180 – 219 ms time windows ($V=207$, $p=3.7E-3$ and $V=205$, $p=4.6E-3$, respectively; Table 5.9), and approached significance in the 273 – 293 time window ($V=201$, $p=7.0E-3$; Table 5.9). While this comparison did not reach significance in for the other models (i.e., temporal-only and parietal-only), it is worth noting that in none of those cases did the models explain the word data better than the ES data. In fact, in the 122 – 137 time window, both the parietal ($V=202$, $p=6.4E-3$) and temporal ($V=200$, $p=7.8E-3$) models trended towards significance in being a better fit for the ES data. An analysis on a very early control time window, 3 – 29 ms (Table 5.9), confirmed that this pattern is not a universal feature of fits to ES vs. word ERPs; rather, in this early window, there was no hint of a significant difference between the fit of the models to ES vs. word responses, confirming that this effect is likely restricted to the N1 and P2 windows ($V_{parietal}=125$, $p_{parietal}=0.53$; $V_{temporal}=122$, $p_{temporal}=0.56$; $V_{both}=130$, $p_{both}=0.46$). Unlike the talker change experiment,

the fit of the models did not correlate significantly with WM; that is, participants with high WM did not exhibit a different pattern than participants with low WM.

Table 5.9 Residual variance for source analysis models. Average RV for ES vs. spoken words in different time windows, with parietal, temporal, and “both” models tested separately. Bold italics denotes the windows where the RV for ES and words was significantly different after controlling for multiple comparisons. *Note that a lower RV indicates more variance explained by the model and thus indicates a better fit.*

Time win	Model	ES	95% CI	Word	95% CI
3-29 ms	parietal	58.58	49.63 - 67.54	58.19	50.80 - 65.59
	temporal	47.06	38.29 - 55.83	45.37	38.18 - 52.56
	both	38.53	31.31 - 45.74	38.51	31.14 - 45.87
122-137 ms	parietal	49.78	42.81 - 56.74	63.35	54.90 - 71.79
	temporal	38.50	32.40 - 44.60	51.96	43.27 - 60.66
	<i>both</i>	<i>32.08</i>	<i>26.17 - 37.99</i>	<i>44.98</i>	<i>36.53 - 53.43</i>
180-219 ms	parietal	50.99	43.79 - 58.19	63.30	55.26 - 71.34
	temporal	37.77	31.23 - 44.32	51.28	41.82 - 60.74
	<i>both</i>	<i>31.56</i>	<i>24.72 - 38.40</i>	<i>44.49</i>	<i>35.56 - 53.42</i>
273-293 ms	parietal	52.33	45.12 - 59.55	64.35	55.21 - 73.48
	temporal	39.25	32.11 - 46.38	53.73	43.13 - 64.33
	both	31.20	24.72 - 37.68	46.80	36.76 - 56.83

5.4 Discussion

This discussion will be in three parts. The first will discuss the results in the context of the previous experiment, i.e., how the results inform our interpretation of the results on environmental sounds. The second part of the discussion will discuss the results in the context of previous talker normalization work. Finally, the third part of the discussion will extend the talker normalization work—including source analysis and role of WM—to the processing of ES in sentence context.

5.4.1 Discussion – follow-up to ERP work on ES vs. words

Similar to Experiment 3, in the present experiment, we found evidence for a congruency-sensitive N400 characterized by stronger central negativities in mismatch conditions. However,

unlike Experiment 3, in the present experiment this significance window was quite short (Table 5.1, Figure 5.1). This may have been due to a closer temporal correspondence between the stimuli in the present experiment. For example, it is likely that even when two different people say a word at a comfortable talking pace, these two stimuli are closer in length than the word and matched environmental sound. Averaging across stimuli of different lengths may have led to the expanded N400 window we observed in Experiment 3. Also in agreement with experiment 3, we found no latency difference between the two types of sentence endings. This indicates that there is not enough of a difference in difficulty between processing the two types of stimuli to lead to a delayed N400, which was expected based on Experiment 3.

Unlike experiment 3, however, we did not find any convincingly long-lived effects of constraint on the N400 (Table 5.2). This was unexpected; constraint effects are well documented in N400 studies (e.g., Kutas & Hillyard, 1984). However, it may be the case that the constraint effects disappeared due to the repetition in the study design. To obtain enough artifact-free trials, sentences repeated with different endings over the course of the experiment. While in Experiment 3, the environmental sounds might have provided enough of a distraction to prevent participants from memorizing what endings go with which sentence stems, it is possible that the sentences seemed more repetitive in the current experiment, where all endings were verbal even though the talker changed. If due to repetition, participants were able to predict which endings should follow which stems even in the low-constraint condition, this would effectively make all the sentences high-constraint. Scalp topographies support this conjecture; the general sentences in this experiment look more like their specific counterparts, and like the specific sentences in Experiment 3: characterized by a largely central N400 negativity without much of a frontal shift (Figure 5.2).

In order to replicate the frontal shift of the N400 in low-constraint conditions, this experiment should be repeated with more distinct stimuli and less repetition.

Also, similarly to Experiment 3, there were main effects of congruency in the P600 time window. Specifically, as expected, incongruent endings were associated with stronger posterior positivities in this time window. In contrast to Experiment 3, however, these effects were more long-lasting (47 ms as opposed to 27 ms, Tables 5.3.1 vs. 4.3.1). This is likely just noise, but it is important to consider what it might mean if the changing talker paradigm reliably elicited longer-lived P600 effects than an environmental sound paradigm. As explained in chapter 4, according to some authors the P600 reflects syntactic (as opposed to semantic) processes. As we might expect little to no syntactic processing for environmental sounds, the longer P600 window in the present experiment could reflect longer-lasting or heightened syntactic processing taking place when all stimuli are verbal. As stated in chapter 4, however, if the P600 does indeed reflect syntactic processing, it is curious that it occurs at all for environmental sounds. One possibility that as far as we know has not yet been proposed is that the P600 potential, as measured at the scalp, actually reflects a sum of composite processes, some of which are syntactic and some of which have more to do with repair or integration difficulty. A final possibility is that the P600 reflects detection of a pattern or expectation violation that is especially sensitive to syntactic violations, but also responds to non-syntactic violations in the absence of a syntactic structure.

Finally, we observed widespread differences across the whole timecourse of the ERP based on target type. The precise differences in topography did not always match the differences seen in Experiment 3. For example, the acoustic change complex responsible for the early part of the long target type significance window in Experiment 3 was not present in Experiment 4, even though there were significant differences between talker in these same time windows. In general, however,

this finding supports the interpretation that widespread and long-lasting ERP effects can be due to acoustic and/or source differences, and need not necessarily be a result of obligatory differences between ES and verbal processing.

In summary, there were important similarities between responses to ES and to a different talker at the end of the sentence, including presence of congruency-sensitive N400s and P600s that do not interact with target type, in addition to widespread main effects of target type that persist across nearly the whole time period analyzed. Some of the differences, such as the short N400 time window and the lack of constraint effects in Experiment 4, can be explained by the stimulus set and the resulting high amount of repetition. Ideally, future experiments should include a larger stimulus set and less repetition to avoid this pitfall.

There were also obvious differences between responses to ES and words spoken by a different talker. The most obvious of these is that the salient ACC that occurs in response to ES at the end of spoken sentences was not apparent in response to a changing talker. This may have occurred because spoken words produced by a new talker are not conceptually or acoustically different enough from the preceding spoken sentence frame to elicit this response. However, it is important to note that this does not necessarily indicate different processing mechanisms for ES and words, as words occurring after silence almost always elicit an ACC, as do changes in extended vowels (Kim, 2015), so this event-related potential is not specific to non-language. Therefore, our data still support the interpretation that ES and words are understood in sentence context via largely similar mechanisms, and that differences in the brain responses to these two types of stimuli are largely resultant from acoustic differences and the attentional effects that those differences engender. Future experiments might further probe this question by using speech embedded in noise

(which might be designed to match spectral properties of matched environmental sounds) as stimuli.

5.4.2 Discussion – follow-up to previous talker normalization work

Before this experiment, neural signatures of talker normalization had never been shown in whole, fluent sentences where participants are not required to make behavioral responses. Based on previous literature, we expected there to be effects of talker as early as the N1 and P2, and potentially earlier. Moreover, we expected these effects to be mediated by working memory (WM).

Our expectations were confirmed by significant differences between same and different talker in the time windows normally containing the N1 and P2 (124 – 147 and 235 – 278 ms post target onset, respectively; Table 5.1). This indicates that the process of adapting to a different talker begins as early as the N1, in agreement with previous literature (Kaganovich et al., 2006; Zhang, Peng, & Wang, 2013). However, when the scalp topographies are examined more closely, it becomes clear that these are not typical N1 and P2 topographies; they consist of central positivities that are left-skewed (124 – 147 ms) or shifted frontally (235 – 278 ms) (Figure 5.4 a and d). As can be clearly seen in Figure 5.4d, neither a canonical N1 nor P2 is readily visible in the voltage traces. This presents some difficulty in reconciling the present results with previous results, which often present stimuli (e.g, words or vowels) after a brief period of silence (e.g., Kaganovich et al., 2006; Zhang, Peng, & Wang, 2013), thus ensuring the presence of a canonical auditory N1-P2. Nevertheless, the presence of main effects of talker in these time windows does suggest that talker normalization-related differences are present in our paradigm as early as 124 ms after target onset.

As stated above, we expected the talker-related differences to be mediated by WM. Surprisingly, our randomization analyses including WM as a between-subject factor (d' on a 3-back task) did not identify any significant interactions between WM and talker (Table 5.5). There

were, however, several windows in which there were significant main effects of WM (Table 5.5), some of which overlapped with windows including main effects of talker. Notably, there are main effects of WM in both N1 and P2 windows, as well as even earlier. This suggests that extremely early after hearing the last word in a sentence, people are already processing what they hear differently based on their WM capacity.

Why might WM be affecting ERPs so early after word onset? One explanation is simply that as people are faced with the task of making sense of the sentence as a whole, they maintain the sentence stem in WM as they are listening for the ending word; because this process requires WM, we find differences associated with individual working memory capacity. However, previous experiments examining correlations between WM capacity and ERPs have implicated slower, later potentials (post-300 ms) as time windows where ERPs differ between high and low WM participants (King & Kutas, 1995; Nittono, Nageishi, Nakajima, & Ullsperger, 1999). In light of this information, a more likely explanation is that participants are managing their attentional focus in order to glean the most relevant information from the signal they are hearing, and that this focusing of attention requires WM (cf. Heald & Nusbaum, 2014). This corresponds well with previous research showing that reallocation of attention affects ERPs as early as the N1 and P2 (e.g., Picton & Hillyard, 1974), and that attentional processes that affect early sensory processing also interact with the maintenance of the perceived information in WM (Awh & Jonides, 2001). Furthermore, it has also been shown that WM capacity predicts N1 and P2 amplitudes in tasks requiring auditory selective attention (Giuliano, Karns, Neville, and Hillyard, 2014). Presumably, then, people with higher WM capacity are better able to focus their attention on relevant cues, which leads to the main effects of WM that are observed in the present experiment as participants listen to words. Finally, it has also been suggested that WM may not affect the deployment of

attention, but rather that it is the other way around: attention acts as a “gatekeeper”, focusing on the aspects of the input that are going to be retained in WM (cf. Awh, Vogel, & Oh, 2006). Interestingly, the present results appear to refute this idea; if the deployment of attention is not influenced by WM, and WM capacity traditionally affects post-300 ms potentials, we would likely not see main effects of WM on the earlier N1 and P2 windows, which are known to be affected by attention. Thus, it seems most likely that WM is mediating attentional mechanisms which, in turn, are influencing early potentials.

While this explanation makes sense with regards to the main effects of WM we found, it does not explain why there are no interactions between talker and WM. If adapting to a different talker taxes WM as suggested by previous research (e.g., Nusbaum & Morin, 1992), then we would expect to see stronger WM effects in the different talker condition. There are a couple potential explanations for this finding. One is that there is something about our paradigm – for example, listening to completely continuous sentences rather than isolated words, or having participants listen passively rather than responding to “target” stimuli – that narrows the gap between the same talker and a different talker. Perhaps the task of passively listening to and understanding sentences is easier than monitoring a stream of stimuli for a particular target, and thus reduces the overall difficulty (and therefore the necessity of recruiting WM) of adapting to a different talker. However, this explanation is difficult to reconcile with the findings of widespread main effects of talker, including during the N1 and P2 time windows (Table 5.1). Clearly the processing of the two talkers is meaningfully different, even at early time points. This could happen if the same WM-dependent processes are mediating the processing of words said by both talkers, but to a greater extent for the different talker. In order to better address this question, we performed exploratory source analyses

on time windows that showed main effects of talker and of WM, in which we tested models based on parietal, temporal, or both sets of coordinates identified by Wong et al. (2004).

Although there were early effects of talker and WM, the source analyses did not indicate any differences in the fit of the models to same vs. different talker data in any window before the N1. While it is possible that there are talker-related, attention-related, and/or WM-related effects in these time periods that come from different sources in the brain, it is worth noting that previous work suggests that attentional effects on stimuli processing do not begin until later (e.g., Picton & Hillyard, 1974; Vogel, Luck, & Shapiro, 1998).

In the N1 time window, we observed that there was a significant correlation between the fit of the parietal model and WM. Specifically, as WM increased, the parietal model explained the different talker topography increasingly better than the same talker topography. This was not observed for the temporal sources model, nor in other time windows (Figure 5.6). This result explains why we failed to find differences in fit of the parietal model between same and different talker topographies in the N1 time window (Table 5.7): differences will not be apparent unless enough of the participants have WM on the high or low end of the distribution to make the effect size large enough to detect statistically. This result has many implications in light of previous talker normalization work. First of all, it serves as further confirmation that these areas indicate attentional processes as theorized by Wong et al. As suggested by previous work already described in this Discussion, the N1 time window is thought to be an important locus for attentional effects, and thus the coincidence of parietal effects with this time window further confirm that attentional modulation via superior parietal areas is happening when the talker changes. Furthermore, it is important that these parietal areas only explain the different talker data better than the same talker data in high-WM participants. This finding strongly supports the idea that WM is used for the

refocusing of attention to relevant features of the stimulus. It appears that only participants with high enough WM recruit these superior parietal areas to a greater extent in the changing talker condition, perhaps because they are better able to quickly direct their attention to acoustic features that differentiate the two talkers. For the participants with lowest WM, this pattern is actually reversed; the parietal model explains the same talker responses better than the responses to the changing talker. This suggests that participants with low WM are perfectly capable of recruiting attentional mechanisms via parietal areas, but they may not be able to focus on the stimulus features of the new talker's speech. In other words, they may be recruiting their attention but failing to direct it to the cues that distinguish the new talker from the old one (cf. Heald & Nusbaum, 2014b).

Finally, in the P2 time window, the bilateral superior temporal areas identified by Wong et al. (Table 5.7) explained the scalp topography for the different talker condition significantly better than the same talker condition. Unlike the parietal effects in the N1 window, this effect did not correlate significantly with WM. This is interesting for several reasons. First, it provides some further temporal separation to the results from Wong et al. Whereas because of the relatively poor time resolution of fMRI it is difficult to say whether the parietal or temporal sources are active sequentially or simultaneously, the present experiment suggests that in adapting to a new talker, the superior parietal regions are active first, in a WM-dependent manner, and the superior temporal areas come online later. Furthermore, it is interesting that the fit of parietal models is WM-dependent, but this is not the case for the models based on superior temporal sources. Wong et al. note that these temporal areas have been implicated in speech recognition. Taken together with the results from the parietal sources in the N1 window, then, this suggests that adapting to a new talker

first requires a WM-dependent re-allocation of attention, after which more speech-specific mechanisms take over.

In summary, our analyses confirm Wong et al. (2004) in that they implicate parietal and temporal sources in talker normalization. They extend the previous work by indicating that in processing a change in talker, parietal sources are active in a WM-dependent manner in the N1 time window, and that speech-related temporal sources come online later in the P2 time window, in a fashion that is not recognizably WM-dependent. These results potentially indicate a processing stream that first includes WM recruitment to reallocate attention, followed by speech-specific mechanisms that are less WM-dependent. These results pave the way for future work to further examine the timecourse of attentional and WM effects on talker normalization. As the source analyses were exploratory, they should not be taken as conclusive evidence, but rather as a strong suggestion that WM-mediated attentional effects are at play even in continuous, full sentences when there is a change in talker. We believe that fruitful avenues for further study could include 1) replicating Wong et al. (2004) with high density EEG, and 2) further examining talker normalization in full sentences with a broader set of stimuli, and possibly by combining fMRI and EEG.

5.4.3 Discussion – WM and talker-normalization-related sources in the processing of ES

The results of the exploratory analyses strongly suggest that, for sentences in which the talker changes, working memory is utilized in the reallocation of attention in order to understand the words said by the new talker. This feature of WM – its use in the reallocation of attention – is thought to be a domain-general resource; not a speech-specific WM role (Engle, 2010; Giuliano et al., 2014). However, it is possible that, due to the specialization of language, adaptation to acoustic change is dealt with differently for speech than for other auditory stimuli. If this is the

case, we might expect that understanding an ES in sentence context might draw on different resources, rather than WM-mediated attentional reallocation. On the other hand, if language relies on cognitive-general mechanisms common to other types of auditory processing as well, we might expect a transition from speech to ES to involve these same mechanisms.

From our analyses including WM as a covariate and performing source analysis for Experiment 3, we can draw several conclusions. The first, and simplest conclusion is that WM is indeed important for processing ES in sentence context, particularly in time windows around the N1 and P2, as shown by the significant interactions between WM and target type (Table 5.8). Though the patterns are not identical, the importance of WM in these time windows is shared with the talker change experiment. The second conclusion we can draw is that both parietal and temporal “speech-related” sources (as described by Wong et al., 2004) are differentially active between ES and words in these WM-related time windows, such that activity in these areas better describes the ERPs in response to ES. This suggests that these areas are active in mediating the transition to ES from speech, and that this activity is related to WM. We can therefore answer “yes” to the first part of our question “is WM recruited to process ES in sentence context in a way similar to its recruitment for spoken words?” WM does indeed appear to be recruited to process ES in sentence context.

The second part of this question requires a bit more explanation, as there are some differences between the results for Experiments 3 and 4 that make it less than obvious whether WM is similarly involved in both cases. I will address these differences one at a time. First of all, there is the issue of WM being involved as a main effect in the talker change experiment, but as an interaction with target type in the ES vs. words experiment. One way this could happen is if such patterns are present for all stimuli, but the difference in WM recruitment is greater between

ES vs. words than between the same vs. different talker. In this case, the greater WM recruitment for ES would dwarf the small effect for spoken words, leading to a weak, distributed topography as we see in Figure 5.7. In contrast, in the talker change experiment, the difference between the same and different talker is likely too small to be picked up as an interaction in an omnibus analysis, because the WM-related processes are so similar for the two stimulus types, so one topography does not overwhelm the other. Therefore, WM shows up as a main effect in this experiment.

These findings fit well with Heald & Nusbaum's (2014a) view of speech as an active cognitive process, in which attention and listening goals continuously guide processing of the acoustic input. According to this view, processing speech even from a single talker involves WM-dependent mechanisms as part of normal active processing. While a change in talker might recruit WM to a slightly greater extent, it makes sense that WM is recruited for both the same and different talker, explaining the main effect of WM. In contrast, coping with the large increase in variability and/or ambiguity accompanying a switch to ES from speech might place a much larger load on WM, accounting for the large gap between speech and ES, and leading to a statistical interaction. This might happen for a few reasons. For example, as argued by Nusbaum & Schwab (1986) if there is increased variability (e.g., a change in talker, or a change from speech to ES) there is greater uncertainty about identifying the sound that has been heard. This is because the listener is already familiar with the speech patterns of the first talker or first source, so the meaning of the acoustic input is easier to interpret. However, changing the talker or the source could lead to temporary ambiguity about the input, which could suddenly have a number of different meanings, leading to a many-to-many mapping problem. Because these alternative interpretations of the input have to be stored in WM before one is eventually chosen, a change in source increases the WM

load. In this case, it is easy to see how a change to ES from speech would involve a larger WM load than a change between two talkers; while there is indeed quite a bit of variability between talkers, ES vary considerably more than speech along many acoustic parameters, leading to greater uncertainty. Another possibility, which is not mutually exclusive with that already described, is that when a change in acoustic source happens, the listener must reorient their attention to certain relevant aspects of the stimulus in order to best make sense of the input. WM is thought to be recruited in this attention-reallocation process (Engel, 2010), as is supported by the results in the present experiment. Importantly, according to these explanations, there are not qualitatively different processes happening for ES and words. Instead, WM-dependent processes happen for both speech and non-speech, just to a greater extent for non-speech.

There is also the question of the slightly different source localization data. In the talker-change experiment, in the N1 window, there an effect where parietal sources better explain the responses to a different talker for high-WM participants but not low-WM participants. Then, in the P2 window, temporal sources explain the different-talker responses better than the same-talker responses. On the other hand, for the experiment with ES, the data are not so complex; in both the N1 and P2 windows, models including both parietal and temporal sources explain the responses to ES better than the responses to speech. Regarding the parietal-then-temporal sequence found in the talker change experiment, as opposed to the simultaneous temporal-and-parietal activity found in the ES experiment: rather than indicating qualitatively different processes, these results likely indicate similar processes that have slightly different timecourses. Whereas in the talker change experiment, differences in parietal activity appeared to be restricted to the N1 time window, and differences in temporal activity appeared in the P2 window, in the ES experiment both pairs of sources were implicated in both time windows. One reason this may have happened is that the

lengths of the ES were more variable and did not closely match their corresponding words. Therefore, some temporal “smearing” may have occurred when ERPs to the various stimuli were averaged together. Another, perhaps more interesting reason, is that perhaps the greater recruitment of WM necessary to comprehend an ES in sentence context involves longer-lasting activity in these brain regions, causing them to overlap temporally.

Finally, there is the question of the individual-difference-related effect reversal in the N1 time window which occurs for only the talker change experiment. Why might the regions implicated by Wong et al. (2004) better explain the response to a changing talker only for high-WM participants, whereas in Experiment 3, they explain the response to ES better for all participants regardless of WM? One plausible explanation is that ES are so obviously different from speech that attention is successfully captured and redirected in response to ES even for low-WM participants. In contrast, when the stimulus involves a change in talker, low-WM participants may have more difficulty with reorienting their attention to relevant aspects of the stimulus because it so closely resembles the preceding speech. Though not direct proof, this theory is supported by the presence of a clear acoustic change complex (ACC) in response to non-speech at the end of a sentence; this ACC was absent from the ERP to a different talker. As discussed in Chapter 4, the ACC is an index of attention capture and can mark a change in perception of the source of an auditory stimulus. This suggests that participants immediately registered a change in source for ES but not for words said by a different talker. This lends credence to the idea that the difference in source was not as obvious for the talker change experiment, and therefore attention reallocation may not have proceeded as cleanly, particularly for low-WM participants.

Importantly, this explanation again points towards a quantitative, not qualitative, difference between the understanding of ES and words in sentence context. At least from our source

localization work, it appears that words and ES draw on the same neural resources for processing in the context of a sentence, though the degree to which these mechanisms are active may vary depending on how distinct the stimuli are from the preceding sentence. This explanation—that differences between words and ES are largely differences of degree in WM recruitment and attention reallocation—fit well with the 60 ms processing lag between ES and words found in Experiment 2, which, as explained in Chapter 3, is approximately the delay expected from an attention reallocation process. This idea – that the differences between ES and words involve differences in the extent of activation, rather than distinct processing streams – is incompatible with a modular or encapsulated account of language processing. According to these accounts, the same mechanisms should not be responsible for understanding ES and speech. Transitioning from speech to ES in a single sentence should be extremely difficult, and should not involve simply a difference in the extent of recruitment of cognitive-general mechanisms like WM. Therefore, in order to account for the present findings, we either need to make alterations to existing models of speech perception, or we need to create new models that can account for WM-mediated understanding of a wide range of auditory stimuli (not just speech).

Chapter 6: General Conclusions

6.1 Purpose of work

The broad goal of this work was to address the debate about specialized versus cognitive-general mechanisms for understanding language. As outlined in Chapter 1, there exists a wide range of theories explaining how humans process and understand language. Some of these hold that human language capabilities depend on highly specialized mechanisms that evolved to handle only language which is required due to specific computational complexities of language (e.g., Chomsky, 1986; Liberman et al., 1967), while others take the perspective that language processing is carried out by flexible, cognitive-general mechanisms (e.g., Leech et al., 2009; Potter et al., 1986).

According to the former view, language processing relies on largely different neural systems than the processing and understanding of other meaningful auditory and visual stimuli. In the most extreme versions of these theories, understanding spoken or written language relies on specialized, language-only “modules” that do not interact with other processing pathways. Arguments in favor of such models often point out the seemingly innate predisposition of human infants to learning language, using this observation to argue for the existence of specialized neural language circuitry at birth, rather than for the motivating power of behavioral relevance (e.g., Dehaene-Lambertz et al., 2010; Dehaene-Lambertz et al., 2006).

According to the alternative view, while language abilities may appear highly specialized, they are not qualitatively different from other kinds of auditory processing. Proponents point to lesion studies which find that language-system lesions often affect auditory processing of environmental sounds as well (Dick et al., 2001; Saygin, 2003). In this view, the language system appears highly specialized because the average person (i.e., the average study participant) has

been gaining expertise with language since childhood. This position is supported by data showing that comparable expert-level auditory processing skills (e.g., music) seem to overlap with canonical “language” networks in participants with extensive experience (e.g., professional musicians; Dick et al., 2011; Donnay et al., 2014; Leech et al., 2009).

As we reasoned in Chapter 1, if language understanding depends on highly specialized mechanisms, making sense of a non-language stimulus (e.g., environmental sound) embedded in language should be difficult, or at the very least require extra processing to translate it into a language format that the language system can accept. However, the studies in this dissertation argue against this view, as they find both behavioral and neural evidence that words and ES are understood in sentence context in remarkably similar ways. Both behavioral results (recognition points; response times) and ERP (N400, P600) results suggest that ES and spoken words interact with sentence context, including contextual constraint, in much the same way. Moreover, results from source analyses, as well as analyses including WM as a covariate, suggest that the same cognitive-general mechanisms underlie the ability to quickly process both ES and words in sentence context. These results support the idea that speech perception, while undeniably a specialized ability, nevertheless relies on cognitive-general underpinnings, such as attention, WM, and top-down effects on processing (in the case of language, this can include contextual constraint). Therefore, it should be possible to explain auditory understanding of speech, ES, and potentially other types of meaningful auditory stimuli (e.g., music) using a single, broad framework. Current theories explaining how speech is understood are almost never discussed broadly enough to extend to environmental sounds or other types of auditory stimuli. Therefore, there is a need to either modify existing models, or develop new ones, in order to explain the current results.

6.2 Explaining the present results: developing a new framework that can incorporate ES and words

The results from the present studies point towards a speech understanding model that accounts for a large overlap with the neural circuitry required for understanding environmental sounds. Indeed, the WM data and source analysis results suggest that, rather than involving vastly different processing mechanisms, switching between ES and speech merely involves a different degree and/or time course of recruitment of cognitive-general resources such as WM, along with posterior temporal auditory processing areas (section 5.4.3).

There are many different possibilities for how such a model might work, and what types of processing steps it might involve. However, as an example, let us take the speech understanding model outlined by Friederici (2011). This four-stage model includes: 1) acoustic-phonological analysis, in which basic acoustic features of the signal are recognized and characterized, 2) initial syntactic processes in which a grammatical structure for the sentence is built, 3) computation of semantic and syntactic relations (this stage involves N400-related processes), and 4) integration of semantic and syntactic information or “interpretation”. For each of these stages, Friederici describes brain areas that have been implicated in these processes using electrophysiology and neuroimaging research. This model can serve as an example for a theoretical framework upon which to build a model that integrates auditory speech and non-speech understanding.

6.2.1 Accomplishing early acoustic signal processing without phonemes

Some of the stages in Friederici’s model (and their underlying neural substrates) are already well-situated to be extended to the processing of environmental sounds. For example, stage 1, involving acoustic-phonological analysis and temporally corresponding to the N1 ERP, could be generalized to involve simply acoustic analysis (which may or may not involve phonemes when

the stimuli are verbal). This is not much of a stretch, given that neuroimaging studies show that many of the areas involved in this processing stage are generally not speech-selective, but rather are activated by many different types of sounds (e.g., Heschl's gyrus, Mummery, Ashburner, Scott, & Wise, 1999; a region of the STG lateral to HG, Hall et al., 2002). In terms of the present data, results from our experiments suggest WM involvement in the time window corresponding to this stage of processing, with likely activity in parietal areas implicated in attention allocation. Importantly, this WM involvement and recruitment of parietal areas was seen for both ES and words in sentence context. Therefore, an addition to stage 1 of this model could include attentional modulation of the acoustic analysis mediated by the parietal areas described by Wong et al. (2004).

It should be noted that a reliance on acoustic features rather than phonemes *per se*, as described above, presents a departure from many extant models of word recognition and understanding. For example, in the TRACE model (McClelland & Elman, 1986) as well as in updates such as jTRACE (Strauss et al., 2007), there are three layers of “nodes” (i.e., theoretical processing units) that correspond to information critical to word recognition. The first layer is the “feature” layer, which represents different acoustic features of speech (e.g., acuteness or diffuseness of a speech sound). The second and third layers are phoneme and word detector layers, respectively. Excitatory connections run between nodes in different layers that are consistent with each other; for example, an excitatory connection might exist between the /p/ phoneme and the word “pen”. Inhibitory connections run between mutually exclusive items within the same layer. In this way, different feature nodes combine their activation together to specify and activate an individual phoneme, and different phonemes come together to activate a word, while suppressing phonemes that are not present in the signal. As we can see from this model, a crucial part of the computational framework for word recognition is a layer of

phonemes, specifically as part of a “feature → phoneme → word” architecture. TRACE is not unique in this respect; other models (e.g., Shortlist (Norris, 1994; Norris & McQueen, 2008), COHORT (Marslen-Wilson, 1987)) include either explicit phoneme layers, or initial processes that are highly dependent on phoneme identity. In the context of a cognitive-general model which might also explain recognition and understanding of non-speech such as environmental sounds, it is problematic that this “feature-phoneme-word” architecture has no clear parallel. For environmental sounds, there are no common constituent “phonemes”, although it might be possible to identify a common set of acoustic features (e.g., frequency/frequencies, speed of onset, rate of frequency change) that could be important in a computational model of non-speech recognition. However, these acoustic properties are not phonemes, and indeed TRACE differentiates acoustic features from phonemes by including them in their own separate layer. Therefore, many existing models of word recognition cannot be directly extended to non-speech sounds without redesigning their entire computational framework.

It is also worth noting that the results of our talker change experiment (Experiment 4) argue against models such as TRACE in which acoustic features quickly give way to processing stages involving standardized, language-specific units such as phonemes. Such a model would predict that, neurally, there should not be major differences between responses to different talkers beyond early stages of the ERP where acoustic differences might be captured. As soon as phonemes are extracted from the acoustic signal, the processing should be exactly the same. Instead, as shown in Experiment 4, there are widespread differences in the raw voltage extending all the way to 800 or 900 ms post-onset, though the N400s (as shown by the difference between matches and mismatches) are statistically identical. These results closely mirror those of Experiment 3, in which raw voltage differences were widespread, but N400 [mismatch – match] difference waves were

identical. It is difficult to reconcile such results with models in which all words are processed via a phoneme-dependent pathway, while all ES are processed through an entirely different system.

However, there are some models that instead rely more on acoustic features rather than phonemes, which would be a more promising extension to environmental sounds. These models come out of a long history of questioning the necessity of phonemes in explaining speech perception. For example, in the LAFS model, a computational model of speech perception by Klatt (1979), the acoustic speech signal proceeds through a filter bank before feeding directly into a lexical access process. There is no intervening phonetic stage or “phoneme recognition layer”. Similarly, others have questioned whether phonemes are a meaningful perceptual unit, and proposed phonetic segments or diphones as alternatives (see Pisoni & Luce, 1987, for a review) although like phonemes, these units are difficult to extend to environmental sounds without some fundamental reconceptualization. Therefore, in order to extend existing speech perception models to allow the rapid processing of environmental sounds, perhaps models like Klatt’s, in which acoustic features and not phonemes are important for recognition, should be used as a basis.

Of course, while environmental sounds cannot consist of phonemes, an alternative is that they do consist of meaningful sound categories that assist in perception and understanding. As an example, Warren & Verbrugge (1984) noted that listeners could distinguish between glass bouncing and breaking, i.e. they could categorize environmental sounds based, presumably, on common acoustic properties that conveyed this meaning. If such results do indeed reflect the existence of individual meaning-conveying acoustic features, we might imagine an update to speech perception models where phonemes are merely a special case of such “auditory meaning categories,” which in reality exist, in some form, for all meaningful sounds that we understand. In

such a case, “acoustic-phonetic analysis” in Friederici’s model might be replaced with a more general form of “acoustic pattern” or “acoustic categorization” analysis.

6.2.2 Syntactic analysis and its meaning in a framework that accommodates ES

After acoustic analysis in Friederici’s model comes stage 2: syntactic analysis. This is the only stage that does not conceptually relate to ES easily. It is also the only stage that is not thoroughly addressed by the experiments in the present work, as we did not manipulate syntax and indeed, kept the sentence frames constant between speech and non-speech. However, it is still possible to think of modifications that might allow this stage to apply to broader auditory processing beyond speech.

Stage 2 involves analysis of the grammatical structure of the sentence. While forward-predictive syntax can easily be imagined for sequences of environmental sounds (e.g., doorbell → dog barking → footsteps → door opening), it is more difficult to think of analogs to more complex syntactical structures in which the sequence of words does not match the sequence of events it is describing due to reverse associations in the syntax (e.g., “the evidence the lawyer examined...”).¹ Therefore, stage 2 of Friederici’s model is the stage most likely to require modification in order to accommodate the easy processing of ES in sentence frames.

There are several possibilities for such modifications. Most simply, stage 2 may not be a necessary processing stage, as Friederici acknowledges; therefore, it may be possible to eliminate

¹ It is important to note that, while it may be difficult to think of examples where reverse associations could apply to ES, they are not a unique feature of language. As an example, take the German 6th chord. In Western classical and romantic harmony, this chord has a pre-dominant function, resolving to the dominant (V chord). However, the intervals of the German 6th are, in fact, identical to those of a dominant seventh chord. In this way, a German 6th in C major/minor can also function as a dominant seventh in D-flat major or minor. Many interesting chord progressions revolve around this dual interpretation of the German 6th as a dominant seventh. It is not until the resolution to the next chord that it is interpretable, in retrospect, as a German 6th or a dominant seventh, thereby retroactively affecting the perception of the resultant tonal center.

it from an ES-and-speech model entirely. Another alternative is that the areas mediating syntax processing (e.g., anterior STG) are keeping track of complex structures, which can involve syntax for language, but more generally involves complex statistical structures for other types of auditory stimuli. There is some support for this theory from animal studies; the anterior STG is involved in the encoding of complex auditory sequences in non-human primates (Rauschecker and Scott, 2009). In addition, as already mentioned, while Broca's area is sensitive to syntax in speech, it has also been found to be involved in musical syntax (Maess et al., 2001). Thus, perhaps "syntax-sensitive" neural mechanisms are more domain-general than previously thought, and stage 2 of Friederici's model could be reconceptualized as "structural analysis" or "sequence analysis".

A final possibility is that syntactic mechanisms are only sometimes recruited when they can be used to make sense of an otherwise ambiguous stimulus. This idea is supported by Friederici's observation that, in native speakers of a language, syntactically straightforward sentences elicit very little activation in "syntax areas" during time windows corresponding to stage 2 of the model. Perhaps very little activity in syntax networks is necessary to process relatively simple syntactic structures. This idea of occasionally-recruited syntax mechanisms is also supported by our current finding that there is no P600 interaction between congruency and target type. This is important because, as discussed in section 4.4, large P600s are associated with syntactic violations. Because environmental sounds cannot fit into a syntactic structure, one might expect that even ES congruent with the sentence frame would elicit large P600s, because they are incapable of filling a syntactic role. This, however, is not what we find; instead, ES follow the same pattern as words, with stronger P600s in response to semantic violations, and weak P600s in response to congruent endings. In this case, it seems as though the system is capable of ignoring syntactic rules when it determines that they are irrelevant for aiding understanding – in this case,

when the sentence ending is not a word. These findings support an account in which stage 2 of Friederici's model is reserved for syntactically difficult cases, and is not an obligatory feature of all auditory processing or even all language processing.

6.2.3 Semantic processing in speech and ES understanding, and the role of context

Stage 3 of Friederici's model (semantic and syntactic relations) is among the easiest to extend to ES understanding, not least because many of the implicated brain areas have multimodal associations. As an example, lesions to areas in the anterior temporal cortex thought to be involved in constructing meaning from a sentence can cause deficiencies in picture processing and memory (Friederici, 2011). Similarly, areas in posterior STG thought to mediate integration processes in speech have also been implicated in audiovisual integration and face processing (Amedi et al., 2005; Haxby et al., 2000).

The present experiments are particularly rich in evidence that the processes underlying semantic analysis are similar for ES and words. For example, determining whether a sentence is “understandable” or “nonsense” (in Experiment 2) relies on a semantic understanding of the sentence and the ending item, which could be either an ES or a word. Response times were almost the same for ES and words in Experiment 2: the marginally significant processing cost of only ~60 ms was consistent with attention reallocation similar to that found for a changing talker (Nusbaum & Morin, 1992; Heald & Nusbaum, 2014b; Wong et al, 2004), or for changing octaves or timbres in music perception (Van Hedger et al., 2015), and was certainly not consistent with covert naming. This finding suggests that the computation of semantic relations proceeds at the same speed for ES and words, and does not require “translation” of an ES into its corresponding word.

Friederici identifies the N400 as the main ERP associated with stage 3 of the processing stream as an index of the computation of semantic relations. The N400 results in the present studies

suggest that this stage progresses in the same way for ES and speech. In Experiment 3, we compared the N400 to sentences ending in either ES or speech. We found statistically indistinguishable [mismatch – match] N400s to ES and words, which suggests that the processes and brain areas that underlie the construction of meaning from a sentence or sequence of auditory stimuli are either general enough to encompass both ES and word processing, or are at least flexible enough to handle both types of stimuli. Given the multimodality of the brain areas Friederici identifies as involved in semantic analysis (as discussed above), it seems plausible that these areas could be carrying out semantic analysis for ES as well as spoken words.

There is, however, one aspect of Friederici's model which needs to be better elucidated to explain the current results: in its current form, the model does not fully explain context effects, including constraint effects. Effects of constraint on word recognition and understanding are well-documented, and indicate close interplay between semantic information and the understanding of new words in the unfolding sentence (see sections 1.3.1 and 1.3.2). The results of Experiments 1 and 2 agreed with these previous studies and extended the findings to environmental sounds: high constraint reduced recognition points and response times for ES and spoken words equally, showing that the processes for understanding ES and words are equally affected by constraint. Therefore, a new model should be able to incorporate linguistic context effects for both words and ES.

In Friederici's model as-is, semantic context mainly affects processing at late stages after semantic analysis, and might not even be important unless the acoustic signal is ambiguous (in which case an extra context-dependent step would take place). In such a view, it is difficult to explain why the same stimulus might be responded to more slowly in a less constraining context (as we found in the present Experiment 2), when the spoken words we used were not masked by

noise or otherwise rendered difficult to understand. If clearly spoken English words heard by native speakers are acoustically ambiguous enough to trigger the context-dependent step, then it might make more sense to think of context-dependent processing as a normal feature of speech understanding, and all speech signals as acoustically ambiguous to some degree. Friederici's view is also in conflict with the finding—in Experiment 3 as well as in previous literature—that the N400 is sensitive to constraint (e.g., Kutas & Hillyard, 1984). If it is true that the N400 indexes semantic analysis, then it stands to reason that the constraint-sensitivity of the N400 indicates some involvement of context in semantic analysis. This goes against the idea of contextual information becoming important only later in processing.

It remains unclear how context affects speech understanding, although there are different theories distinct from Friederici's. As described in Chapter 1, predictive coding frameworks, like that of Lupyan and Clark (2015) (Figure 1.1) theorize that context creates top-down neural predictions for incoming signals at different levels. These predictions can then be used to facilitate processing and to compute prediction errors, the latter of which can assist in adaptation (i.e., learning). While attempts to explain speech perception in light of predictive coding are relatively new, older models have also incorporated context effects. As described in section 3.4, the TRACE model includes bidirectional connections between layers, easily allowing for top-down modification of processing. In order to explain context effects on speech perception, one could modify this model by adding a fourth, highest-level layer including concepts and semantic associations, which can modify word perception in a top-down fashion (as described in section 3.4).

While the TRACE model cannot extend to ES due to its obligatory reliance on phonemes (see section 6.2.1), predictive coding models such as Lupyan and Clark's might be combined with

Friederici's framework for a more complete picture of the semantic analysis stage of auditory understanding. For example, perhaps Stage 3 could include processing modifications that are dependent on conceptual-level top-down predictions based on the listener's current understanding of the semantic content of the sentence. In this way, semantic processing in the areas suggested by Friederici (e.g., posterior STG) could be modified from sentence to sentence based on the associated concepts evoked by the preceding verbal context. There is precedence for this idea; for example, an fMRI study of image perception suggested that verbal context can quickly and dynamically alter neural representations of individual images (Çukur et al., 2013).

6.2.4 Interpretation of the meaning of a sentence as a whole, even when it includes ES

Like stages 1 and 3 of processing, stage 4 of Friederici's model is also easily extensible to ES understanding. Friederici identifies both Broca's area and surrounding parts of the left IFG, along with left STG, as the main candidates for involvement in Stage 4, "interpretation". Importantly, Broca's area and the surrounding areas in IFG, while often described as "speech" areas (e.g, Santi & Grodzinsky, 2007), have been implicated in much more cognitive-general processes, including patterns in musical sequences (Maess et al., 2001) and cognitive-general working memory (e.g., Ranganath, Johnson, and D'Esposito, 2003). Similarly, left STG (indeed, STG bilaterally) has been shown to be responsive not only to speech sounds, but to environmental sounds as well (Giraud and Price, 2001; Leech and Saygin, 2011; Lewis et al., 2004). Therefore, because of the multimodality of these brain areas, it does not seem out of the realm of possibility that ES or other meaningful non-speech could be processed by this network as the listener develops a general understanding of the meaning of the auditory input.

Friederici identifies the P600 as the main ERP associated with the interpretation stage. At first, this may appear to present a problem given that the P600 is often thought to index syntactic

processes (e.g., Osterhout & Holcomb, 1992), which Friederici places earlier in stage 2 of her model. However, regarding the P600, Friederici takes an updated view that this ERP reflects not simply syntactic processing, but the integration of semantic and syntactic information into a coherent whole that can then be understood. This view is consistent with our findings that semantic violations (whether they consist of spoken words or necessarily syntax-free ES) elicit larger P600s (Experiments 3 and 4). In this case, it is likely that integration of syntactic information with semantic does not need to take place for ES, because they do not form a particular part of speech, and therefore the P600 in the case of ES reflects simply semantic interpretation processes.

In terms of extending this interpretation stage to ES as well as words: if these two types of stimuli proceed through this processing stage in a similar fashion, we would expect the P600 to look largely similar for ES and words, and, most importantly, to interact with context in the same way. Indeed, this is what we find: Experiment 3 shows that the P600 does not exhibit an interaction between congruency and target type. Such an interaction would be expected if ES and words proceed differently through stage 4, as they should, in that case, be combined with the meaning of the preceding sentence frame in different ways. Therefore, our ERP results support a view in which ES and words proceed through later interpretation stages in the same way.

6.3 Final Conclusions

It would obviously be incorrect to state that speech processing is exactly the same as all other auditory processing. However, the present results call into question whether differences between speech and other types of sounds are fundamental features of speech. Rather, they suggest that differences between speech and other auditory skills might be a result of large differences in experience between speech and the other skills measured, which might, in turn, lead to differences in the time course or extent of recruitment of cognitive-general neural mechanisms.

There are, of course, other differences between speech and other types of auditory processing: like music, speech has both motor and perceptual components, as most people both produce and perceive speech regularly (in contrast to environmental sounds, most of which are not routinely directly produced by humans). This has been emphasized in motor theories, in which the motor speech production system plays an integral role in speech perception (e.g., Liberman & Mattingly, 1985). More recently, however, fMRI studies have shown that the extent of motor area recruitment during speech perception varies widely among individuals (Szenkovits, Peelle, Norris, & Davis, 2012). In light of such results, it seems reasonable that motor involvement might not be a necessary component of a broad auditory understanding model, although perhaps under certain conditions, or in certain listeners, motor activity can be used to help disambiguate an ambiguous signal, or to focus attention in particular ways. The present studies are not equipped to address these questions, as they do not examine speech production or motor brain areas. Perhaps in the future, non-speech examples of specialized sound production with heavily-practiced motor components—such as music or birdsong—will be a promising area in which to push the boundaries of motor theories, and examine the role of motor system involvement in auditory perception beyond speech.

The present experiments call for an account of auditory processing in which meaning is understood via largely cognitive-general mechanisms, involving temporal auditory processing regions, as well as depending heavily on WM. While such neural resources may be active to different degrees for different types of stimuli (e.g., a novel talker or an environmental sound), they appear to form a single, flexible network that can handle a wide variety of auditory stimuli, rather than distinct processing pathways. Of course, within this framework, there is still room for a multitude of variations of specific theories still consistent with this perspective. The present experiments cannot distinguish, for example, between the existence of a broad-reaching auditory

processing network that handles many types of sounds including language, and the existence of a language network that is flexible enough to handle other types of stimuli. The former is closer to the example of an expansion of Friederici's model, in which all the stages are specific examples of the operation of a broad auditory processing system. The latter could occur, still within a non-encapsulated framework if, for example, exposure to language since childhood shapes the development of pathways in the brain, making them uniquely adapted to processing language. In this case, language abilities could at first glance appear quite specialized. However, on closer examination, it would become clear that this specialization is a result of continued experience and plasticity, and moreover that these language networks are capable of handling non-speech sounds given enough experience. Some important predictions follow from such a theory. For example, the more experience someone has with a non-language auditory stimulus class, the more responses to these stimuli should start to look like neural responses to language. Indeed, this is supported by the substantial body of music expertise work described in Chapter 1. However, a couple pieces of evidence argue against even this degree of specialization. In an experiment where participants were trained to categorize computer noises that did not resemble speech, neuroimaging showed activation patterns for the computer noises become speech-like over the course of hours, rather than years as would be expected if such plasticity takes years of experience to develop (Leech et al., 2009). Similarly, given the extreme discrepancy in the extent of the average person's experience with environmental sounds versus language, our results show remarkable similarities between these two types of stimuli.

The truth is likely closer to some combination of the two in which a broad auditory processing network exists at birth, and is then honed over the course of the lifetime by experience with particular behaviorally-relevant classes of stimuli. This could lead to a limited degree of

specialization via a process in which certain small cortical regions develop a preference for particular classes of stimuli (e.g., speech, non-speech), while remaining 1) capable of responding to different classes of stimuli despite having a preferred type, and 2) nested inside the larger generalized architecture, making EEG results look quite similar due to spatial proximity. This type of organization is suggested by results from Leech and Saygin (2011), who find distributed preferential activation for words vs. ES in spatially close, but not identical, sub-regions of temporal cortex. Additional neuroimaging and electrophysiology studies are, of course, necessary to determine to what extent this is the case.

In summary, the present experiments show, with both behavioral and electrophysiological data, that processing costs for switching to ES from speech a) presents little to no difficulty, b) relies on remarkably similar neural mechanisms as processing spoken words in the same context, and c) involves cognitive-general mechanisms, such as WM, that are also recruited in speech-only tasks such as processing a novel talker or even processing naturally unfolding speech from the same talker (cf. Heald and Nusbaum, 2014a). While intriguing, these results are only the beginning of a potentially large body of future research that could probe the limits of this system with stimuli of varying recognizability or familiarity. Moreover, the present experiments test only ES in sentence context; logical next steps could include examining words in the context of an environmental sound auditory scene, or manipulating experience by exposing participants to a subset of environmental sounds repeatedly. There are endless possibilities for future behavioral, electrophysiological, and neuroimaging work that can further clarify the characteristics of a cognitive-general system for auditory understanding—a system that has been only briefly examined in the present work.

References

- Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models. *Journal of Memory and Language*, 38(4), 419–439. <https://doi.org/10.1006/jmla.1997.2558>
- Amedi, A., von Kriegstein, K., van Atteveldt, N.M., Beauchamp, M.S., Naumer, M.J. (2005). Functional imaging of human crossmodal identification and object recognition. *Exp Brain Res* 166: 559-571.
- Ardal, S., Donald, M. W., Meuter, R., Muldrew, S., & Luce, M. (1990). Brain responses to semantic incongruity in bilinguals. *Brain and Language*, 39(2), 187–205.
- Awh, E., Vogel, E. K., & Oh, S.-H. (2006). Interactions between attention and working memory. *Neuroscience*, 139(1), 201–208. <https://doi.org/10.1016/j.neuroscience.2005.08.023>
- Awh, Edward, & Jonides, J. (2001). Overlapping mechanisms of attention and spatial working memory. *Trends in Cognitive Sciences*, 5(3), 119–126. [https://doi.org/10.1016/S1364-6613\(00\)01593-X](https://doi.org/10.1016/S1364-6613(00)01593-X)
- Ballas, J. A., & Mullins, T. (1991). Effects of Context on the Identification of Everyday Sounds. *Human Performance*, 4(3), 199–219. https://doi.org/10.1207/s15327043hup0403_3
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using {lme4}. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Bates, E., & Dick, F. (2002). Language, gesture, and the developing brain. *Developmental Psychobiology*, 40(3), 293–310. <https://doi.org/10.1002/dev.10034>
- Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, 403(6767), 309–312. <https://doi.org/10.1038/35002078>
- Bennington, J. Y., & Polich, J. (1999). Comparison of P300 from passive and active tasks for auditory and visual stimuli. *International Journal of Psychophysiology*, 34(2), 171–177. [https://doi.org/10.1016/S0167-8760\(99\)00070-7](https://doi.org/10.1016/S0167-8760(99)00070-7)
- Bonhage, C. E., Mueller, J. L., Friederici, A. D., & Fiebach, C. J. (2015). Combined eye tracking and fMRI reveals neural basis of linguistic predictions during sentence comprehension. *Cortex*, 68, 33–47. <https://doi.org/10.1016/j.cortex.2015.04.011>
- Bornkessel-Schlesewsky, I., & Schlewsky, M. (2008). An alternative perspective on “semantic P600” effects in language comprehension. *Brain Research Reviews*, 59(1), 55–73. <https://doi.org/10.1016/j.brainresrev.2008.05.003>
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10(4), 433–436.

- Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and Affordances in Syntactic Ambiguity Resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(3), 687–696. <https://doi.org/10.1037/0278-7393.30.3.687>
- Chen, Y.-C., & Spence, C. (2011). Crossmodal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity. *Journal of Experimental Psychology: Human Perception and Performance*, 37(5), 1554–1568. <https://doi.org/10.1037/a0024329>
- Chomsky, N. (1969). *Aspects of the Theory of Syntax*. MIT Press.
- Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin, and Use*. Greenwood Publishing Group.
- Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to Segment Speech Using Multiple Cues: A Connectionist Model. *Language and Cognitive Processes*, 13(2–3), 221–268. <https://doi.org/10.1080/016909698386528>
- Comerchero, M. D., & Polich, J. (1999). P3a and P3b from typical auditory and visual stimuli. *Clinical Neurophysiology*, 110(1), 24–30. [https://doi.org/10.1016/S0168-5597\(98\)00033-1](https://doi.org/10.1016/S0168-5597(98)00033-1)
- Cotton, S., & Grosjean, F. (1984). The gating paradigm: A comparison of successive and individual presentation formats. *Perception & Psychophysics*, 35(1), 41–48.
- Coulson, S., King, J. W., & Kutas, M. (1998). Expect the Unexpected: Event-related Brain Response to Morphosyntactic Violations. *Language and Cognitive Processes*, 13(1), 21–58. <https://doi.org/10.1080/016909698386582>
- Çukur, T., Nishimoto, S., Huth, A.G., Gallant, J.L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, 16: 763-770.
- Cummings, A., Čeponienė, R., Koyama, A., Saygin, A. P., Townsend, J., & Dick, F. (2006). Auditory semantic networks for words and natural sounds. *Brain Research*, 1115(1), 92–107. <https://doi.org/10.1016/j.brainres.2006.07.050>
- Dahan, D., & Magnuson, J. S. (2006). Chapter 8 - Spoken Word Recognition A2 - Traxler, Matthew J. In M. A. Gernsbacher (Ed.), *Handbook of Psycholinguistics (Second Edition)* (pp. 249–283). London: Academic Press. Retrieved from <http://www.sciencedirect.com/science/article/pii/B9780123693747500092>
- Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time Course of Frequency Effects in Spoken-Word Recognition: Evidence from Eye Movements. *Cognitive Psychology*, 42(4), 317–367. <https://doi.org/10.1006/cogp.2001.0750>
- Dahan, D., & Tanenhaus, M. K. (2004). Continuous Mapping From Sound to Meaning in Spoken-Language Comprehension: Immediate Effects of Verb-Based Thematic Constraints.

- Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2), 498–513.
<https://doi.org/10.1037/0278-7393.30.2.498>
- Dehaene, Stanislas. (2011). The massive impact of literacy on the brain and its consequences for education. *Human Neuroplasticity and Education. Pontifical Academy of Sciences*, 19–32.
- Dehaene, Stanislas, & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, 15(6), 254–262. <https://doi.org/10.1016/j.tics.2011.04.003>
- Dehaene-Lambertz, G., Montavont, A., Jobert, A., Alliol, L., Dubois, J., Hertz-Pannier, L., & Dehaene, S. (2010). Language or music, mother or Mozart? Structural and environmental influences on infants' language networks. *Brain and Language*, 114(2), 53–65. <https://doi.org/10.1016/j.bandl.2009.09.003>
- Dehaene-Lambertz, Ghislaine, Hertz-Pannier, L., & Dubois, J. (2006). Nature and nurture in language acquisition: anatomical and functional brain-imaging studies in infants. *Trends in Neurosciences*, 29(7), 367–373. <https://doi.org/10.1016/j.tins.2006.05.011>
- DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8(8), 1117–1121. <https://doi.org/10.1038/nn1504>
- Dick, F., Bates, E., Wulfeck, B., Utman, J. A., Dronkers, N., & Gernsbacher, M. A. (2001). Language deficits, localization, and grammar: evidence for a distributive model of language breakdown in aphasic patients and neurologically intact individuals. *Psychological Review*, 108(4), 759–788.
- Dick, F., Lee, H. L., Nusbaum, H., & Price, C. J. (2011). Auditory-Motor Expertise Alters “Speech Selectivity” in Professional Musicians and Actors. *Cerebral Cortex*, 21(4), 938–948. <https://doi.org/10.1093/cercor/bhq166>
- Dick, Frederic, Krishnan, S., Leech, R., & Saygin, A. P. (2016). Environmental Sounds. In *Neurobiology of Language* (pp. 1121–1138). Elsevier. <https://doi.org/10.1016/B978-0-12-407794-2.00089-4>
- Donnay, G. F., Rankin, S. K., Lopez-Gonzalez, M., Jiradejvong, P., & Limb, C. J. (2014). Neural Substrates of Interactive Musical Improvisation: An fMRI Study of ‘Trading Fours’ in Jazz. *PLOS ONE*, 9(2), e88665. <https://doi.org/10.1371/journal.pone.0088665>
- Engle, R.W. (2010). Role of Working-Memory Capacity in Cognitive Control. *Current Anthropology* 51 Supplement 1, S17-S26.
- Fagard, J., Sirri, L., & Rămă, P. (2014). Effect of handedness on the occurrence of semantic N400 priming effect in 18- and 24-month-old children. *Frontiers in Psychology*, 5. <https://doi.org/10.3389/fpsyg.2014.00355>

- Federmeier, K. D., & Kutas, M. (1999). A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. *Journal of Memory and Language*, 41(4), 469–495. <https://doi.org/10.1006/jmla.1999.2660>
- Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, 1146, 75–84. <https://doi.org/10.1016/j.brainres.2006.06.101>
- Fodor, J. A. (1983). *Modularity of Mind: An Essay on Faculty Psychology*. Cambridge: MIT Press.
- Frey, A., Aramaki, M., & Besson, M. (2014). Conceptual priming for realistic auditory scenes and for auditory words. *Brain and Cognition*, 84(1), 141–152. <https://doi.org/10.1016/j.bandc.2013.11.013>
- Friederici, A. D. (2011). The Brain Basis of Language Processing: From Structure to Function. *Physiological Reviews*, 91(4), 1357–1392. <https://doi.org/10.1152/physrev.00006.2011>
- Frishkoff, G. A., & Tucker, D. M. (2001). *Anatomy of the N400: brain electrical activity in propositional semantics*. Institute of Cognitive and Decision Sciences, University of Oregon.
- Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, 12(5–6), 613–656.
- Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the P600. *Language and Cognitive Processes*, 25(2), 149–188. <https://doi.org/10.1080/01690960902965951>
- Giraud, A.L., Price, C.J. (2001). The constraints functional neuroimaging places on classical models of auditory word processing. *J Cogn Neurosci* 13: 754-765.
- Giuliano, R.J., Karns, C.M., Neville, H.J., Hillyard, S.A. (2014). Early auditory evoked potential is modulated by selective attention and related to individual differences in visual working memory capacity. *J Cogn Neurosci* 26(12): 2682-2690. https://doi.org/10.1162/jocn_a_00684
- Grodzinsky, Y. (2000). The neurology of syntax: Language use without Broca's area. *Behavioral and Brain Sciences*, 23(01), 1–21.
- Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, 28(4), 267–283. <https://doi.org/10.3758/BF03204386>
- Guthrie, D., & Buchwald, J. S. (1991). Significance Testing of Difference Potentials. *Psychophysiology*, 28(2), 240–244. <https://doi.org/10.1111/j.1469-8986.1991.tb00417.x>
- Gygi, B., Kidd, G. R., & Watson, C. S. (2007). Similarity and categorization of environmental sounds. *Perception & Psychophysics*, 69(6), 839–855.

- Hagoort, P., Baggio, G., & Willems, R. M. (2009). Semantic Unification. In *The Cognitive Neurosciences* (Vol. 4, pp. 819–836). Boston: MIT Press.
- Hall, D.A., Johnsrude, I.S., Haggard, M.P., Palmer, A.R., Akeroyd, M.A., Summerfield, Q. (2002). Spectral and Temporal Processing in Human Auditory Cortex. *Cerebral Cortex* 12(2): 140 - 149.
- Hansen, J. C., & Hillyard, S. A. (1984). Effects of Stimulation Rate and Attribute Cuing on Event-Related Potentials During Selective Auditory Attention. *Psychophysiology*, 21(4), 394–405. <https://doi.org/10.1111/j.1469-8986.1984.tb00216.x>
- Hauk, O., & Pulvermüller, F. (2004). Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology*, 115(5), 1090–1103. <https://doi.org/10.1016/j.clinph.2003.12.020>
- Haxby, J.V., Petit, L., Ungerleider, L.G., Courtney, S.M. (2000). Distinguishing the functional roles of multiple regions in distributed neural systems for visual working memory. *NeuroImage* 11: 380-391.
- Heald, S. L. M., & Nusbaum, H. C. (2014a). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience*, 8. <https://doi.org/10.3389/fnsys.2014.00035>
- Heald, S. L. M., & Nusbaum, H. C. (2014b). Talker variability in audio-visual speech perception. *Frontiers in Psychology*, 5. <https://doi.org/10.3389/fpsyg.2014.00698>
- Heffner, H. E., & Heffner, R. S. (1984). Temporal lobe lesions and perception of species-specific vocalizations by macaques. *Science (New York, N.Y.)*, 226(4670), 75–76.
- Hendrickson, K., Walenski, M., Friend, M., & Love, T. (2015). The organization of words and environmental sounds in memory. *Neuropsychologia*, 69, 67–76. <https://doi.org/10.1016/j.neuropsychologia.2015.01.035>
- Hickok, G. (2012). The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model. *Journal of Communication Disorders*, 45(6), 393–402. <https://doi.org/10.1016/j.jcomdis.2012.06.004>
- Hillyard, S. A., & Picton, T. W. (1978). On and off components in the auditory evoked potential. *Perception & Psychophysics*, 24(5), 391–398.
- Hockett, C. F. (1959). Animal “languages” and human language. *Human Biology*, 31(1), 32–39.
- Hochstetter, K., Berg, P., & Scherg, M. (2009). BESA Research Tutorial 4: Distributed Source Imaging. MEGIS Software GmbH.
- Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., ... Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods*, 45(4), 1099–1114. <https://doi.org/10.3758/s13428-012-0304-z>

- Jaeger, T. F. (2008). Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of Memory and Language*, 59(4), 434–446. <https://doi.org/10.1016/j.jml.2007.11.007>
- Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. (Second). London: Pearson.
- Kaan, E., Harris, A., Gibson, E., & Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes*, 15(2), 159–201.
- Kaganovich, N., Francis, A. L., & Melara, R. D. (2006). Electrophysiological evidence for early interaction between talker and linguistic information during speech perception. *Brain Research*, 1114(1), 161–172. <https://doi.org/10.1016/j.brainres.2006.07.049>
- 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. <https://doi.org/10.1016/j.jml.2004.10.002>
- Kim, J.-R. (2015). Acoustic Change Complex: Clinical Implications. *Journal of Audiology & Otology*, 19(3), 120–124. <https://doi.org/10.7874/jao.2015.19.3.120>
- King, J. W., & Kutas, M. (1995). Who did what and when? Using word- and clause-level ERPs to monitor working memory usage in reading. *Journal of Cognitive Neuroscience*, 376–395.
- Klatt, D. (1979). Speech perception: a model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, 7, 279–312.
- Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in psychtoolbox-3. *Perception*, 36(14), 1–16.
- Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122(2), 148–203. <https://doi.org/10.1037/a0038695>
- Koenig, T., Kottlow, M., Stein, M., & Melie-Garcia, L. (2011). Ragu: A Free Tool for the Analysis of EEG and MEG Event-Related Scalp Field Data Using Global Randomization Statistics. *Computational Intelligence and Neuroscience*, 2011, e938925. <https://doi.org/10.1155/2011/938925>
- Koenig, T., & Melie-Garcia, L. (2009). 8. Statistical analysis of multichannel scalp field data. In *Electrical Neuroimaging* (pp. 169–190). Cambridge University Press.
- Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, 1146, 23–49. <https://doi.org/10.1016/j.brainres.2006.12.063>

- Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31(1), 32–59. <https://doi.org/10.1080/23273798.2015.1102299>
- Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science (New York, N.Y.)*, 207(4427), 203–205.
- Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annual Review of Psychology*, 62, 621–647. <https://doi.org/10.1146/annurev.psych.093008.131123>
- Kutas, M., & Hillyard, S. A. (1980). Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological Psychology*, 11(2), 99–116. [https://doi.org/10.1016/0301-0511\(80\)90046-0](https://doi.org/10.1016/0301-0511(80)90046-0)
- Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947), 161–163.
- Kutas, M., & Van Petten, C. (1994). Psycholinguistics electrified. *Handbook of Psycholinguistics*, 83–143.
- Kwon, H., Kuriki, S., Kim, J. M., Lee, Y. H., Kim, K., & Nam, K. (2005). MEG study on neural activities associated with syntactic and semantic violations in spoken Korean sentences. *Neuroscience Research*, 51(4), 349–357. <https://doi.org/10.1016/j.neures.2004.12.017>
- Lawrence, M. A. (2013). ez: Easy analysis and visualization of factorial experiments. R package version 4.2-2. Retrieved from <http://CRAN.R-project.org/package=ez>
- Le, T. H., Pardo, J. V., & Hu, X. (1998). 4 T-fMRI study of nonspatial shifting of selective attention: cerebellar and parietal contributions. *Journal of Neurophysiology*, 79(3), 1535–1548. <https://doi.org/10.1152/jn.1998.79.3.1535>
- Leech, R., Holt, L. L., Devlin, J. T., & Dick, F. (2009). Expertise with Artificial Non-speech Sounds Recruits Speech-Sensitive Cortical Regions. *The Journal of Neuroscience*, 29(16), 5234–5239. <https://doi.org/10.1523/JNEUROSCI.5758-08.2009>
- Leech, R., & Saygin, A. P. (2011). Distributed processing and cortical specialization for speech and environmental sounds in human temporal cortex. *Brain and Language*, 116(2), 83–90. <https://doi.org/10.1016/j.bandl.2010.11.001>
- Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, 5(4), 356–363. <https://doi.org/10.1038/nn831>
- Lewis, A. G., & Bastiaansen, M. (2015). A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex*, 68, 155–168. <https://doi.org/10.1016/j.cortex.2015.02.014>

- Lewis, J. W. (2005). Distinct Cortical Pathways for Processing Tool versus Animal Sounds. *Journal of Neuroscience*, 25(21), 5148–5158. <https://doi.org/10.1523/JNEUROSCI.0419-05.2005>
- Lewis, J.W., Wightman, F.L., Brefczynski, J.A., Phinney, R.E., Binder, J.R., DeYoe, E.A. (2004). Human brain regions involved in recognizing environmental sounds. *Cerebral Cortex* 14(9): 1008-1021.
- Lieberman, A. M. (1970). The Grammars of Speech and Language. *Cognitive Psychology*, 1(4), 301–323.
- Lieberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6), 431–461.
- Lieberman, A.M., & Mattingly, I.G. (1985). The motor theory of speech perception revised. *Cognition*, 21(1), 1–36. [https://doi.org/10.1016/0010-0277\(85\)90021-6](https://doi.org/10.1016/0010-0277(85)90021-6)
- Luce, P. A., & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and Hearing*, 19(1), 1–36.
- Luck, S. J. (1998). Sources of Dual-Task Interference: Evidence From Human Electrophysiology. *Psychological Science*, 9(3), 223–227. <https://doi.org/10.1111/1467-9280.00043>
- Lupyan, G., & Clark, A. (2015). Words and the World: Predictive Coding and the Language-Perception-Cognition Interface. *Current Directions in Psychological Science*, 24(4), 279–284. <https://doi.org/10.1177/0963721415570732>
- Maess, B., Koelsch, S., Gunter, T.C., Friederici, A.D. (2001). Musical syntax is processed in Broca's area: an MEG study. *Nat Neurosci* 4: 540-545.
- Magnuson, J. S., Tanenhaus, M. K., & Aslin, R. N. (2008). Immediate effects of form-class constraints on spoken word recognition. *Cognition*, 108(3), 866–873. <https://doi.org/10.1016/j.cognition.2008.06.005>
- Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, 25(1), 71–102.
- Maurer, G., Fromkin, V. A., & Cornell, T. L. (1993). Comprehension and Acceptability Judgments in Agrammatism: Disruptions in the Syntax of Referential Dependency. *Brain and Language*, 45(3), 340–370. <https://doi.org/10.1006/brln.1993.1050>
- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18(1), 1–86. [https://doi.org/10.1016/0010-0285\(86\)90015-0](https://doi.org/10.1016/0010-0285(86)90015-0)
- McClelland, J., Mirman, D., & Holt, L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences*, 10(8), 363–369. <https://doi.org/10.1016/j.tics.2006.06.007>

- McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science*, 12(3), 369–378. <https://doi.org/10.1111/j.1467-7687.2009.00822.x>
- McRae, K., Hare, M., Elman, J. L., & Ferretti, T. (2005). A basis for generating expectancies for verbs from nouns. *Memory & Cognition*, 33(7), 1174–1184. <https://doi.org/10.3758/BF03193221>
- Meteyard, L., Bahrami, B., & Vigliocco, G. (2007). Motion Detection and Motion Verbs: Language Affects Low-Level Visual Perception. *Psychological Science*, 18(11), 1007–1013. <https://doi.org/10.1111/j.1467-9280.2007.02016.x>
- Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, 66(4), 545–567. <https://doi.org/10.1016/j.jml.2012.01.001>
- Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90(2), 227–234. <https://doi.org/10.1037/h0031564>
- Mirman, D., McClelland, J. L., & Holt, L. L. (2006). An interactive Hebbian account of lexically guided tuning of speech perception. *Psychonomic Bulletin & Review*, 13(6), 958–965.
- Morin, T. M., & Nusbaum, H. C. (1989). Cues to perceptual normalization of talker differences. *The Journal of the Acoustical Society of America*, 86(S1), S100–S100. <https://doi.org/10.1121/1.2027256>
- Morris, A. L., & Harris, C. L. (2002). Sentence context, word recognition, and repetition blindness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(5), 962–982. <https://doi.org/10.1037//0278-7393.28.5.962>
- Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, 47(4), 379–390.
- Mummery, C. J., Ashburner, J., Scott, S. K., & Wise, R. J. (1999). Functional neuroimaging of speech perception in six normal and two aphasic subjects. *The Journal of the Acoustical Society of America*, 106(1), 449–457.
- Murray, M. M., Brunet, D., & Michel, C. M. (2008). Topographic ERP Analyses: A Step-by-Step Tutorial Review. *Brain Topography*, 20(4), 249–264. <https://doi.org/10.1007/s10548-008-0054-5>
- Nittono, H., Nageishi, Y., Nakajima, Y., & Ullsperger, P. (1999). Event-related potential correlates of individual differences in working memory capacity. *Psychophysiology*, 36(6), 745–754.
- Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition*, 52(3), 189–234. [https://doi.org/10.1016/0010-0277\(94\)90043-4](https://doi.org/10.1016/0010-0277(94)90043-4)

- Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, 115(2), 357–395. <https://doi.org/10.1037/0033-295X.115.2.357>
- Nowak, M. A., & Komarova, N. L. (2001). Towards an evolutionary theory of language. *Trends in Cognitive Sciences*, 5(7), 288–295. [https://doi.org/10.1016/S1364-6613\(00\)01683-1](https://doi.org/10.1016/S1364-6613(00)01683-1)
- Nowak, M. A., Plotkin, J. B., & Jansen, V. A. A. (2000). The evolution of syntactic communication. *Nature*, 404(6777), 35006635. <https://doi.org/10.1038/35006635>
- Nusbaum, H. C., & Magnuson, J. S. (1997). Talker normalization: Phonetic constancy as a cognitive process. *Talker Variability in Speech Processing*, 109–132.
- Nusbaum, H. C., & Morin, T. M. (1992). Paying Attention to Differences Among Talkers. In *Speech Perception, Production and Linguistic Structure*. IOS Press.
- Nusbaum, H. C., & Schwab, E. C. (1986). The role of attention and active processing in speech perception. In *Pattern Recognition by Humans and Machines: Speech Perception* (Vol. 1, pp. 113–157). San Diego: Academic Press.
- Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, 17(4), 273–281. <https://doi.org/10.1080/17470216508416445>
- Orgs, G, Lange, K., Dombrowski, J., & Heil, M. (2007). Is conceptual priming for environmental sounds obligatory? *International Journal of Psychophysiology*, 65(2), 162–166. <https://doi.org/10.1016/j.ijpsycho.2007.03.003>
- Orgs, Guido, Lange, K., Dombrowski, J.-H., & Heil, M. (2006). Conceptual priming for environmental sounds and words: An ERP study. *Brain and Cognition*, 62(3), 267–272. <https://doi.org/10.1016/j.bandc.2006.05.003>
- Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31(6), 785–806. [https://doi.org/10.1016/0749-596X\(92\)90039-Z](https://doi.org/10.1016/0749-596X(92)90039-Z)
- Pascual-Marqui, R. D. (2002). Standardized low resolution brain electromagnetic tomography (sLORETA): technical details. *Methods & Findings in Experimental & Clinical Pharmacology*, 24(SupplD), 5–12.
- Patel, A. D., Gibson, E., Ratner, J., Besson, M., & Holcomb, P. J. (1998). Processing Syntactic Relations in Language and Music: An Event-Related Potential Study. *Journal of Cognitive Neuroscience*, 10(6), 717–733. <https://doi.org/10.1162/089892998563121>
- Petkov, C. I., & Jarvis, E. (2012). Birds, primates, and spoken language origins: behavioral phenotypes and neurobiological substrates. *Frontiers in Evolutionary Neuroscience*, 4. <https://doi.org/10.3389/fnevo.2012.00012>

- Pfennig, A. R., Hara, E., Whitney, O., Rivas, M. V., Wang, R., Roulhac, P. L., ... Jarvis, E. D. (2014). Convergent transcriptional specializations in the brains of humans and song-learning birds. *Science*, 346(6215), 1256846–1256846. <https://doi.org/10.1126/science.1256846>
- Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, 11(3), 105–110. <https://doi.org/10.1016/j.tics.2006.12.002>
- Picton, T. W., & Hillyard, S. A. (1974). Human auditory evoked potentials. II: Effects of attention. *Electroencephalography and Clinical Neurophysiology*, 36, 191–200. [https://doi.org/10.1016/0013-4694\(74\)90156-4](https://doi.org/10.1016/0013-4694(74)90156-4)
- Pisoni, D. B., & Luce, P. A. (1987). Acoustic-phonetic representations in word recognition. *Cognition*, 25(1–2), 21–52.
- Potter, M. C., Kroll, J. F., Yachzel, B., Carpenter, E., & Sherman, J. (1986). Pictures in sentences: Understanding without words. *Journal of Experimental Psychology: General*, 115(3), 281.
- Potts, G. F., & Tucker, D. M. (2001). Frontal evaluation and posterior representation in target detection. *Cognitive Brain Research*, (11), 147–156.
- Rabovsky, M., & McRae, K. (2014). Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. *ResearchGate*, 132(1), 68–89. <https://doi.org/10.1016/j.cognition.2014.03.010>
- Ranganath, C., Johnson, M.K., D'Esposito, M. (2003). Prefrontal activity associated with working memory and episodic long-term memory. *Neuropsychologia* 41: 378-389.
- Rauschecker, J.P., Scott, S.K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nat Neurosci* 12(6): 718-724.
- Regel, S., Meyer, L., & Gunter, T. C. (2014). Distinguishing Neurocognitive Processes Reflected by P600 Effects: Evidence from ERPs and Neural Oscillations. *PLOS ONE*, 9(5), e96840. <https://doi.org/10.1371/journal.pone.0096840>
- Reuter-Lorenz, P. (2002). New visions of the aging mind and brain. *Trends in Cognitive Sciences*, 6(9), 394.
- Revill, K. P., Aslin, R. N., Tanenhaus, M. K., & Bavelier, D. (2008). Neural correlates of partial lexical activation. *Proceedings of the National Academy of Sciences*, 105(35), 13111–13115. <https://doi.org/10.1073/pnas.0807054105>
- Russell, G. S., Jeffrey Eriksen, K., Poolman, P., Luu, P., & Tucker, D. M. (2005). Geodesic photogrammetry for localizing sensor positions in dense-array EEG. *Clinical Neurophysiology*, 116(5), 1130–1140. <https://doi.org/10.1016/j.clinph.2004.12.022>

- Santi, A., Grodzinsky, Y. (2007). Working memory and syntax interact in Broca's area. *NeuroImage* 37(1): 8-17.
- Saygin, A. P. (2003). Neural resources for processing language and environmental sounds: Evidence from aphasia. *Brain*, 126(4), 928–945. <https://doi.org/10.1093/brain/awg082>
- Saygin, Ayşe Pinar, Dick, F., & Bates, E. (2005). An on-line task for contrasting auditory processing in the verbal and nonverbal domains and norms for younger and older adults. *Behavior Research Methods*, 37(1), 99–110.
- Schneider, T. R., Engel, A. K., & Debener, S. (2008). Multisensory Identification of Natural Objects in a Two-Way Crossmodal Priming Paradigm. *Experimental Psychology*, 55(2), 121–132. <https://doi.org/10.1027/1618-3169.55.2.121>
- Shillcock, R. C., & Bard, E. G. (1993). Modularity and the Processing of Closed-class Words. In *Cognitive Models of Speech Processing: The Second Sperlonga Meeting* (pp. 163–185). East Sussex: Lawrence Erlbaum Associates Ltd.
- Shintel, H., & Nusbaum, H. C. (2007). The sound of motion in spoken language: visual information conveyed by acoustic properties of speech. *Cognition*, 105(3), 681–690. <https://doi.org/10.1016/j.cognition.2006.11.005>
- Shintel, H., Nusbaum, H. C., & Okrent, A. (2006). Analog acoustic expression in speech communication. *Journal of Memory and Language*, 55(2), 167–177. <https://doi.org/10.1016/j.jml.2006.03.002>
- Simos, P. G., Basile, L. F. H., & Papanicolaou, A. C. (1997). Source localization of the N400 response in a sentence-reading paradigm using evoked magnetic fields and magnetic resonance imaging. *Brain Research*, 762(1), 29–39. [https://doi.org/10.1016/S0006-8993\(97\)00349-1](https://doi.org/10.1016/S0006-8993(97)00349-1)
- Simpson, G. B., Peterson, R. R., Casteel, M. A., & Burgess, C. (1989). Lexical and sentence context effects in word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(1), 88–97. <https://doi.org/10.1037/0278-7393.15.1.88>
- Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, 82, 1–17. <https://doi.org/10.1016/j.jml.2015.02.004>
- Strauss, T. J., Harris, H. D., & Magnuson, J. S. (2007). jTRACE: a reimplementation and extension of the TRACE model of speech perception and spoken word recognition. *Behavior Research Methods*, 39(1), 19–30.
- Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, 18(6), 645–659. [https://doi.org/10.1016/S0022-5371\(79\)90355-4](https://doi.org/10.1016/S0022-5371(79)90355-4)

- Szenkovits, G., Peelle, J.E., Norris, D., Davis, M.H. (2012). Individual differences in premotor and motor recruitment during speech perception. *Neuropsychologia* 50(7): 1380-1392. DOI: 10.1016/j.neuropsychologia.2012.02.023
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217). Retrieved from <https://pdfs.semanticscholar.org/7654/91c9cfc746275000297fc0ae919b5c4440e0.pdf>
- Tanner, D., Morgan-Short, K., & Luck, S. J. (2015). How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. *Psychophysiology*, 52(8), 997–1009. <https://doi.org/10.1111/psyp.12437>
- Trask, R. L. (2003). *Language: The Basics*. Routledge.
- Tyler, Lorraine K., Voice, J. K., & Moss, H. E. (1996). The interaction of semantic and phonological processing. In *In Eighteenth Annual Conference of the Cognitive Science Society*.
- Tyler, Lorraine K., & Wessels, J. (1983). Quantifying contextual contributions to word-recognition processes. *Perception & Psychophysics*, 34(5), 409–420.
- Tyler, Lorraine K., & Wessels, J. (1985). Is gating an on-line task? Evidence from naming latency data. *Perception & Psychophysics*, 38(3), 217–222. <https://doi.org/10.3758/BF03207148>
- Tyler, Lorraine Komisarjevsky, & Marslen-Wilson, W. (1986). The effects of context on the recognition of polymorphemic words. *Journal of Memory and Language*, 25(6), 741–752. [https://doi.org/10.1016/0749-596X\(86\)90047-1](https://doi.org/10.1016/0749-596X(86)90047-1)
- Uddin, S., Heald, S. L. M., Van Hedger, S. C., Klos, S., & Nusbaum, H. C. (2018). Understanding environmental sounds in sentence context. *Cognition*, 172, 134–143. <https://doi.org/10.1016/j.cognition.2017.12.009>
- Uddin, S., Heald, S. L. M., Van Hedger, S. C., & Nusbaum, H. C. (2018). Hearing sounds as words: Neural responses to environmental sounds in sentence context. *Brain & Language* 179: 51-61. <https://doi.org/10.1016/j.bandl.2018.02.004>
- Van Hedger, S. C., Heald, S. L. M., & Nusbaum, H. C. (2015). The effects of acoustic variability on absolute pitch categorization: Evidence of contextual tuning. *The Journal of the Acoustical Society of America*, 138(1), 436–446. <https://doi.org/10.1121/1.4922952>
- van Petten, C., & Rheinfelder, H. (1995). Conceptual relationships between spoken words and environmental sounds: Event-related brain potential measures. *Neuropsychologia*, 33(4), 485–508. [https://doi.org/10.1016/0028-3932\(94\)00133-A](https://doi.org/10.1016/0028-3932(94)00133-A)

- Vogel, E. K., Luck, S. J., & Shapiro, K. L. (1998). Electrophysiological evidence for a postperceptual locus of suppression during the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance*, 24(6), 1656–1674.
- Vouloumanos, A., Kiehl, K. A., Werker, J. F., & Liddle, P. F. (2001). Detection of Sounds in the Auditory Stream: Event-Related fMRI Evidence for Differential Activation to Speech and Non-speech. *Journal of Cognitive Neuroscience*, 13(7), 994–1005. <https://doi.org/10.1162/089892901753165890>
- Warren, W. H., & Verbrugge, R. R. (1984). Auditory perception of breaking and bouncing events: a case study in ecological acoustics. *Journal of Experimental Psychology. Human Perception and Performance*, 10(5), 704–712.
- Wong, P. C. M., Nusbaum, H. C., & Small, S. L. (2004). Neural bases of talker normalization. *Journal of Cognitive Neuroscience*, 16(7), 1173–1184. <https://doi.org/10.1162/0898929041920522>
- Yantis S., Serences, J.T. (2003). Cortical mechanisms of space-based and object-based attentional control. *Current Opinion in Neurobiology* 13(2): 187-193. DOI: 10.1016/S0959-4388(03)00033-3
- Zhang, C., Peng, G., & Wang, W. S.-Y. (2013). Achieving constancy in spoken word identification: Time course of talker normalization. *Brain and Language*, 126(2), 193–202. <https://doi.org/10.1016/j.bandl.2013.05.010>
- Zwaan, R. A., & Pecher, D. (2012). Revisiting Mental Simulation in Language Comprehension: Six Replication Attempts. *PLOS ONE*, 7(12), e51382. <https://doi.org/10.1371/journal.pone.0051382>
- Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science*, 13(2), 168–171. <https://doi.org/10.1111/1467-9280.00430>

Appendices

Appendix 1: Stimuli

Table A1. Paired environmental sounds and spoken words used in the current studies.

Sound	Word
baby laughing	"baby"
camera shutter	"camera"
car engine revving	"car"
cashregister ch-ching	"cashregister"
cat meowing	"cat"
churchbells ringing	"churchbells"
clock ticking	"clock"
coin dropping onto hard surface	"coin"
cow mooing	"cow"
crow cawing	"crow"
dog barking	"dog"
creaky door closing	"door"
doorbell ringing	"doorbell"
drum set	"drums"
frog croaking	"frog"
guitar being strummed	"guitar"
gunshot	"gunshot"
helicopter	"helicopter"
car horn	"horn"
papers being ruffled	"paper"
phone ringing	"phone"
octave played on piano	"piano"
rooster crowing	"rooster"
saxophone notes	"saxophone"
servicebell ringing	"servicebell"
sheep bleating	"sheep"
ambulance/police siren	"siren"
sword being unsheathed	"sword"
toilet flushing	"toilet"
train whistle	"train"
water dripping	"water"
zipper	"zipper"

Appendix 2: Detailed methodology for topographical randomization analyses

RAGU (Randomization Graphical User interface) is an open-source Matlab-based program that compares the collected data with a null (estimated control) distribution generated from resampled randomization as a nonparametric significance test. It uses all 128 electrodes' average re-referenced voltage as a “map” of scalp topography; then it randomizes the data between experimental conditions and re-calculates scalp topography, creating a null distribution of topographies based on 5,000 randomization runs. This simulated null distribution can then be compared with the actual collected data. This avoids biases associated with *a priori* assumptions about which time windows and electrodes should be included in analysis (see Koenig et al., 2011; Murray et al., 2008). The individual subject-level data that we used as direct input to RAGU (ascii files of waveforms averaged within participants and conditions) is available on Open Science Framework (<https://osf.io/asw48/>). We used RAGU to perform two TANOVAs (topographic analyses of variance, not related to an ANOVA). The first (1) compared the effects of constraint (low (general) or high (specific)) and target type (ES or spoken word) on scalp topography of [mismatch – match] differences across the entire period from stimulus onset to 900 ms afterwards, using 5,000 simulations. The second (2) compared the effects of congruency (match or mismatch) and target type (ES or spoken word) on raw scalp topographies, and (as explained in *Methods*) was intended to be more exploratory, as raw differences between ES and words are not particularly informative for a host of reasons.

RAGU measures differences between topographical maps in different conditions using *generalized dissimilarity*, here abbreviated as *s*. This is a measure of the difference in spatial (scalp) distribution of EEG activity—which in principle implies changes in the underlying pattern

of neural activity—over electrodes and across conditions. Briefly, RAGU’s TANOVA calculates generalized dissimilarity s across the experimental conditions:

$$s = \sum_{i=1}^c \sqrt{\frac{\sum_{j=1}^n (\bar{v}_{ij} - \bar{\bar{v}}_j)^2}{n}}$$

where c is the number of conditions, n is the number of electrodes, $\bar{v}_{ij}$ is the mean voltage of condition i at electrode j across subjects, and $\bar{\bar{v}}_j$ is the mean voltage at electrode j across all subjects, with conditions averaged together (from Koenig & Melie-García, 2009). Once RAGU has calculated the generalized topographic dissimilarity s from the data, it scrambles the data between conditions and re-calculates s 5,000 times to generate a null distribution of generalized dissimilarities that would be expected if there were no differences between conditions. A p value is calculated from this output as the proportion of the 5,000 simulations that show equal or greater differences between conditions compared to the observed (i.e. unscrambled) data. The 95% CI can also be calculated directly from the quantiles of this distribution.

This measure of the generalized topographic dissimilarity in EEG s can be used to examine main effects and interactions of experimental conditions separately using RAGU. This is done by first normalizing the voltages across the scalp at each time point, then calculating s for the conditions corresponding to the first main factor of interest (again, at each time point). For example, in the second analysis in the current experiment, after normalizing the data, RAGU calculated s between match and mismatch endings to examine the main effect of congruency (sentences ending in sounds and words were averaged together for this step of the analysis). When repeated at each time point, this step determines at which time points in the ERP (measured beginning at target onset) there are significant changes in the distribution of scalp measured electrical activity as a result of the match or mismatch status of the sentence ending.

Once this has been done, the mean scalp topography in each condition is subtracted from the individual data matching that condition. As a concrete example: the mean topography in response to a mismatch sentence ending was subtracted from both the mismatch/sound and mismatch/word conditions for all subjects. This removes variance due to the first main effect examined, allowing RAGU to repeat the procedure to obtain the main effect of target type by calculating s between sound and word endings. Finally, topographies associated with target type were subtracted, and the interaction could be calculated as remaining variance in scalp topography. The output of this analysis is a time series showing a p value across time for the main effects and interactions included in the model.

Because a distribution of 5,000 randomizations is compared to the original topographical maps at every time point, there are multiple comparisons across time. To avoid drawing conclusions from significance windows that are false positives due to these multiple comparisons, we reasoned, similarly to previous authors, that real ERP effects should show up in multiple consecutive windows (e.g., Guthrie & Buchwald, 1991). Consequently, as a conservative estimate, only significance windows 40 ms or longer in duration were treated as robust effects, unless there was an *a priori* reason to expect an effect in a certain time window.

Appendix 3: AAHC microstate analysis

The algorithms for microstate analysis, as well as the explanations of the equations used, come largely from Murray et al. (2008) in which their group's work on developing the theory behind microstate analyses is summarized and explained. Microstate analysis is simply an unsupervised pattern classification algorithm applied to topographical maps obtained from EEG. In this case, the data fed into the classifier is the averaged topographical map in each condition and at each time point in the time-locked event-related potential. As in k-means clustering, the

algorithm begins by treating each topographical map as a unique cluster. The spatial correlation C is then calculated between all the various topographical maps, both within conditions at different time points, and across conditions:

$$C_{u,v} = \frac{\sum_{i=1}^n u_i \cdot v_i}{\|u\| \cdot \|v\|}$$

where u_i and v_i are average-referenced potentials of the i th electrode either at two different time points, or in two different conditions. Topographical maps are then redefined by averaging together the template maps with the highest spatial correlation with each other, and the “microstate” that defines this new cluster is defined as the centroid of the maps in the cluster. The next step is where AAHC clustering differs from k-means: the global explained variance (GEV) of all the clusters is calculated, and the cluster with lowest GEV is eliminated before the process is repeated iteratively to reduce the number of clusters.

$$GEV = \frac{\sum_{t=1}^{tmax} (GFP_u(t) \cdot C_{u,T_t})^2}{\sum_{t=1}^{tmax} GFP_u^2(t)}$$

where t is the time point, $GFP_u(t)$ is the global field potential i.e. standard deviation across all scalp electrodes for condition u at time point t , T_t is the centroid of the cluster to which the topography of condition u at time t belongs, and C_{u,T_t} is the spatial correlation between the topography of condition u at time t and the centroid of the cluster that it belongs to. Thus, we can see that the closer the correspondence between the topographical data and the representative map of the microstate category it fits into, the higher the GEV . Therefore, the clusters that are iteratively dropped from the clusters by the AAHC algorithm are those that have the poorest correspondence between the topographies and the representative maps of the microstates to which they have been assigned.

Appendix 4: Questionnaires

1



Demographic Information

How old are you? _____ years

Are you: ☐ Male ☐ Female

Are you: ☐ Left-handed ☐ Right-handed

Are you a student? ☐ Yes ☐ No

If yes, what year?

What is your major / what department are you in?

Year Participant ID#
 -

Language Experience Questionnaire

Age: _____

Gender: _____

Primary Language: _____

1. Do you speak any secondary languages? YES NO

If so, please list them below, along with your level of proficiency for each:

_____	Basic	Proficient	Semi-Fluent	Fluent
_____	Basic	Proficient	Semi-Fluent	Fluent
_____	Basic	Proficient	Semi-Fluent	Fluent
_____	Basic	Proficient	Semi-Fluent	Fluent

2. In what country were you born? _____

3. Please list the different areas in which you have lived, including the country, region/state, and number of years. Please write your answers chronologically (earliest to most recent).

_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____

(City, State/Province/Region, Country)

(Number of Years)

Music Experience Questionnaire

1. Do you play (or have you played) any musical instruments? YES NO

1b. At what age (approximately) did you first begin musical instruction? _____

1c. Do you still actively play a musical instrument? YES NO

2. If you circled YES in Question 1, please list all the musical instruments you have played, including the number of years you have played them:

Instrument _____ Number of Years _____

Instrument _____ Number of Years _____

Instrument _____ Number of Years _____

Instrument _____ Number of Years _____

Instrument _____ Number of Years _____

Voice _____ Number of Years _____

3. Do you know how to read music? YES NO

4. Have you ever played in any ensembles or sung in any choirs? YES NO

5. If you answered YES to Question 4, please list the ensembles and/or choirs with which you played/sung.

Ensemble/Choir _____ Number of Years _____

Ensemble/Choir _____ Number of Years _____

6. Have you taken any music courses at the high school or collegiate level? YES NO

7. If you answered YES to Question 6, please list the approximate number of music courses you have completed

8. Do you have perfect pitch (the ability to name any note without a reference)?

YES NO

9. Approximately how many hours do you spend making music per week? (circle one):

Less than one hour 1-2 hours 3-4 hours 5-6 hours More than 6 hours

10. Approximately how many hours do you spend listening to music per day (circle one):

Less than one hour 1-2 hours 3-4 hours 5-6 hours More than 6 hours

11. Approximately what percentage of music you listen to falls under the following genres?

Pop/Rock _____

Dance/Techno _____

Jazz _____

R&B _____

Hip Hop _____

Classical _____

Folk _____

Non-Western/Ethnic _____

Other (please specify) _____

12. Approximately how many concerts do you attend per year? _____

Participant Questionnaire

PARTICIPANT ID: _____

DATE: _____

REMINDERS:

- ☐ No Alcohol or Drug use starting night before
- ☐ No caffeinated beverages for 48 hours prior to the experiment
- ☐ No previous exposure to this task

1. What time do you usually go to sleep?
2. Do you fall asleep easily?
3. How deeply do you sleep (Light, Medium, Deep)?
4. What time do you usually wake up?
5. Do you usually feel well rested?
6. Have you ever sought medical attention for a sleep disorder?
7. Do you have any disabilities that disrupt your sleep?
8. Are you currently taking any medications to help you sleep?
9. Are you taking any other medications?
(excluding oral contraceptives)
10. Do you have a history of substance abuse or diagnosed major mental illness?
11. How many caffeinated beverages do you drink each day?



THE UNIVERSITY OF
CHICAGO

CENTER FOR COGNITIVE AND SOCIAL
NEUROSCIENCE

Date: _____

ID#: _____

Subject: _____

Choose the number of the description that best fits your current state:

- 1 feeling active and vital; alert; wide awake
- 2 functioning at a high level, but not at peak; able to concentrate
- 3 relaxed; awake; not at full alertness; responsive
- 4 a little foggy, not at peak; let down
- 5 foggiess; beginning to lose interest in remaining awake; slowed down
- 6 sleepiness; prefer to be lying down; fighting sleep, woozy
- 7 almost in reverie; sleep onset soon; lost struggle to remain awake
- X asleep

SS