UNIVERSITY OF STIRLING

Disfluency as. . . er. . . delay: An investigation into the immediate and lasting consequences of disfluency and temporal delay using EEG and mixed-effects modelling.

by

Jennifer A.E. Bouwsema

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in

Psychology

School of Natural Sciences

October 2014

Abstract

Difficulties in speech production are often marked by disfluency; fillers, hesitations, prolongations, repetitions and repairs. In recent years a body of work has emerged that demonstrates that listeners are sensitive to disfluency, and that this affects their expectations for upcoming speech, as well as their attention to the speech stream. This thesis investigates the extent to which delay may be responsible for triggering these effects.

The experiments reported in this thesis build on an Event Related Potential (ERP) paradigm developed by Corley et al., (2007), in which participants listened to sentences manipulated by both fluency and predictability. Corley et al. reported an attenuated N400 effect for words following disfluent *ers*, and interpreted this as indicating that the extent to which listeners made predictions was reduced following an *er*. In the current set of experiments, various noisy interruptions were added to Corley et al.,'s paradigm, time matched to the disfluent fillers. These manipulations allowed investigation of whether the same effects could be triggered by delay alone, in the absence of a cue indicating that the speaker was experiencing difficulty.

The first experiment, which contrasted disfluent *ers* with artificial beeps, revealed a small but significant reduction in N400 effect amplitude for words affected by *ers* but not by beeps. The second experiment, in which *ers* were contrasted with speaker generated coughs, revealed no fluency effects on the N400 effect. A third experiment combined the designs of Experiments 1 and 2 to verify whether the difference between them could be characterised as a context effect; one potential explanation for the difference between the outcomes of Experiments 1 and 2 is that the interpretation of an *er* is affected by the surrounding stimuli. However, in Experiment 3, once again no effect of fluency on the magnitude of the N400 effect was found. Taken together, the results of these three studies lead to the question of whether *er's* attenuation effect on the N400 is robust.

In a second part to each study, listeners took part in a surprise recognition memory test, comprising words which had been the critical words in the previous task intermixed with new words which had not appeared anywhere in the sentences previously heard. Participants were significantly more successful at recognising words which had been unpredictable in their contexts, and, importantly, for Experiments 1 and 2, were significantly more successful at recognising words which had featured in disfluent or interrupted sentences. There was no difference between the recognition rates of words which had been disfluent and those which were affected by a noisy interruption. Collard et al., (2008) demonstrated that disfluency could raise attention to the speech stream,

and the finding that interrupted words are equally well remembered leads to the suggestion that any noisy interruption can raise attention. Overall, the finding of memory benefits in response to disfluency, in the absence of attenuated N400 effects leads to the suggestion that different elements of disfluencies may be responsible for triggering these effects.

The studies in this thesis also extend previous work by being designed to yield enough trials in the memory test portion of each experiment to permit ERP analysis of the memory data. Whilst clear ERP memory effects remained elusive, important progress was made in that memory ERPs were generated from a disfluency paradigm, and this provided a testing ground on which to demonstrate the use of linear mixed-effects modelling as an alternative to ANOVA analysis for ERPs. Mixed-effects models allow the analysis of unbalanced datasets, such as those generated in many memory experiments. Additionally, we demonstrate the ability to include crossed random effects for subjects and items, and when this is applied to the ERPs from the listening section of Experiment 1, the effect of fluency on N400 amplitude is no longer significant.

Taken together, the results from the studies reported in this thesis suggest that temporal delay or disruption in speech can trigger raised attention, but do not necessarily trigger changes in listeners' expectations.

Acknowledgements

I have been surprised, perhaps naively, to learn how much a thesis is a collective project. There are many people without whom this work would not have been possible, but whose names do not appear on the cover.

Firstly, thanks are due to Professor David I. Donaldson, and Dr. Martin Corley, who have provided four years of sound advice, guidance, expertise and stimulating discussions. Their enthusiasm for the subject and for the development of their students has become increasingly evident throughout my time working with them. I am grateful to Ekaterina Klepousniotou for introducing me to psycho-linguistics, and to Lucy MacGregor, upon whose work this project was based, for her patience in showing me the ropes of EEG, and her enthusiasm and encouragement in persuading me to pursue a Ph.D. with David and Martin.

Much of the work in this thesis has only been possible with the excellent technical assistance of Catriona Bruce, whose competence and expertise keeps the Psychological Imaging Laboratory together; and Eric Bouwsema, whose willingness to step in at short notice to set up and support the running of R on cloud-based servers made possible the implementation of mixed-effects modelling on large ERP datasets. Thanks are also due to Ric Sharp, who kindly provided some code to allow the extraction of single trials from the raw EEG.

I am indebted to Adam Milligan and Kirstin Bouwsema, whose patient and diligent proof-reading has reduced my opportunity for typographic embarrassment. Thanks are due to Elli Drake, who provided the voice of the experiments. I am grateful to Dave Williamson for many stimulating and insightful conversations about this work, and particularly value his input on the computational challenges encountered.

Last but not least, the very fact of my finally having arrived at the point of writing acknowledgements owes a great deal to my family, who have been unwavering in their support; to Laura Hamlet, whose encouragement, commiseration, and coffee-drinking abilities are unparalleled; to Sarah Peters, with whom many hours have been spent in companionable silence, writing theses at extreme latitudes in both hemispheres; and to Kevin Bouwsema, who has succeeded against the odds of making the year spent writing this thesis one of the happiest of my life.

Contents

References 301

List of Figures

List of Tables

Chapter 1

Thesis Overview

For most people, speaking and listening in their mother tongue comes as naturally as eating or breathing, and we rarely, if ever, give much thought to the processes that facilitate our communication. We often think of speech as analogous to writing; a well formed stream of words, conveying a message. A moment's thought shows, however, that this view is overly simplistic. If we imagine a speaker uttering a stream of words without rhythm, pause, or intonation, perhaps similar to early text-to-speech programmes, we quickly realise that this stream would be much harder to understand than if it were uttered normally. Once we realise this, it is natural to begin to wonder about the comprehension benefits of other elements in the communication stream. What role, if any, is played by gesture, expression, voice tone, rhythm, hesitation and disfluency? In this thesis, I will focus on disfluency; a phenomenon of imperfect language production, and specifically, how disfluency impacts the listener.

In Chapter 2, I will introduce the reader to disfluency and the body of research surrounding listeners' responses to disfluent speech. Chapter 3 provides a background to understanding ERP methodology, which is the primary experimental tool used in this

work, and Chapter 4 provides detailed methods for the three experiments reported in this thesis. The immediate impact of disfluent interruptions on listeners' speech processing is examined in Chapters 5, 6 and 7, which report the ERP results of three experiments designed to capture online effects of participants hearing disfluent speech, while the longer term significance of disfluency is addressed in Chapter 8, where the results of memory tests are examined. Chapter 9 introduces an alternative analysis methodology with the potential to increase the efficiency of data-collection in long ERP experiments, such as those presented in this thesis, and asks whether this methodology can tell us anything more about the data presented. Finally, Chapter 10 provides a discussion of the findings of the experiments reported in this thesis, and considers them in the context of each other and of the existing literature.

Chapter 2

Introduction to Disfluency

2.1 Introduction

Disfluencies are typically classified as phenomena interrupting the flow of speech and not adding propositional content to an utterance (Fox Tree, 1995). These may include pauses, repetitions and false starts, as well as fillers such as *er* and *um*. Exclusive of silent pauses (which are often difficult to classify as fluent or otherwise) disfluencies are estimated to affect 6% of words uttered in spontaneous speech (Fox Tree, 1995). Despite the relative prevalence of disfluency, it might be tempting to dismiss it as irrelevant to speech comprehension, particularly given that listeners are highly adept at filtering it out. For example, in transcription or reproduction tasks, listeners often move disfluencies to clause boundaries (e.g. Duez, 1985; Martin, 1967: Martin and Strange, 1968) or omit them altogether. In recent years, however, a body of work has been published demonstrating that disfluency does, in fact, have some significant and interesting effects on listeners' online language comprehension, and subsequent representations of spoken language. This chapter will explore what disfluency is, how and when it is produced,

and discuss its immediate and lasting effects on listeners. Finally attention will turn to what features of disfluency drive these effects on listeners, which is the theoretical question at the heart of this thesis.

2.1.1 The Origins of Disfluency

To understand disfluency, it is logical to begin with the speaker, as it is here that disfluency originates. For the purposes of the work in this thesis, a basic Leveltian three stage model of language production is assumed (Levelt, 1983, 1989). First, the speaker must plan the utterance, which involves conceptualising the utterance in pre-verbal form. Second, the speaker must transform their preverbal message into a verbal plan; they must select appropriate word forms and syntactic structures, as well as the relevant phonological and articulatory forms. The third and final stage in this process is the actual articulation of the message, in which the speaker physically produces the speech.

As speech production is an incremental process, and speakers are not only formulating, but also conceptualising utterances while speaking (Marslen-Wilson & Tyler, 1980), it is unsurprising that difficulties sometimes occur. These difficulties may often result in disfluency.

2.2 Types of disfluency

Disfluency is often classified into a number of categories. These not only represent different manifestations of disfluency, but have also been theorised as indexing difficulties at different stages of the language production process. Below is a brief overview of some of the main types of disfluency produced in spontaneous speech.

Hesitations

Hesitations are perhaps the most instinctively recognised manifestation of disfluency. Hesitation encompasses three forms, which often (but not always) co-occur. Lexical fillers such as *er*, *erm*, *eh*, *uh*, *uhm*, *mm*[1] ,pauses and prolongations are all classified as hesitations. Pauses may be silent, or contain a filler, and are often preceded by prolongations of syllables in the words leading up to the pause, for example lengthening of the definite article *the*, so that it is pronounced *thee*. Some prolongation also follows filled hesitations; word duration tends to be an average of 1.19 times longer than when the same word appears in the same sentence context fluently spoken (Bell et al., 2003).

Pauses which are completely silent present something of a challenge to disfluency research, as they are difficult to classify as fluent or disfluent. Whilst they may reflect speaker difficulty, the natural prosody of an utterance may also lead to fluent, intended silent pauses. Fraundorf and Watson (2008) associated fillers and silent pauses with detected difficulty in upcoming speech; that is, when speakers (who are monitoring the internal speech production) detect a problem in speech which has not yet been uttered. Fraundorf and Watson posit that fillers are more likely to be produced when the detected problem is at the level of conceptualisation or translating a pre-verbal message into its form, and silent pauses are more likely when speakers detect upcoming difficulties at grammatical or articulatory levels.

[1]There is no generally accepted standard orthography for filled pauses, although as a rule of thumb, North American authors tend to transcribe fillers as *uh, uhm* whereas British authors favour *er, erm*. Throughout this thesis I use the transcription provided by the authors of each paper referred to, and no particular distinction is made between different orthographic representations of fillers, except where this is explicitly made clear.

Overt Repairs

Repairs occur when a speaker detects an error after it has been articulated, and stops mid-speech, then corrects themselves by providing new information to replace the erroneous speech already uttered. Repairs often take the form of modified repetitions of phonemes, words or phrases. These may repair problems in articulation, lexical choice, or the structure of a phrase or whole utterance. Levelt (1983) defined three distinct stages of an overt repair: a reparandum; an editing term or pause; and the repair. For example, in the following utterance;

Have you got some...er... any clean socks?

'some' is the reparandum, which is the material to be corrected. The reparandum is followed by an edit interval, which in this example is marked by a filled pause. In real speech, however edit intervals may also be unmarked. The repair follows the edit interval, and is made up of information to replace the erroneous speech in the reparandum.

Repetitions

Repetitions describe repeated phrases, words or syllables which do not add to the propositional content of an utterance. Repetitions can be sub-classified according to whether they represent perceived upcoming difficulty, for example as a speaker searches for the next word of an utterance, or whether they instead reflect an attempt to re-establish fluency following hesitation or repair (Heike, 1981).

2.3 When Are Disfluencies Produced?

As mentioned in the previous section, disfluency can originate at various stages throughout the speech production process. The type of disfluency produced depends to a certain extent on the stage at which the speaker detects their error. If errors are detected after the speech has been articulated then the necessary repairs will be overt. If speakers detect problems in speech which has not yet been articulated, then they may initiate a covert repair. This repair may involve suspending fluent speech until they can access the conceptual meaning they wish to communicate, or until a correct speech plan is ready to be articulated. If a speaker suspends fluent speech, they may produce a repetition or hesitation while they re-formulate their speech plan (Levelt, 1983). This interpretation relies on the speaker's self-monitoring process having access to the original internal speech plan (Levelt, 1983).

Although disfluency can originate at different points in production, there are certain patterns to when disfluency tends to occur in natural speech, and these will now be briefly described.

Difficulty at the conceptualisation of an utterance can lead to speakers becoming disfluent. This is, perhaps, unsurprising. Speakers rate themselves as less likely to know the correct answer to general knowledge questions when they have been disfluent (Hart, 1965) and indeed, filled pauses are more common in incorrect answers to general knowledge questions than in correct answers (V. Smith & Clark, 1993). Schnadt and Corley (2006) sought to experimentally manipulate conceptualisation demands by asking participants in a Network task (Levelt, 1983) to describe the route of a marker moving between pictured objects. Making it more difficult for participants to access the names

of these objects (by visual blurring of the images) increased participants' rates of disfluency, particularly with regard to hesitations and prolongations.

Hesitations and disfluency can also occur when the speaker has difficulty formulating an utterance. This can particularly be the case when the speaker's cognitive load is high. A number of factors have been demonstrated to increase disfluency rates, including lexical choice and syntactic burden.

Lexical choice was implicated by Schachter, Christenfeld, Ravina and Bilous (1991), who recorded the rate of disfluencies produced by lecturers in humanities, social sciences and natural sciences when discussing their fields of expertise. The lowest rate of filler production was found for lecturers from natural sciences, whilst the highest was observed in humanities lecturers. Importantly, when interviewed on general topics, there was no difference in their rates of disfluency. A later corpus study confirmed the assumption made by Schachter et al. (1991) that there exists a greater range of linguistic options for discussing the humanities than social sciences, and that in turn social sciences has a greater range of linguistic options than do the natural sciences (Schachter, Rauscher, Christenfeld, & Tyson Crone, 1994). Taken together, these studies support the idea that disfluency rates are higher when speakers have more options for expressing their ideas.

Difficulty in selecting words can also lead to disfluency when a word is harder to access due to having low lexical frequency (Levelt, 1983) or when context makes the word improbable. This effect of contextual probability is seen even when lexical frequency is controlled (Beattie & Butterworth, 1979).

Syntactic burden is also implicated in giving rise to higher disfluency rates stemming from formulation difficulty. Fillers occur more often before complex syntactic structures (Maclay & Osgood, 1959), and are most likely to occur at phrase and clause boundaries

(Boomer, 1965; Hawkins, 1971). If, as suggested by Butterworth (1975), new idea units need syntactic formulation at the point where they are introduced, then the prevalence of fillers at these points in an utterance indicates cognitive load as a cause of difficulty in formulation, which is expressed as disfluency.

Is disfluency intended as a signal?

The previous section briefly outlined some of the situations in which disfluencies are often produced. Now, I will move on to consider one of the more contentious claims in disfluency research; the idea that disfluency is produced deliberately as a communicative signal to listeners.

Speakers tend to be more disfluent in dialogue than monologue situations and when addressing other humans compared to addressing machines, for example answer-phones (Oviatt, 1995). Why this should be is not entirely clear, if disfluency is simply a symptom of production difficulty. This has led to the suggestion that speakers produce disfluency intentionally. This would be to claim that disfluencies are not so much symptoms of difficulty, as communicative signals to the listener.

One way in which fillers may serve as communicative signals is to maintain control of the floor. This motivation for fillers was proposed by Maclay and Osgood (1959), who suggested that a speaker pausing long enough to detect their own silence will produce a signal (such as a filler) to indicate that he is still in control, and does not wish to be interrupted. This view has been developed further by Goffman (1981), who proposes that a speaker momentarily unable or unwilling to produce a desired word or phrase will give audible cues that he is engaged in "speech-productive labour". In line with these views, disfluency would be considered to be part of a collateral message, providing supporting

information to assist the listener in their interpretation of the primary message and allowing a speaker to comment on their own performance.

Another approach to the *disfluency-as-signal* view is to regard fillers as words in their own right. This claim is based largely on their apparent adherence to phonetic, semantic, structural and prosodic rules. This interpretation of disfluency has primarily been put forward by Clark and Fox Tree (Clark, 1994; Clark & Fox Tree, 2002; Fox Tree & Clark, 1997), who point out that fillers are not wholly automatic in their production; speakers do have some control over them, and different fillers are produced in different situations. As the focus of this thesis is to investigate the effects of disfluency on listeners, the question of speaker-intentionality does not contribute significantly to the discussion, and so will not be addressed further.

2.4 Are Listeners Affected by Disfluency?

One might reasonably assume that disfluency should cause significant problems for listeners, who are required to filter it out of speech to recover the fluent, intended message of the speaker. This is especially the case if one thinks of speech comprehension as analogous to Artificial Speech Recognition (ASR) programmes, which are badly disrupted by disfluency. Human speech comprehension is, however, somewhat more subtle and sophisticated than ASR technology, and consequently, the way listeners respond to disfluency is somewhat counter-intuitive.

Evidence shows that rather than being badly disrupted by disfluency, listeners are in fact very poor at (consciously) detecting and reporting it. Lickley (1995) asked participants to listen to spoken monologues containing naturally occurring disfluencies. Listeners were provided with transcripts of the monologues, from which the disfluencies had been

removed, and required to follow the transcripts as they listened to the monologues, marking the points where the two differed. Even though participants simultaneously followed the transcript and the spoken material, they identified only half of the filled pauses (55%). Moreover, the detection rate for filled pauses within sentences was even lower, at 35%.

Despite participants' poor performance at reporting the locations of disfluency, it seems that they are not impervious to it. Disfluency has variously been shown to affect meta-judgements about the speaker (Brennan & Williams, 1995; Fox Tree, 2002; Christenfeld, 1995), syntactic parsing (Bailey & Ferreira, 2003; Maxfield, Lyon, & Silliman, 2009), lexical access and referent resolution (Arnold, Tanenhaus, Altmann, & Fagnano, 2004; Arnold, Hudson-Kam, & Tanenhaus, 2007; Brennan & Schober, 2001), semantic integration (Corley, MacGregor, & Donaldson, 2007; MacGregor, Corley, & Donaldson, 2010) and listener attention (Collard, Corley, MacGregor, & Donaldson, 2008). Disfluencies have also been shown to affect subsequent memory, both for lexical items (Collard et al., 2008; Corley et al., 2007; MacGregor et al., 2010) and at the discourse level (Fraundorf & Watson, 2011). Clearly there is considerable evidence that disfluencies are not simply ignored or edited out of speech before it is processed.

As disfluency's effect on the listener will be the main subject of this thesis, these effects will now be discussed in further detail.

2.4.1 Disfluency and Judgments of the Speaker

Smith and Clark (1993) demonstrated that speakers were more likely to be disfluent when they were unsure of the answers to general knowledge questions. Listeners are probably sensitive to this pattern, as demonstrated by Brennan and Williams (1995).

Listeners were asked to explicitly rate their confidence in general knowledge question answers which had been manipulated to be either fluent, or preceded by filled or silent pauses. Answers which were preceded by disfluency, were rated as less reliable than fluent answers, with filled pauses affecting confidence more than silent pauses. Filled pauses can also lead listeners to believe that speakers are being evasive and dishonest (Fox Tree, 2002), and these disfluencies can cause courtroom defendants to give a stronger impression of being guilty than they would otherwise have done (Hosman & Wright II, 1987).

Perhaps more interesting, though, than the relatively intuitive finding that disfluency can affect listeners' meta-judgements about a speaker's state, is how disfluency affects the processing of speech itself.

2.4.2 Disfluency and Syntactic Structure

We have already mentioned that disfluency is more common at boundaries between syntactic units than within units (Boomer, 1965; Hawkins, 1971), and if listeners use this distribution pattern to guide their syntactic judgements, this may go some way to explaining the findings of Bailey and Ferreira (2003). Bailey and Ferreira made use of garden path sentences, which are a common experimental tool in psycholinguistic research, as they present a temporary syntactic ambiguity for the listener. For example, in the following sentence there is a temporary ambiguity:

After the kidnappers returned the princess decided to cancel the party.

The syntactic parser will initially assign *princess* a role as the object of the first clause. The parser runs into problems when it encounters the next word, *decided*, which is illegal

in the context created by the initial parsing of the first part of the sentence. In fact, to correctly parse this sentence, *princess* must be assigned the subject role for the second clause. Failure to initially select this dispreferred interpretation means that reanalysis is needed when the following verb is encountered. If this reanalysis is not successful, the comprehender will deem the sentence ungrammatical. The difficulty of this reanalysis can be manipulated by the presence of relative clauses or modifiers between the head noun of an ambiguous phrase and the disambiguating word, and this is known as the 'head noun effect' (Ferreira & Henderson, 1991).

The head noun effect can also be elicited by disfluent filled pauses. Bailey and Ferreira (2003) used filled pauses in the place of relative clauses and modifiers, and found no significant difference in the percentage of sentences judged grammatical in the fluent condition and when the filled pause was placed before the ambiguous head noun (83% and 85% respectively), but that far fewer sentences were judged grammatical when the filled pause intervened between the head noun and disambiguating verb (60%). Importantly, replacing the filled pauses with time matched non-speech interruption (e.g., dogs barking, doorbells ringing) elicited the same basic pattern, suggesting that in sentence parsing at least, the delay constituent of disfluency is enough to trigger a change of tactic by the listener.

2.4.3 Disfluency and Listeners' Predictions

Having discussed that listeners are sensitive to the relationship between disfluency and high cognitive load, and use this pattern to inform syntactic judgements and meta-judgements about the speaker, it is logical to ask whether this listener sensitivity extends to other speaker states co-occurring with disfluency. Specifically this section discusses the evidence that listeners make use of the relationship between disfluency speaker difficulty

resulting from lexical choice (Schachter, Christenfeld, Ravina, & Bilous, 1991), or items which are difficult to name (Schnadt & Corley, 2006), have low lexical frequency (Levelt, 1983), or low contextual probability (Beattie & Butterworth, 1979).

Arnold , Fagnano, Maria and Tanenhaus (2004) tested whether listeners were sensitive to a relationship between speaker disfluency and lexical access (how difficult an item is to name) by measuring listeners' eye movements while they followed instructions involving items which were either given (had previously been mentioned in the discourse) or were discourse-new (not previously mentioned). New items are more difficult for speakers to access than given items (Arnold, Wasow, Losongco, & Ginstrom, 2000), and so Arnold and colleagues hypothesised that if listeners were sensitive to this, they would use disfluency as a cue indicating that the upcoming item would be discourse new. Eye-fixations, which are believed to reflect lexical access, can be used to track the time course of continuous speech processing (Tanenhaus & Trueswell, 1995), and so the authors reasoned that if listeners were able to use disfluency to predict discourse-new items, this would be reflected in eye-fixations.

Participants followed auditory instructions to manipulate an array of four objects displayed on a computer screen. For example:

Now put the(e uh) candle below the grapes.

The results supported the idea that listeners used disfluency to predict upcoming items. Following hesitations, participants' gaze moved to items which were discourse-new, whereas in fluent conditions, they were more likely to fixate on previously mentioned items. Arnold et al. (2004) suggest that this may indicate that listeners are either subconsciously aware of the distribution of fillers, and that this guides their predictions for

the following portion of speech, or that they are aware of the types of problems likely to cause disfluency, and likewise, this guides their expectations.

Similar evidence for listeners' sensitivity to the relationship between disfluency and discourse-new items comes from Barr and Seyfeddinipur (2010). This mouse-tracking study found listeners to strongly anticipate items which were discourse new in utterances incorporating a filler (*um*). In an extension of Arnold's findings, Barr and Seyfeddinipur found that this effect was speaker specific; listeners' predictions were based on which items were new to an individual speaker within the discourse.

Further evidence for predictive processing in relation to disfluencies comes from Arnold, Hudson, Kam and Tananhaus (2007), who found participants to be biased to expect referents which are difficult to name following disfluency. Arnold and colleagues manipulated lexical access with regard to the referent itself, rather than its status in discourse. They achieved this manipulation by contrasting concrete objects (e.g., an ice cream cone) with abstract images (e.g., "a funny squiggly shape", sic p. 918). As in Arnold et al. (2004), participants' eye movements were tracked while they followed instructions relating to four objects presented on a computer screen in front of them. Two of the objects were familiar and concrete; two were abstract. Objects were presented in two colours, with one familiar and one abstract object in each colour. Participants followed instructions such as

Click on the(e uh) red ice-cream cone

As soon as the colour phrase was uttered, participants were able to rule out two potential referents, and so eye-movement analysis focussed on the period from when the

colour became clear. Eye movements revealed a bias towards the unfamiliar, difficult-to-describe shapes when auditory instructions were disfluent, whereas in the case of fluent instructions, listeners looked at the familiar and unfamiliar objects equally frequently.

To verify whether listeners' bias for familiar/unfamiliar objects was driven by their predictions about speaker difficulty, Arnold et al. (2007) repeated the experiment, but told half of the participants that the speaker suffered from object agnosia, making if difficult for them to describe familiar items. Under these conditions the bias for unfamiliar items under disfluent conditions disappeared, suggesting that the predictions listeners were making about upcoming items was driven by some level of top-down judgement of the disfluency in the auditory instructions. In a third experiment, Arnold et al. (2007) repeated the same procedure, but rather than telling participants that the speaker suffered from object agnosia, they spliced in segments of background noise (building noise, beeping, *etc.*) just before the prolonged definite article (*thee uh...*) to make it sound as though the disfluency was in response to the distracton, rather than difficulty for the speaker in describing the target. Results revealed that these noisy interruptions did not alter the pattern reported in the first experiment. Unlike the 'knowledge' that the speaker suffered from object agnosia, the indication that an external stimulus had disrupted speech did not affect listeners' tendency to expect abstract referents following disfluency. On the basis of the overall pattern of results, Arnold et al. suggest that listeners can only use top-down inferences to a certain extent. Another interpretation would be to consider that in reality, speakers may prefer to talk over such noise than to allow it to interrupt their flow, and if listeners are sensitive to this, they may continue to interpret disfluency as indicating conceptual or formulation difficulty.

Along with the results of Barr and Seyfeddinnipur (2010), Arnold et al. (2007) clearly demonstrate that the effect of disfluency on listeners' prediction is not simply to produce

a one-size-fits-all departure from the referent which is most accessible to the listener. Instead, these prediction effects are complex and sophisticated, with room for a degree of top-down contextual understanding.

The role of disfluency in guiding prediction has also been demonstrated using scalp-measured Event Related Potentials (ERPs), which are a continuous on-line measure of cortical activity and can be used to measure cognitive processes in the absence of a secondary task. ERPs are a central methodology for the work in this thesis, and a fuller expanation of ERPs can be found in Chapter 3. For now though, we will move on to consider how ERPs have contributed towards demonstrating that listeners' predictions are affected by disfluency.

The N400 is a language-related ERP component sometimes colloquially described as indexing linguistic surprise. It consists of a negative going voltage wave which is maximal over the centre and towards the rear of the scalp, and which peaks around 400ms after the onset of a linguistic stimulus. The amplitude of this negative going wave is sensitive to the ease of semantic integration of a target word into its context. The more difficult a word is to integrate, the larger the N400 amplitude. Take, for example, in the following sentences:

Everyone's got bad habits, I'm always biting my ***nails***

That coffee was too hot, I burnt my ***nails***

The word *nails* is much more difficult to integrate into the second sentence than the first. Measuring the ERP for this final word, a much bigger N400 effect would be expected for the word *nails* in the second sentence. A more detailed discussion of the N400 can be found in Section 3.5.2.

Corley, MacGregor and Donaldson (2007) examined changes in N400 amplitude as participants listenened to recordings of sentences such as those in the example above. These utterances were either fluent, or contained a disfluent filler directly before the final (target) word. Technical constraints meant that it wasn't possible to directly compare the ERPs for fluent and disfluent utterances; instead they measured the difference in N400 between predictable and unpredictable words, and compared this difference across fluency conditions. They reported that the size of the N400 effect (difference between predictable and unpredictable) was reduced for targets preceded by a filler, indicating that the perceived difference in contextual fit of predictable and unpredictable words was attenuated by the disfluency (See Figure 2.1). A similar effect was reported when the experiment was repeated using silent pauses instead of fillers before target words (MacGregor et al., 2010). For the sake of clarity, it is worth mentioning that due to the technical limitations which prevented the direct comparison of ERPs from fluent and disfluent utterances, it is not possible to determine whether the presence of disfluent fillers made unpredictable words easier to integrate or predictable words more difficult. Instead, these results should be thought of as demonstrating that unpredictable and predictable words were treated more similarly when they were preceded by a disfluency.

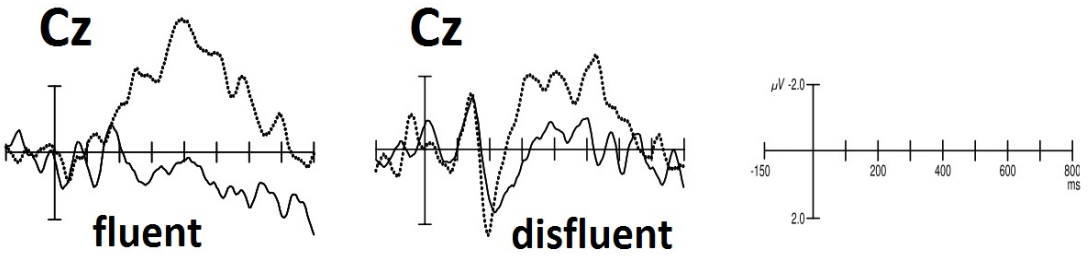

FIGURE 2.1: Figure adapted from Corley et al. (2007), showing the N400 effect at the CPz electrode for fluent (left) and disfluent (right) utterances. ERPs to predictable words are shown with a solid line; ERPs to unpredictable words with a dotted line. Negative is plotted upwards. Fluent utterances elicit a larger N400 effect than disfluent utterances, indicating that the processing of predictable and unpredictable words was more similar following a disfluency than in fluent sentences.

2.4.4 Disfluency and Listeners Attention

In addition to precipitating changes in listeners' expectations for upcoming speech, disfluency appears to have an effect on their attention. A link between disfluency and attention was suggested by Fox Tree (2001), who noticed that reaction times to targets in pre-recorded speech varied according to whether targets were preceded by an *uh*, an *uhm*, or whether these fillers had been excised. Reaction times were found to be faster for targets preceded by an *uh* than where the *uh* had been removed. This advantage was not found for utterances containing an *uhm*. Fox Tree posited that this discrepancy may arise because *uh* usually signals a short delay, causing listeners to 'heighten their attention', whereas *uhm* signals a longer delay, and in this circumstance, such heightened attention is not beneficial. An alternative interpretation might be to suggest that *uh* and *uhm* both heighten attention, but that this attention decays in the longer pause following an *uhm*.

Before going on to discuss the effects of disfluency on attention any further, it is useful to define what is meant by attention. The concept of *attention* is widely accepted and used in everyday-life, and perhaps because, rather than in spite of, this, attention is not always clearly defined in the psychological literature. There is a reasonable crossover, however, between the lay definition and psychological definition of attention. This crossover is useful in that it bounds what authors mean by *attention* in publications (such as Fox Tree, 2001; see above) where authors do not specify whether they are using a strict psychological definition or a lay definition. Attention generally refers to a selective and finite psychological resource (Pashler, 1998b), which at any one time is assigned to particular stimuli or events (Pashler, 1998a). Among other things, the allocation of attention enhances a person's awareness of, and memory for, items concerned,

giving these items a psychological priority over other, competing items. Items to which attention can be allocated include stimuli, external activities and internal cognitive processes. Attention may be voluntarily directed towards these items, or grabbed by items which are particularly salient (Collard, 2009).

Of particular relevance to the studies pertinent to this thesis is how attention interacts with auditory speech processing. Studies of attention orienting in speech processing have sometimes used a paradigm in which a participant is required to attend to one speech stream whilst ignoring another. Manipulations are then made in the distractor stream to see what causes participants to lose their focus on their 'target' stream. Loss of focus on the 'target' stream is interpreted as indicating that attention was grabbed by the competing distractor stream. With the exception of the participant hearing their own name in the competing stream (Moray, 1959), attention is more likely to be grabbed by changes to the physical properties of the stimulus than the content. Whilst participants are unlikely to notice the speech in the competitor stream being played backwards (Cherry, 1953), their attention is affected by a change in pitch (Cherry, 1953), or by a language change (Scharf & Buus, 1986), which presumably significantly alters the prosody of the stream.

A specific investigation of attention and disfluency was made by Collard, Corley, Macgregor and Donaldson, (2008), who demonstrated that fillers modulate attention using an auditory oddball paradigm within an ERP experiment. Participants listened to a set of sentences, half of which contained a filled pause before the final, highly predictable target word[2]. Half of the target words were artificially compressed, by amplifying the mid-range frequencies, making them sound similar to speech through a poor quality telephone line, and so acoustically deviant from the rest of the recording. Acoustically

[2]The sentences and auditory recordings used in Collard et al. (Collard et al., 2008) were the same recordings as were used in Corley et al., (Corley et al., 2007).

deviant targets in fluent sentences elicited a Mis-Match Negativity (MMN) ERP component, which is associated with change detection (Schroger, 1997), followed by a large P300 effect, thought to index the orientation of attention (Polich, 2004). Crucially, in disfluent sentences, incongruous stimuli elicited an MMN, indicating that listeners were aware of the acoustic mis-match, but the P300 effect disappeared. To explain the pattern of ERP effects, Collard et al., reasoned that attention had already been engaged by the disfluent filler *er* in the stimuli, and so there could be no further orientation of attention to the deviant target.

Further examples of the influence of disfluency on attention can be seen in studies which demonstrate advantages for disfluency, even when the disfluency is not predictive of upcoming material. For example, Corley and Hartsuiker (2011) asked participants to follow auditory instructions and press buttons corresponding with images on a computer screen. They found that the *ums* speeded responses to all targets, not just less accessible targets; here - those with low lexical frequency. The increased reaction speed for all items preceded by *um* suggests that attention was heightened following *um*, but no specific predictions about the referent were made. However, one thing to bear in mind when assessing the results of this study is that the conditions of this experiment may have changed the way that listeners processed the stimuli. This experiment used the same utterance and the same acoustic tokens over and over, changing only the target word, whereas the previously described experiments used more naturalistic stimuli. The repetition of stimuli may have allowed participants to make specific predictions about the stimuli. Additionally, in fluent utterances there were three possibilities directly before the target word; the easier target, the more difficult target, or a disfluency. Hearing a disfluency consequently rules out one of these possibilities, meaning that participants could be more prepared to respond to the target word, knowing it would be the next

token in the sentence.

2.4.5 Disfluency and Listeners' Memory

Disfluency does not only affect listeners' immediate, on-line language processing, but appears also to have lasting consequences. Given that attention is necessary for successful memory encoding of stimuli (*c.f.* Craik, Govoni, Naveh-Benjamin and Anderson, 1996; Mulligan, 1998; Chun and Turk-Browne, 2007), it is not altogether surprising that disfluency, which has consequences for listener attention, also has some bearing on memory.

Following disfluent discourse, items are more likely to be subsequently remembered if at presentation they were affected by filled pauses (Corley et al., 2007) or silent pauses (MacGregor et al., 2010). These studies employed surprise memory tests following listening tasks to ensure that participants were not using memory strategies to prepare for a test, and so provide an interesting insight into the way disfluency affects memory for speech which listeners have heard and comprehended in a naturalistic, content-focussed manner. Interestingly, no memory benefit is found for words affected by disfluent repetition (MacGregor, Corley, & Donaldson, 2009), indicating that not all types of disfluency affect the listener in the same way.

A different approach to investigating the relationship between disfluency and memory was adopted by Fraundorf and Watson (2011), who also investigated the effect of disfluency on memory. Fraundorf and Watson focussed on how disfluent fillers might modulate memory at the discourse level, by requiring participants to retell previously heard excerpts of a story. Their results showed that participants benefited from disfluent fillers,

and were more likely to remember plot points which had been disfluent than those which had been fluent, or affected by coughs.

Taken together, the results of these studies indicate that disfluencies have consequences for listeners lasting well beyond immediate processing effects.

2.4.6 Summary of the Effects of Disfluency on Listeners

Listeners are sensitive to speaker disfluency on a number of different levels. Disfluency affects judgements about the speaker, the level of attention allocated to speech, predictions about upcoming speech, judgement of syntactic structure and memory for words and concepts affected by disfluency. Despite this, listeners perform very poorly when asked to explicitly identify disfluency, indicating that these immediate and lasting consequences of disfluency happen without participants necessarily being aware of disfluency at the surface level of utterances.

The natural progression from having identified these effects is to enquire why these effects are found, and what drives them. It is here that discussion will now turn.

2.5 What Drives the Effects of Disfluency on Listeners?

The specific mechanism by which disfluency affects comprehension and subsequent memory remains unclear. Across the disfluency literature, three broad accounts of disfluency processing emerge, and evidence for each will be presented below. These three accounts can be described as the "predictive processing hypothesis", the "attentional orienting hypothesis", and the "delay hypothesis". Briefly, the predictive processing hypothesis

states that listeners use either perspective taking, or their prior experience of the distribution of disfluencies, to predict that upcoming material will be discourse-new or in some way problematic for the speaker. An attentional orienting account, conversely, states that disfluencies raise the level of attention oriented to the speech stream, resulting in faster responses to any material heralded by a disfluency, regardless of whether the disfluency is predictive of problematic material. The third account usually states that any delay in the speech stream allows comprehension processes, including those at the discourse level, to unfold, leading to a processing benefit for the listener. It may be possible to propose a fourth account, based on a cooperative view of language use, as suggested in Section 2.3. In this case, listeners would respond to disfluency based upon an inherent understanding of disfluency itself (as a signal or word) rather than based upon sensitivity to the distribution of fillers in speech.

Although these three accounts of the effects of disfluency are often explicitly or implicitly presented as competing (*c.f.* Fraundorf and Watson [2011], for a recent example), careful consideration suggests this to be a false contrast. To suggest that only one of these accounts can explain the results of the body of studies on disfluency is to ignore much of the data. It is not clear that these accounts are mutually exclusive. For example, the predictive processing account may depend on raised attention to allow changes in predictive processes, or attention raising may depend on delay. It seems entirely possible, that a combination of these three accounts would best account for the variety of results found within disfluency research. Whilst it seems apparent that attentional and predictive changes do occur as a result of disfluency, the important question appears to be rather whether it is the form of the disfluency, or simply the delay it introduces which triggers these effects.

Before continuing, it is important to define what is meant by the *form* of a disfluency

in this context, as the precise meaning will be pivotal to the arguments discussed in this thesis. I use the word *form* to refer to the specific indicator of speaker difficulty. Indication of speaker difficulty may be gleaned from disfluent fillers, prolonged syllables or anything else which indicates that the speaker is having trouble producing their intended utterance. This definition stands in contrast to the extra time, or *delay*, that disfluency may add to that same utterance. A number of studies, some of which are previously mentioned in this chapter, have provided evidence on both sides of the form versus delay debate.

2.5.1 Is Delay Sufficient?

One account for the effects elicited by disfluency states that it is the delay which disfluency adds to an utterance driving the observed effects. This hypothesis states that any interruption, be it a filler, a silent pause, or simply noise should allow comprehension processes to unfold, leading to processing advantages for the listener. As is highlighted below, however, evidence in support of this hypothesis is somewhat mixed.

Support for the delay account is claimed by Brennan and Schober (2001), who demonstrated a disfluency advantage for reaction time in a referent selection task. This advantage persisted in cases where the speaker interrupted themselves with a disfluency, but then continued with the original material, rather than with a corrected and altered version of the pre-disfluency material, *e.g.:*

Move the the yel — uh — yellow square.

Listeners benefited from the disfluency, even when it did not predict the upcoming referent by ruling out the previously mentioned alternative. This is not surprising,

given that disfluency raises attention to the speech stream (Collard et al., 2008), which one might assume to speed reactions to all targets. Importantly, silent pauses were as effective as fillers in speeding reaction time. The authors interpret the finding that fillers and silent pauses both speed reactions as indicating that the form is less important than the delay it brings. It should, however, be remembered that silent pauses may themselves be interpreted as a disfluency, associated with planning difficulties (Maclay & Osgood, 1959), and if the control delay condition can also be interpreted as an explicit signal of speaker difficulty, then this study does not answer the question of what triggers disfluency effects. The difficulty in selecting an appropriate control condition goes some way towards explaining the mixed nature of the evidence in favour of a disfluency-as-delay hypothesis.

Corley and Hartsuiker (2011) attempted to address this by comparing filled and silent pauses to an artificially created sine-wave beep filled interruption in a referent selection task. Referents were images of objects whose names had either high or low lexical frequency, and low frequency images were blurred, making them significantly more difficult targets than high frequency items. Reaction time was faster for easy targets than difficult targets and faster for conditions where the interruption directly preceded the referent, compared to when the interruption occurred earlier in the sentence. This was the case regardless of the form of interruption. There was, however, no interaction between the position of an interruption and target difficulty. Participants did not gain any more benefit from interruptions occurring before hard-to-name targets than before easy targets. This would indicate that participants benefited from interruptions and disfluencies even when they did not occur in a typical disfluency distribution, and so did not allow participants to use the disfluent interruption to predict difficult-to-name referents.

Some reservations must be borne in mind, however, when extrapolating the findings of this study to speech comprehension more generally. Although it seems highly likely that some extra time will be beneficial to listeners, this study used utterances that were far from natural speech. In particular, the same acoustic tokens were used over and over again, which makes it seem reasonable to assume that listeners were not listening naturally for understanding, but simply waiting for the target word. Two possible target images were displayed on a computer screen while the stimulus utterances were played. In this context, therefore, it is possible that participants did not carefully study the images until they were required. Such a strategy would allow participants to avoid holding the targets in their working memory longer than necessary, and in this case, they would benefit from some extra time directly before hearing the target to allow lexical activation of the two possible items, unlike in natural conversation.

An alternative way to account for the delay advantage reported by Corley and Hartsuiker (2011) is to consider the acoustic properties of the stimuli. Because the utterances were created by splicing together tokens from separately recorded utterances, participants may have been aware of some jarring at the point where the utterances were spliced together, as a result of misleading co-articulation. Mis-leading co-articulation may have caused a delay in recognising targets in the fluent condition. In delay conditions, however, the extra time between recordings would have allowed phonetic expectations to subside, making for an easier transition into the target. ERPs have demonstrated this clash between phonological expectation and realisation to produce an effect at the neural level; a Phonological Mismatch Negativity (PMN), which is indeed attenuated when a delay intervenes between spliced parts of an utterance (MacGregor et al., 2010). It could be that this mismatch delays word recognition by increasing cognitive load, leading to the longer reaction times reported by Corley and Hartsuiker.

Whilst evidence from behavioural studies is difficult to interpret unambiguously, further support for the view that delay may be critical can be found in the combined findings of a number of ERP studies. Across the field, not all disfluencies appear to elicit the same "disfluency effects". Neither the reduction in the N400 effect reported by Corley (2007), nor the attendant memory improvement, were replicated when the same paradigm was used to investigate predictable and unpredictable words preceded by disfluent repetitions, rather than fillers (MacGregor et al., 2009). By way of disfluency, the utterances in this study contained repetitions of previously uttered speech. This indicated speaker difficulty, but did not provide any time period in which there was no linguistic input. In the absence of delay, disfluency effects were not found.

The equivalence between disfluent fillers and noisy interruptions in disrupting syntactic parsing has also been cited as evidence for the disfluency-as-delay hypothesis (Bailey & Ferreira, 2003). Two cautionary notes apply here though. If permitting the parser to linger on an incorrect parse allows the incorrect interpretation to solidify, as demonstrated by the Head Noun Effect (Ferreira & Henderson, 1991), then it is not clear that the effects reported in syntactic studies are in any way disfluency effects, and so the comparison of disfluency to noisy interruptions loses some of its currency. Additionally, Fraundorf and Watson (2011) question whether speakers would pause speech for such interruptions. It may be the case that speakers would not allow such interruptions to cause a break mid-phrase unless they were having production difficulties and were prepared to admit a disfluency to the speech-stream anyway. Thus, even if the parsing effects Bailey and Ferreira report are in some way disfluency specific, listeners' sensitivity to a speaker's unwillingness to pause mid clause may still facilitate correct parsing if the listener interprets the interruption as disfluency on the part of the speaker.

2.5.2 Does Form Dictate Function?

The effects of disfluency on prediction appear to be somewhat context driven. Listeners' use of disfluency to anticipate difficult-to-describe objects disappears when they believe the speaker has object agnosia (Arnold et al., 2007), and the use of disfluency to predict new referents is speaker specific (Barr & Seyfeddinipur, 2010). These finely tuned effects suggest that participants are making use of either perspective taking or an awareness of the distribution of disfluency to guide their expectations. This would seem to indicate that listeners are making use of their knowledge that disfluency indicates speaker difficulty, although one could say that there is no evidence to demonstrate that these same effects could not have been precipitated by delay. Unfortunately, however, the design of these experiments does not allow for a distinction to be drawn between form and delay.

Returning to Fox Tree (2001), this study makes some specific claims about the role of the form of fillers, suggesting that *um* and *uh* differ from one another in terms of their effects on listeners. In a pair of studies, one conducted in Dutch, and one in English, participants heard recordings of spontaneously produced speech, while monitoring for a target word, which they had seen on a computer monitor immediately prior to sentence onset. As previously mentioned in Section 2.4.4, *uhs* speeded responses but *ums* did not. Fox Tree proposes that it is the differing forms of these fillers which drives this difference. However, this study did not contrast *ums* and *uhs* with truly fluent utterances. The non- *um/uh* conditions were created by digitally excising the *ums* and *uhs*, and any silent pauses and co-articulatory features before and after the fillers remained. If there are systematic differences between the silent pauses heralded by *ums* and *uhs* then it may not be that the fillers themselves affect processing differently, but that the differing effects are caused by their differing "fluent" conditions. Additionally, the remaining

silent pauses and co-articulatory features may well have been sufficient to trigger a "disfluency response", even in the "fluent" conditions. All of these factors mean that the findings of Fox Tree (2001) cannot offer any conclusive evidence on the trigger required to elicit a "disfluency response".

Some stronger evidence that the form of a disfluency may play a key role in modulating memory, at least at the discourse level, can be found in the work of Fraundorf and Watson (2011). In a pair of experiments in which participants were required to retell excerpts from a previously heard story, fillers improved memory for plot points, whereas time matched coughs impaired memory, suggesting that the delay added by the disfluency was not the key feature in modulating memory.

One caution with regard to the Fraundorf and Watson (2011) study is that some inferences are made about the effect of disfluent fillers and delay on processing. These inferences rely on an assumption that there should be a link between easier, more facilitated processing resulting from the benefit conferred by appropriately distributed fillers, and improved subsequent memory (see Fraundorf and Watson, 2011, p. 172). An influence of disfluency on memory has been reported in a number of studies (Corley et al., 2007; Collard et al., 2008; MacGregor et al., 2010), but any assumption that this is a result of the greater ease of semantic processing seems to be flawed. Unpredictable words at the end of sentences are probably not facilitated by highly constraining sentences, such as those used in Corley et al. (2007) and subsequent similar paradigms. However, unpredictable words are better remembered than the words which are more easily processed (predictable) (Corley et al., 2007; Collard et al., 2008; MacGregor et al., 2010). Thus, it is important not to extrapolate the memory effects reported by Fraundorf and Watson (2011) to make assumptions about language processing.

2.6 Summary

Disfluency is a common phenomenon occurring throughout natural speech. It arises when the speaker encounters difficulties in conceptualising, planning or forming utterances. Listeners have been clearly shown to be affected by disfluency on a number of levels. In addition to affecting judgements about the speaker, disfluency brings about changes in listeners' attention and predictions of upcoming speech.

Nevertheless, it is not clear what aspects of disfluency drive these changes. One model might suggest that listeners make use of their own experience to take a speaker's perspective, or that they use their experience of the distribution of disfluency to respond appropriately to indications that the speaker is in trouble. An alternative model would propose that it is not the indication of speaker difficulty which is key, but simply that the delay, added to an utterance when a speaker becomes disfluent, allows listeners' comprehension processes to unfold more fully. Whilst some studies have attempted to address this issue, and others have provided incidental evidence to support one model or the other, a satisfactory explanation remains elusive, and it is this question which the experiments in this thesis seek to address.

Chapter 3

Introduction to Event Related Potentials (ERPs)

3.1 Introduction

Eighty-five years ago, the then controversial claim was made that the electrical activity of the human brain could be measured using an electrode placed on the scalp and recording voltage fluctuations over time (Berger, 1929). It was a further six years before Berger's findings were confirmed by others, including influential physiologists (Adrian, Matthews, & C., 1934; Gibbs, Davis, & Lennox, 1935), leading to the acceptance of the electroencephalogram (EEG) as a real phenomenon. In the ensuing years, EEG has become a key tool in cognitive neuroscience. Although EEG does not give high spatial resolution (which would allow researchers to identify specific areas of the brain responsible for various processing activities), modern recording techniques provide excellent temporal resolution, which affords researchers the opportunity to continuously track tiny changes in cognitive processing. This chapter will give a brief overview of the

use of EEG in cognitive neuroscience, beginning with the neural origins of EEG, and moving on to consider best practice in data collection and methods of analysis, before outlining some of the ERP components most relevant to the experiments reported in this thesis, and the studies which informed their design.

3.2 Neural origins of EEG

3.2.1 The neuron

Although the brain is divided into distinct regions, each responsible for various tasks ranging from the basic maintenance of life through to complex cognitive processing, the most important structures throughout the whole brain are among the smallest; the neurons. To understand what we are measuring when we record EEG, and hence how best to use it, it is helpful to have a basic understanding of neurons, and how the voltage we measure at the scalp is generated.

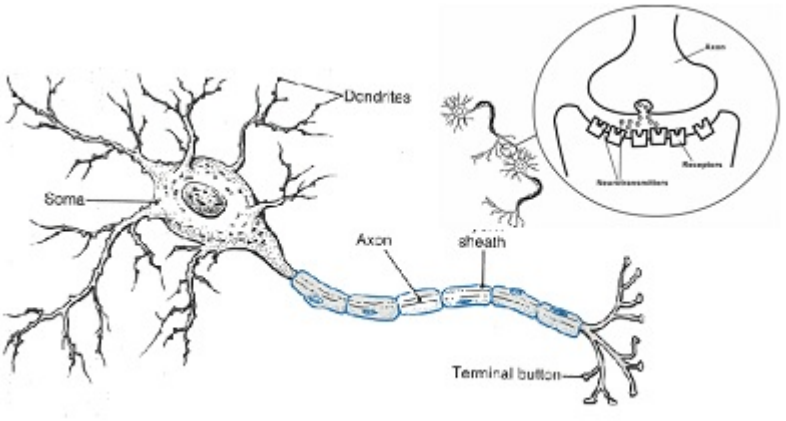

FIGURE 3.1: The basic structure of a typical neuron, comprising an axon, a cell body (soma) and dendrites; and a synapse. Neurons form the functional and structural basis of the nervous system (Cajal, 1909). Information is transmitted between adjacent neurons at synapses. Here, neurotransmitters are released from the terminal buttons to cross the tiny synaptic gap to the dendrites of the next neuron. Diagram adapted from Carson (1992) and http://en.citizendium.org/wiki/File:Synapse.gif.

In their resting state, neurons maintain a negative voltage gradient across their membranes, with the inside of the cell having a negative charge of 60-70mV. Neurons have a semi-permeable membrane, made up of two layers of fat molecules, with larger protein cells embedded in the fat layers. While oxygen, carbon dioxide, urea and water can freely cross the membrane, it is usually impermeable to larger or electrically charged molecules and ions. Importantly, however, the membrane is able to "pump" sodium ions [Na^+] and potassium ions [K^+] against their concentration gradient. [Na^+] is found in higher concentration outside of the cell, while [K^+] is found in higher concentration inside the cell. These ions naturally want to flow from areas of high concentration to areas of low concentration. [K^+] tends to diffuse out of the cell, although it is returned into the cell at the same rate by the sodium-potassium pump, which simultaneously removes [Na^+] from the inside of the cell. [Na^+] is attracted not only by the lower concentration, but also by the negative charge of the inside of the cell. Specialised gates allow ions (particularly sodium [Na^+], potassium [K^+] and chloride [Cl^-]) to pass freely into and out of the cell, but in the cell's resting state, these gates are kept closed. Hence sodium ions are kept out, chloride ions are kept in, and potassium ions are returned into the cell at the same rate as they diffuse out, and the cell maintains its interior negative charge.

Neurons produce two main types of electrical activity, action potentials and post-synaptic potentials. Action potentials are voltage spikes, propagated along the axon to the terminal buttons, and are generated when a stimulus causes sodium [Na^+] gates to open, and sodium ions from outside the cell flood in. This incoming flood of Na^+] reduces the negative charge inside the cell, and so depolarises the cell membrane. Only if this depolarisation reaches a critical threshold level is an action potential generated. The action potential travels along the axon as a wave of depolarisation, which lasts approximately 1ms and is followed by a wave of hyperpolarisation (in which the inside of the cell is

more negative than in its resting state). When the action potential reaches the terminal buttons, at the very end of the axon, it causes the release of neurotransmitters; chemical molecules are capable of crossing the synaptic gap (0.02 microns) to the dendrites of the next neuron (known as the post-synaptic neuron). There, the neurotransmitters bind to large protein molecules, known as neurotransmitter receptors. This binding causes the opening or closing of ion channels (depending on the type of neurotransmitter and synapse) which leads to a change in potential across the membrane of the post-synaptic cell. This potential may last from tens to hundreds of milliseconds, and is known as a post-synaptic potential. Post-synaptic potentials occur almost instantaneously, and are largely limited to the dendrites and the cell body, rather than travelling down the axon. Crucially, the difference in potential between the inside and the outside of the cell forms a tiny dipole (equal and opposite charges, separated by a distance). If this post-synaptic potential is a de-polarisation of the membrane, and it reaches the critical threshold level, an action potential will be generated.

Action potentials cannot usually be measured at the scalp, as they are so small in size and transient. Neurons very rarely fire at exactly the same time, and so the waves of de-polarisation running along the neuronal axons can be considered to be out of phase with one another, and so do not summate but cancel each other out. By contrast, the dipoles of post-synaptic potentials have a longer duration, and summate more easily, making them detectable with scalp electrodes. For this to happen, the voltages of thousands or millions of neurons must summate, which means they must occur at approximately the same time and in neurons which are spatially aligned in an open field, usually in parallel. The signal measured in EEG is generated primarily in cortical pyramidal cells, where neurons are aligned perpendicular to the surface of the cortex (Kutas & Dale, 1997), although the cortex is not flat, but highly folded, so that even if neurons are beautifully

aligned relative to the surface of the cortex, they are not necessarily aligned with reference to the scalp, and as such only a small portion of cortical activity is detectable using scalp-measured EEG. In a closed field, the cell bodies are clustered together, with their dendrites extended outwards in various directions. Because the dipoles generated by these neurons are not aligned, they cancel each other out, meaning that no potential can be measured outside of the structure. This means that activity in regions of the brain with this type of structure (such as the midbrain nuclei or the thalamus) cannot be measured using EEG. As such, EEG represents only a select range of neuronal activity of the brain, and so failure to detect differences in scalp-measured EEG between experimental conditions does not necessarily imply identical neuronal activity.

3.2.2 Volume conduction

Electrical activity (current) from summating dipoles propagates through the various tissues within the head until it reaches the surface. The flow of current is affected by its conductive medium (i.e., the brain, intracranial tissues, fluids), and is particularly affected by the skull, whose high resistance causes a large degree of lateral spreading. Current follows the path of least resistance, so its path is also determined by the shapes of the structures within the head, particularly the brain and the skull. This means that voltages generated in one part of the brain may produce voltages on quite distant parts of the scalp.

Given all the relevant information about a set of dipoles within the head (i.e., their locations, polarities, magnitudes) and the correct algorithms to calculate their propagation, it is relatively easy to calculate the voltage which will be measured on the scalp. This is known as the *forward problem*. However, this is not reversible, hence it is not possible to provide a definite answer to the *inverse problem*, as there are an infinite number of

arrays of dipoles which could produce a given voltage pattern observed at the scalp. For this reason, EEG is poorly suited to questions regarding the precise locations of neural generators.

3.3 Recording EEG

3.3.1 Equipment and Set-Up

Electrode Placement

Voltage, or potential difference, is not a value that can be measured at one site, as it represents the difference in electrical potential between two sites. Therefore, when voltage is measured at an electrode site, this is really the difference in potential between that site and another. However, a simple subtraction along the lines of *potential(A) - potential(B)* will produce a set of values value which also include any environmental noise. To counter this, two sites are chosen against which to measure the recorded voltages. These are known as the reference and the ground, which together provide the 'baseline' or 'background' against which voltage at a given electrode is recorded. It is important to be aware of the choice of reference and ground when assessing any EEG data because of the impact that they have on the recorded data. Changing the position of the electrode and ground electrodes will change the distribution and may change the apparent polarity of measured effects, as the voltage measured at each electrode is not meaningful as an absolute, but rather represents the potential difference between that electrode, and the electrodes selected to act as a ground and a reference.

The ground electrode is usually placed somewhere on the participant's body, assumed not to be affected by neural activity, but which will experience the same level of electrical

noise as the scalp. Throughout the experiment, voltages are recorded from these two sites, and one subtracted from the other. It is assumed that any electrical activity to which they are both subject must be environmental noise, not neuronal activity, and so this provides a new baseline against which to measure the activity at the active electrodes on the scalp. Common ground sites include the mastoids (the bony projection behind the ears), the earlobes, and the end of the nose. Many researchers prefer to use the mastoids, as ear clips and nose electrodes may become uncomfortable and distracting. To avoid any hemispheric bias when mastoids are used, one mastoid is used as a ground electrode and one as an active electrode during recording. An average of the two recordings is then constructed off-line (post-recording), to give a new ground, or baseline, against which recordings from the other active electrodes can be measured (for a fuller explanation, see Luck, 2005).

An agreed system for the placing and naming of electrodes makes it possible to compare data across labs. The most commonly used systems are the International 10/20 system, developed in the 1950s (Jasper & Carmichael, 1958) and the Extended International 10/20 system (American Electroencephalographic Society, 1994). These divide the head into lines of latitude and longitude, and electrodes are placed at 10% and 20% positions along these lines.

Electrode Construction

As electrodes corrode, their impedances change, making them unsuitable for accurate data collection, and so EEG electrodes are designed to have low impedance and corrode very slowly. Most research uses either tin electrodes, or silver electrodes covered in a layer of silver chloride, which are the type used for the experiments in this thesis.

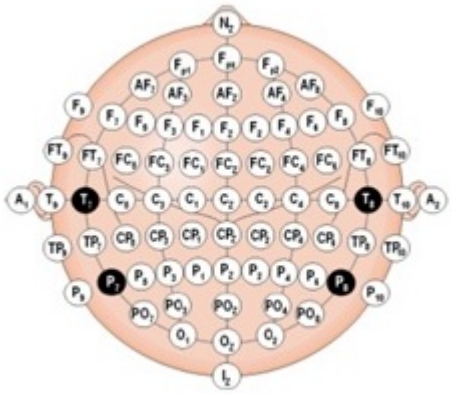

FIGURE 3.2: The Extended International 10/20 system (left) for the arrangement of electrodes on the scalp. Adapted from American Electroencephalographic Society (1994).

Reducing Impedance

Electrical current always follows the path of least resistance, so impedance cannot be allowed to vary across electrodes and is usually reduced as far as possible. In the experiments reported in this thesis, electrode impedances were reduced to below 5kΩ before recording, and monitored throughout the experiment to ensure they did not rise. To reduce impedance, electrode sites were gently abraded to remove the outer layer of dead skin cells, and a conductive gel was used between individual electrodes and the scalp. This also reduces skin potentials (the potential difference between surface- and deep skin layers) which manifest as low frequency noise in the signal. Reducing impedance is important, because failing to do so makes it more difficult for the amplifier to effectively eliminate environmental noise.

Reducing Noise in the Signal

Collecting good EEG data means collecting clean EEG data. When looking at any individual trial, the signal of interest will usually be obscured by unrelated neural activity, and environmental noise (electrical signals in the testing environment, as well as other biological electrical activity). Although averaging together trials does go some way towards extracting the signal from the noise, it is insufficient simply to rely on averaging as the signal to noise relationship is not linear, but increases as a function of the square root of the number of trials.

One of the greatest sources of noise in EEG is the participant themselves. This can be reduced by ensuring that participants are comfortable, and explaining to them the importance of sitting still and relaxed. All participants in the experiments reported in this thesis had the opportunity to see how their own EEG was affected by muscle activity, particularly ocular and facial movement. Additionally, the testing chamber was kept at a comfortable temperature, with air conditioning used to cool the room between testing blocks if necessary. This procedure is particularly important because changing skin potentials resulting from sweating can cause huge fluctuations in voltage, resulting in many trials being rejected at the processing stage.

3.3.2 Amplifying, digitising and filtering

The voltages between the active electrodes, reference and ground are measured as an analogue signal, which is first amplified, then filtered to attenuate frequencies outside of a given range. As with all waves, EEG can be broken down into a series of sine waves of different frequencies. Most of the EEG relevant to questions of cognitive neuroscience occurs between frequencies of 0.01Hz and 30Hz, and so EEG is typically filtered to

attenuate frequencies below 0.01Hz, which often represent very slow potential shifts as a result of participant sweating or changes in electrode impedance; and frequencies above 80Hz, which are likely to contain a significant portion of muscle-related electrical activity. Not all undesirable noise occurs outside of the 0.01Hz—30Hz range. For example, alpha waves, which are typically associated with participants feeling fatigued or relaxed, have a frequency of 8Hz—10Hz. Thus, although alpha contributes significant high amplitude noise to the signal, it is difficult to remove, as applying a filter at this frequency would also remove parts of the signal of interest. Filtering is not entirely unproblematic, and given that filtering removes some of the data, filtering can lead to apparent changes in the EEG data, such as introducing artificial oscillations, or changing the apparent on/off-set times of ERP components. However, filtering is an important part of the EEG collection process, and facilitates the extraction of small signals from often noisy raw data. Careful choice and reporting of filters helps to counter some of the potential difficulties mentioned above.

Finally, the data are digitised at a given resolution and sampling rate. The resolution refers to the number of discrete values which can be produced over a range of voltages (e.g., 16bit resolution means 65536 discrete values). The sampling rate must be chosen to be at least twice the frequency of the highest frequency in the analogue signal to capture all of the data (Nyquist Theorem). A sampling rate which is too low will not only miss high frequencies, but will record them as low frequency artefacts (aliasing). For clarification, see Figure 3.6.

3.3.3 Offline Processing and creating ERPs

Event Related Potentials (ERPs) are EEG time-locked to the event of interest. They are formed by averaging together the EEG from multiple examples of an event; in

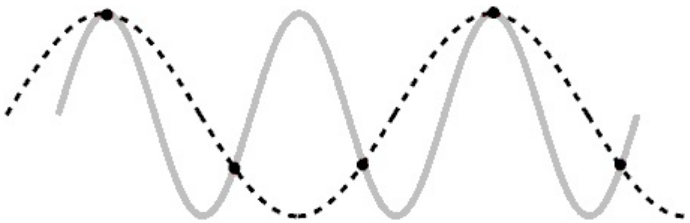

FIGURE 3.3: This figure demonstrates how samples taken at too low a sampling rate (less than twice per cycle of the highest frequency in the data) can result in incorrect interpretations of the frequencies present in the data. Joining the infrequent sampling points (represented by the black dots) on a real wave (in grey) gives the impression of a low frequency wave (dotted line) not present in the analogue data.

cognition studies, this is often the presentation of a stimulus. Raw EEG cannot be used to investigate the neuronal activity linked to specific cognitive events, because scalp measured EEG includes activity from unrelated, on-going neural processes, as well as noise, such as electrical activity from outside the brain. Assuming that this activity is random, or at least, not dependant on the event of interest (e.g., stimulus presentation), then it will tend to zero if a sufficient number of trials are averaged together. The activity which constitutes cognitive processing is assumed to be constant with regard to the event of interest, and hence will not be reduced by averaging.

To obtain an ERP for an event of interest, multiple trials relating to the event of interest are collected, and the data segmented into epochs. These epochs may often start at the presentation of the stimulus, but this is not necessarily the case. Depending on the aim of the experiment, epochs may, for example, be based on the timing of participant responses or particular features within the EEG. To create ERPs from a collection of epochs, data are averaged across epochs at each sampled point. Typically, this is done first within participants, so that an ERP for each condition of interest is created for each participant in the study, and then subsequently, these participant ERPs are averaged together to create grand averages for each condition of interest.

To allow comparison of ERPs across conditions, they are interpreted with regard to a

common baseline, often ERP activity in the 100-200ms period before the beginning of the epoch. As all data in ERPs is relative, absolute polarity and magnitude are not necessarily meaningful. Magnitude and polarity as represented on the ERP waveform depend on a number of factors. These include the electrodes chosen, and the location of the reference and the ground, which tend not to vary within experiments, as well as the underlying neural generators of the activity and activity in the baseline period, which may vary between experimental conditions. For this reason, a lot of ERP research focusses on the differences observed between experimental conditions, rather than absolute magnitudes within an ERP.

Because of the high noise to signal ratio and individual differences between participants, most ERP experiments (including the ones in this thesis) set a lower limit of sixteen trials per participant, per condition, for analysis. This means that unless a participant has contributed 16 good trials to each testing condition, their data cannot be included in the analysis. Sixteen trials is a widely accepted minimum within the field of ERP research, and is based upon the relationship between trial numbers and the signal-to-noise ratio. The signal-to-noise ratio increases as a function of the square root of the number of trials incorporated into an average, and so to double the signal to noise ratio from that achieved with 16 trials would require 64 accepted trials, which constitutes a significant increase in testing . However, significant advantages can be achieved by using a minimum of sixteen trials, compared to accepting participants with fewer trials.

One advantage of this procedure is that it prevents data from being skewed by individuals. Because of the high noise/signal ratio, and the high likelihood of losing data due to artefacts (e.g., amplifier saturation, drift, etc.), a large number of trials must be included in testing. It is not unusual to lose 30% of trials for these reasons. This means that ERP experiments tend to be quite long, and participants may become bored and fatigued.

This has to be borne in mind by researchers designing experiments, particularly when deciding how many experimental conditions to include, or tasks for the participants.

3.3.4 Artefact Correction

Ocular Artefacts

Even relatively clean EEG data requires some processing to minimise the contribution of artefacts, which are not removed by averaging. Here we outline the most significant source of artefact, due to eyeblinks. We also note, however, that in practice a key aspect of good data collection is the continuous monitoring of EEG during recording, which allows problems with artefacts to be identified and corrected on-line.

Muscular movement involved in eye blinks produces large artefacts, relative to neural EEG, which need to be effectively dealt with before ERPs can be created. There are various approaches to removing these artefacts, and each approach has its own problems. The simplest solution is simply to ask participants to keep their eyes focused on a fixation point and not blink during trials. In conjunction with this, it is possible simply to discard all trials containing an eye blink. However, this approach has a number of drawbacks. Firstly, it limits the populations who can be tested as some groups of subjects (e.g. children and some patient groups) cannot easily control their eye-movements, making it difficult to collect sufficient numbers of trials. Secondly, discarding all trials with eye blinks may lead to an unrepresentative sample. Finally, asking participants not to blink adds a separate cognitive task to that being tested, which may change the participants' responses to the experimental task. Fortunately, an alternative approach is available; it is possible to minimise the impact of ocular artefacts in the processing stage.

When the eyes are moved, or blink, voltages are created that are propagated to the scalp electrodes. This voltage can be measured as an electro-oculogram (EOG) by electrodes above and below one eye. Regression techniques, which assume a linear relationship between EOG and EEG, can be used to calculate the propagation factor between the EOG and EEG at each electrode. It is then possible to subtract the corresponding proportion of EOG from EEG at each scalp electrode. Propagation factors are estimated on an individual basis, derived from an average of blinks in the experiment. These regression techniques were used in the experiments described in this thesis. One drawback, however, is that the VEOG electrodes are also collecting neural EEG, so by deleting their influence, some neural EEG is also inevitably lost.

Voltage Drift, Amplifier Saturation and Muscular Activity

Trials affected by sources of contamination other than EOG are usually rejected from analysis. Typically, this means trials affected by voltage drift, amplifier saturation, and muscular activity. Voltage drift results from changes in skin or electrode impedance, usually caused by the participant sweating or electrodes moving. The amplifiers become saturated if the measured voltage exceeds their capacity. Muscular activity produces high frequency activity in the EEG. This can be reduced by ensuring the participant is comfortable and understands the importance of sitting still. Remaining muscular activity contamination can then be attenuated using a low pass filter on the amplifier.

3.4 Interpreting ERPs

ERPs represent a changing pattern of activity over time, and are interpreted relative to activity in a pre-stimulus baseline (often 100ms).

Traditional ERP analysis relies on analysing specific components, i.e., observable patterns of activity with specific polarities, timings and general scalp distributions[1]. To make sense of ERP waveforms, they are often directly compared to one another; as previously mentioned, absolute voltage and polarity are not necessarily meaningful. Any waveform represents contributions from all cognitive processes on-going at the time, so to isolate a component of interest, we rely on subtraction of waveform from another.

3.4.1 What, where and when

If a component demonstrates a quantitative difference in amplitude between conditions, this is interpreted as a difference in the degree to which underlying cognitive processes are engaged. Amplitude is commonly measured in one of two ways. One method is to use the *Peak Amplitude Measure*. The Peak Amplitude Measure involves identifying a time window of interest, and then finding the maximum amplitude within that window for each waveform being measured. The alternative *Mean Amplitude Measure* requires the calculation of the mean amplitude for each waveform within a pre-defined time window. The Mean Amplitude Measure is directly correlated to an area amplitude measure, which makes the mean amplitude measure easier to visualise.

Small differences in the timing of components between experimental conditions are assumed to represent temporal differences in the engagement of neural processes, although these timing differences can only provide an upper bound on the time point at which differences in processing occur. Other processing differences may have occurred earlier, but

[1]It is worth noting that ERPs can also be interpreted using frequency analysis. Neural activity can be characterised as electrical oscillations across a range of frequencies, and hence, EEG can be visualised as a spectrogram, similar to that used in sound analysis. In frequency analysis, changes in the amplitudes at various frequencies are used to make inferences about changing patterns of neural activity. However, as the experiments in this thesis rely on a traditional 'components' based analysis, in keeping with the studies which informed these experiments (particularly Corley et al., 2007; MacGregor, 2008; MacGregor et al., 2010; Collard et al., 2008; Collard, 2009), frequency analysis of ERPs is not discussed further.

may not have been detectable at the scalp. A very large timing difference may suggest a different process being engaged, and be considered to be a different component.

Differences in the topographic distributions of ERP effects are assumed to represent qualitatively different cognitive processes. This is based on the assumption that specific cognitive processes are associated with invariant underlying neuronal activity. However, this logic cannot be reversed; identical scalp distributions are not evidence of identical cognitive processes (see inverse problem, section 3.2.2, page 36).

3.4.2 Amplitude and Topographic Analyses

Most ERP studies use the Analysis of Variance (ANOVA) to test whether their effects are statistically significant. ANOVA provides a test of whether or not the means of groups are equal, and partitions variance in data into components which can be attributed to different sources of variation. ANOVA is not a perfect method for analysing ERP data, as most ERP experiments violate the assumptions of ANOVA, particularly the assumption of sphericity. This assumption states that all pairs of variables are correlated to the same degree, but in scalp recorded data, results are likely to be more strongly correlated between pairs of electrodes that are located closer together. To correct for this variation in correlation strength, and limit the chance of making Type I errors (false positives), ERP studies often use the Greenhouse-Geisser correction to adjust the degrees of freedom, making the test more conservative.

Where there appear to be amplitude differences between ERPs, it is important to assess whether the ERPs have equivalent scalp distributions. As different scalp distributions are assumed to reflect the engagement of different cognitive processes, amplitude comparisons are not valid where topographic differences are also present.

To assess differences in scalp distribution of measured electrical activity, data must be re-scaled. This procedure eliminates amplitude differences across conditions, while preserving the patterns of relative amplitude within conditions. Re-scaling is necessary because changes in the activation of neuronal generators have a multiplicative effect on ERPs, rather than an additive effect, as is assumed by the ANOVA model (McCarthy & Wood, 1985). If data are not normalised, then significant interactions between condition and location may be detected, which actually arise from a change in source strength, rather than activation of different generators, as would be assumed by an interaction with location.

One re-scaling method, commonly used, and employed in this thesis, is the Max/Min method, proposed by McCarthy and Wood, (1985). This technique takes the effect (the difference between conditions), and re-scales it at each electrode relative to all other electrodes. This is achieved by finding the the maximum and minimum value in each condition, then subtracting the minimum from each data point, and finally dividing each data point by the difference between the maximum and the minimum. Although this rescaling technique has been criticised as not robust in the case of data with non-zero baselines (Urbach & Kutas, 2002), the criticism focuses mainly on the dangers of inferring differences between scalp topographies where none exist, as opposed to using re-scaling to confirm similarity between effects, as is the case in the experiments reported in this thesis. (See also Wilding, 2006, for a response to Urbach and Kutas', 2002, criticism of the rescaling technique.)

3.5 Language and Memory Related ERP Components

As the concerns of this thesis are primarily language comprehension and recognition memory, the second half of this chapter will describe and review some ERP components relevant to to the topics and studies covered. ERPs have been extensively used to investigate online language comprehension, as they can give insight into the comprehension processes without the need for a secondary task. With regard to memory, ERPs give researchers the opportunity to better understand the processes underlying the responses participants make. Some of the effects described below are not explicitly measured or investigated in the experiments described in this thesis, but have been included with brief descriptions as they are important to the understanding of key studies which have informed the current work, and are mentioned freely in discussions throughout this thesis.

3.5.1 Auditory Sensory Processing Effects

Mismatch Negativity

The Mis-Match Negativity (MMN) is found in response to auditory stimuli that deviate acoustically from their context, even in situations where the participant is not required to attend to the stimuli. Because of this, it is thought to reflect automatic neural processes associated with the acoustic mismatch between a stimulus and the sensory memory trace resulting from previous stimuli (echoic memory) (Näätänen & Winkler, 1999; Schroger, 1997). The MMN takes the form of a negative deflection in the waveform, largest over central and fronto-central sites, which peaks 100-250ms post-stimulus (Näätänen, 2001; Näätänen, Gaillard, & Mäntysalo, 1978). For illustration purposes, a some sample waveforms demonstrating MMNs are presented in Figure 3.4.

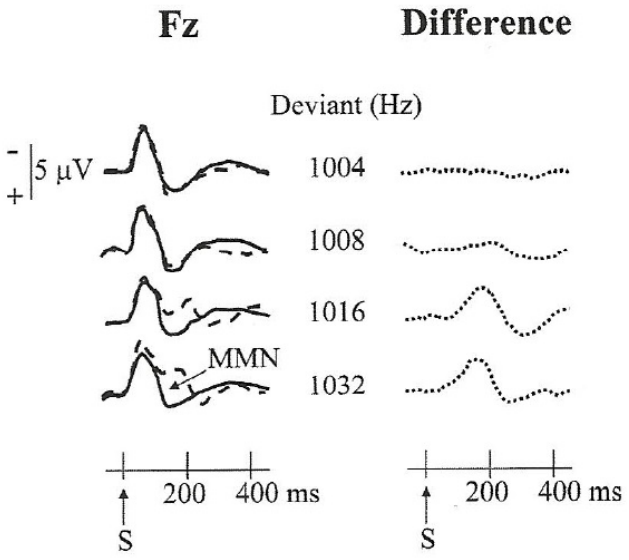

FIGURE 3.4: The waveforms in this figure demonstrate a Mis-Match Negativity, obtained in an experiment in which participants were exposed to standard tones of 1000Hz in 80% of trials, and deviant tones with frequencies with a range of higher frequencies in 20% of trials. On the left are shown ERP waveforms from the standard (solid line) and deviant (dotted line) tones. On the right is shown the difference waveform obtained by subtracting the standard waveform from the deviant waveform, demonstrating an MMN. As can be seen, the greater the deviance in stimuli, the greater the magnitude of the MMN. Diagram adapted from Näätänen and Kreegipuu, (2012); data from Sams, Paavilainen, Alho and Näätänen (2005). ERPs are shown from the Fz electrode. Negative is plotted upwards.

P300

The MMN is often followed by the P300 effect (P3). This comprises a number of distinguishable components; chief among them the frontally maximal P3a and the parietally maximal P3b (Squires, Squires, & Hillyard, 1975). The P300 family of effects are linked to the detection of deviant stimuli, for example a change in pitch or intensity of auditory stimuli. The distinction between the P3a and P3b remains somewhat unclear, but a commonly held view states that the P3a reflects the detection of deviant stimuli, and the P3b reflects orienting of attention to the same, and subsequently updating memory (Polich & Criado, 2006).

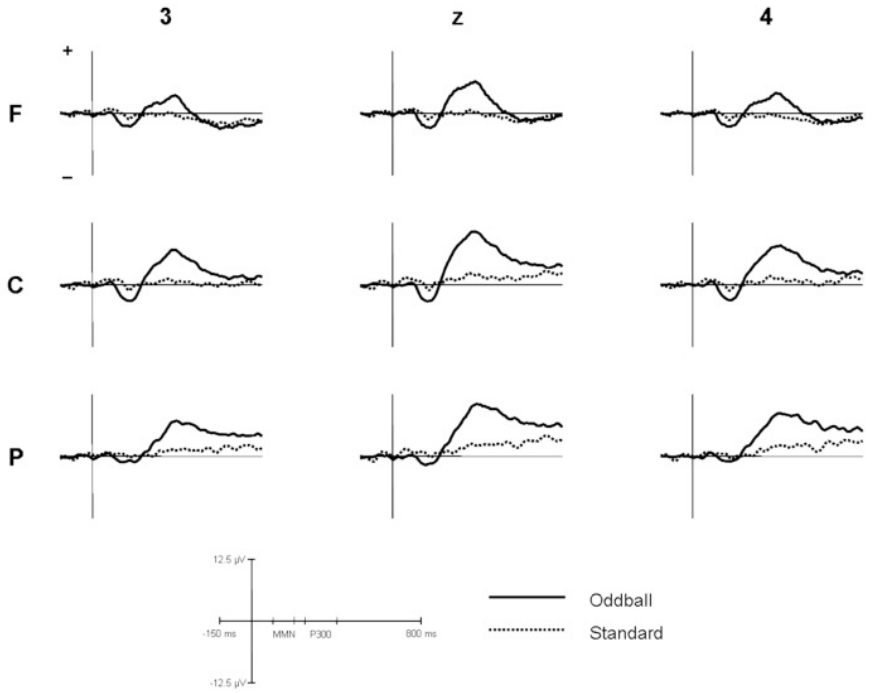

FIGURE 3.5: The waveforms in this figure demonstrate a P300 effect. Shown are the ERPs obtained as participants listen to standard target words (dotted line), and oddballs, in which the target words have been acoustically manipulated (solid line). Following the early MMN, a large difference emerges between the ERPs to standard and oddball targets. This difference, which is maximal between 250ms and 450ms, constitutes a P300 effect. Figure adapted from Collard, 2009, (p.74). Data are shown from frontal, central and parietal locations in the left and right hemispheres, and on the midline (F3, Fz, F4, C3, Cz, C4, P3, Pz, P4). Positive is plotted upwards.

3.5.2 Semantic Processing Effects

The N400

Since it was first reported by Kutas and Hillyard (1980), the N400 has been used as a dependent measure in over 1000 articles concerning a variety of topics, from language processing to face recognition and mathematical cognition. Broadly speaking, the N400 is sensitive to semantic expectation, and so any stimuli which build up a semantic context can elicit an N400 effect. Here, we are mainly concerned with the N400 and language processing.

The N400 was first discovered as a correlate of semantically incongruous (but grammatical) words presented at the ends of sentences, such as:

He took a sip from the *transmitter*. (Kutas & Hillyard, 1980)

In response to these violations of semantic expectation, Kutas and Hillyard observed a large negative going parietal wave, broadly distributed with a parietal maxima, and peaking 400ms after stimulus onset. Subsequent study has demonstrated that the size of the N400 peak can be modulated by manipulating the semantic congruity of targets, and it is primarily this modulation — the N400 effect — which is of interest to the work presented in this thesis.

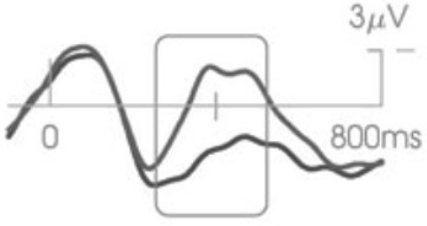

FIGURE 3.6: A sample N400 effect to visually presented stimuli at the central midline electrode, adapted from Kutas, (2009). In the 300-500ms time window (outlined), unpredictable words elicit a more negative ERP than predictable words (negative is plotted upwards).

The N400 has been demonstrated in response to semantic violations in both auditory and written stimuli. In response to written stimuli, the effect has a slight right hemisphere bias, whereas for auditory stimuli, the N400 is bilaterally distributed and can onset considerably earlier than 250ms post-stimulus. For example, Holcomb and Neville (1991) reported N400s occurring as early as 50ms post-stimulus in fluent speech. This early onset may well be due to co-articulatory effects, as listeners are able to use the changing formants of phonemes to detect the identity of upcoming phonemes in the speaker's plan, well before they are uttered. The N400 is also found in response to signed language (Kutas, Neville, & Holcomb, 1987).

With the exception of the slight timing and distributional differences for written and spoken stimuli, the N400 seems to be relatively insensitive to the physical characteristics of stimuli. The N400 has been observed for fluently produced speech (e.g. Hagoort and Brown, 2000), for speech where the target word had been artificially spliced into the utterance (e.g. McCallum, Farmer and Pocock, 1984), and for auditory stimuli whose temporal rhythm is disrupted by a silent pause separating the target from its context (Besson, Faita, Czternasty, & Kutas, 1997). For visually presented stimuli, N400 effects are observed whether stimuli are presented quickly, using the rapid serial visual presentation technique (Van Berkum, Hagoort, & Brown, 1999), or slowly (Kutas & Hillyard, 1980). The N400 is not sensitive to the type of physical property manipulation which would typically elicit an oddball effect (e.g. "She put on her high-heeled **SHOES**"), and is not sensitive to all language manipulations and violations. For example, it was not found in response to simple grammatical violations (see Kutas and Federmeier, 2009, for a review).

It is not only semantic violations which elicit an N400, such as in the example given above, but also words which are less expected given the context. For example, in a

context where "bee" was the most likely word to complete the sentence,

She was stung by a ***bee/wasp***

a larger N400 would be seen to the less expected completion, "wasp", even though it is a perfectly plausible option (Kutas & Hillyard, 1984).

The N400 is not only sensitive to sentence level semantics, but also to the semantic relationships between words presented within lists, which do not build up meaning in the way that a sentence would (e.g. Bentin, McCarthy and Wood, 1985). A reduced N400 is observed for words which have been preceded by a semantically or associatively related prime. For example, the word *nurse* typically elicits a smaller N400 in the pairing: *doctor — nurse*, than in the pairing *table — nurse*. Unsurprisingly, then, lexical repetition also leads to a reduced N400, particularly when the lag between first and second presentations of an item is relatively short (Rugg & Nagy, 1989).

The N400 is also sensitive to semantic build up at a much wider level. In addition to within-sentence comprehension, the N400 magnitude also varies in response to violations of expectation at a discourse level. For example, in the following pair of sentences, the discourse anomalous word "slow" will elicit a larger N400 than the discourse compatible word, "quick", but this effect is significantly reduced if the experimental sentence is not preceded by the discourse.

Jane was to wake her sister and brother at five o-clock in the morning. But the sister had already washed herself, and the brother had even got dressed.

Jane told her brother that he was exceptionally ***quick/slow***.

On a still wider level, the N400 has been deomonstrated to be sensitive to semantic expectation grounded in pragmatic, real world knowledge. A larger N400 is observed when comprehenders encounter words which do not fit with their real world knowledge of the topic at hand. Hagoort, Hald, Bastiaansen and Petersson (2004) exposed Dutch speakers (with a real-world knowledge of Dutch trains) to sentences such as the following:

Dutch trains are ***yellow/white/sour*** and very crowded.

Both the outright semantic violation ("sour"), and the violation of real world knowledge ("white") elicited large N400s, which did not significantly vary from one another, although there is some indication that N400 amplitude is reduced if the dis-preferred ending is of the same category as the most expected ending (Federmeier & Kutas, 1999).

In summary, the N400 can be considered to be a useful index of semantic expectation, which is modulated in response to the context of a target word; when a target word violates sematic expectation, either at the word-pair-, sentence-, context- or global-level, then a larger N400 will be produced than if the word had been a better semantic fit.

Phonological Mismatch Negativity (N200)

As previously mentioned, in auditory stimulus presentation, an early onsetting negativity is observed for targets which are semantically incongruous. Listeners begin semantic processing of words as soon as relevant information is available, and this does not require the entire phonetic form of the word to have become apparent. This was demonstrated by Van Petten, Coulson, Rubin, Plante and Parks (1999), who compared the ERPs to target words which were either the most expected completion for a sentence, shared the

onset phoneme of the most expected target, rhymed with the most expected target, or were fully incongruous. An example sentence is presented below for clarity:

> It was a pleasant surprise to find that the car repair bill was only seventeen ***dollars/dolphins/scholars/hospitals***.

Relative to the preferred ending ("dollars") the rhyming target and the fully incongruous target elicited negative ERP effects onsetting around 150ms, with a distribution similar to the N400. For the shared phoneme target, "dolphins", on the other hand, the negativity did not onset until 400ms. This represented the isolation point; the point at which the identity of the competing lexical candidates became clear. For the targets which did not share an initial phoneme with the most expected final word, the negative ERP effect onset before the isolation point; after it had become clear that the target was not the most likely candidate, but before the lexical identity of the word became clear. Although there is some debate about whether this early onsetting negativity is part of the N400 or a distinct perceptual component reflecting the fit of phonological form into a given context, it does clearly demonstrate that some semantic processing begins before the identity of the word is fully available to the listener.

Late Positive Complex (LPC)

N400 effects are sometimes followed by a positive deflection in the waveform 500-900ms post-stimulus, usually with a frontal focus and occasionally with a left hemisphere bias. This is known as a Late Positive Complex (LPC). The exact functional representation of the LPC is unclear, but it has been suggested that it indexes deliberate memory retrieval and suppression of semantic information (e.g. of the most contextually predictable word when an unpredictable word is encountered) (Federmeier, Wlotko, De Ochoa-Dewald,

& Kutas, 2007). The LPC has also been obtained in response to probe words unrelated to previously presented jokes (Coulson & Wu, 2005). The LPC is reduced for words repeated within a sentence (Van Petten, Kutas, Kluender, Mitchiner, & McIsaac, 1991). An example of an LPC effect can be seen in Figure 3.7.

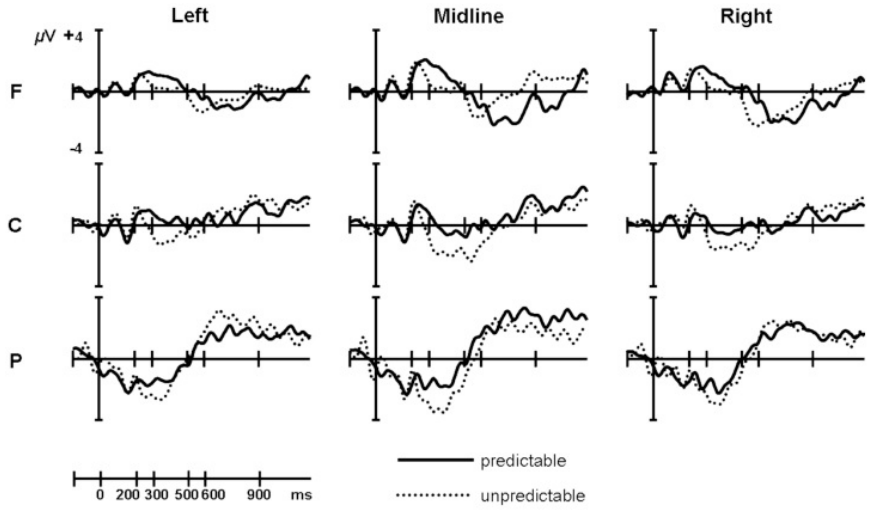

FIGURE 3.7: This figure shows LPC effects obtained as participants listened to utterances which ended with predictable and unpredictable target words. A relative positivity is observed for unpredictable compared to predictable words, between 600ms and 900ms. This positivity is maximal at frontal sites at at parietal electrodes in the left hemisphere. Diagram adapted from MacGregor, 2008, (p.56). Shown here are ERPs measured at the frontal, central and parietal electrodes in the left and right hemispheres, and at the midline.

3.5.3 Syntactic Processing Effects

P600

The P600 is known to index processes associated with grammar and structure. It is found in response to grammatical violations, as in the sentence,

Every Monday he *mow the lawn.

where the word mow elicits a P600 in relative to its grammatical counterpart, mows (Coulson, King, & Kutas, 1998). Like the N400, the P600 is elicited by violations of expectation, as well as violations of rules, and so is found for words that are unexpected, given the preferred reading of a sentence. This can be demonstrated using a garden path sentence, i.e.,

The broker persuaded *to* sell the stock was sent to jail

in which the first "to"[2] elicits a P600 relative to the word "to" in the sentence,

The broker hoped *to* sell the stock (Osterhout & Holcomb, 1992).

Because the P600 indexes grammar, rather than semantics, it can be found even in the context of otherwise meaningless sentences. The P600 takes the form of an increased positivity in the ERP waveform with a mainly posterior scalp distribution, beginning roughly 600ms post-stimulus (see Kaan, Harris, Gibson and Holcomb, 2000).

[2]For clarity, the target words "to" have been presented in italics in the examples given here. This is not usually the case when such garden path sentences are used in an experimental setting

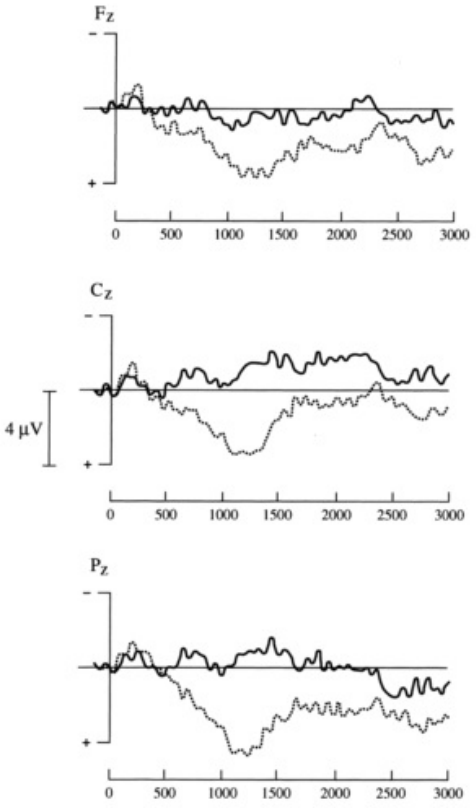

FIGURE 3.8: Example P600 effect adapted from Hagoort and Brown, (2000b). ERPs to syntactically incongruous words (dotted lines) show a sustained relative positivity compared to syntactially correct words (solid lines), which emerges around 500ms after word onset. This positivity is greatest at central and parietal electrode locations. Shown here are ERPs measured at frontal, central and parietal electrodes on the midline. Negative is plotted upwards.

3.5.4 Recognition Memory Effects

When participants in a recognition memory test are presented with items which are old (previously encountered) and new (not yet encountered), certain differences tend to emerge between the ERPs for correctly identified old and new items. The two following ERP effects both refer to these differences, and differ from repetition effects in that they take into account only data from trials where participants correctly identified old and new items.

Left Parietal Old/New Effect (LPONE)

The Left Parietal Old/New Effect (LPONE) is typically evident between 500 and 800ms after stimulus presentation, and consists of a relative positivity to correctly recognised old words, greater in the left than right hemisphere, and over parietal electrode sites.

An example of an ERP waveform and the topographic distribution of the LPONE are included in Figure 3.9. The LPONE is widely accepted as reflecting recollection — the memory for an event and its context. This is demonstrated by experiments in which participants are asked to make context judgements about previously encountered stimuli, for example their modality at presentation. These studies have demonstrated that the magnitude of the LPONE is correlated with the number of accurate context judgements (Wilding, 2000; Wilding & Rugg, 1996; Wilding, Doyle, & Rugg, 1995). Assuming that recollection is the process underlying accurate context judgements, then this evidence supports the link between the LPONE and recollection.

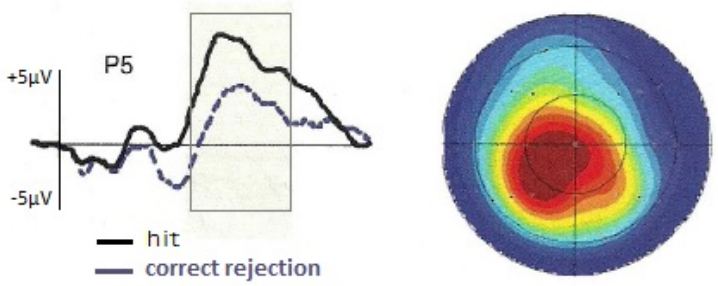

FIGURE 3.9: This figure shows waveforms demonstrating a sample Left Parietal Old/New Effect (LPONE), as measured at the inferior parietal electrode in the left hemisphere (P5), alongside the typical scalp distribution for the LPONE. Correctly remembered old words (hits) elicit a positivity over left parietal electrodes, compared to correctly identified new words (correct rejections). This difference over left parietal electrode sites is generally greatest in the 500-800ms time window (outlined). Diagram adapted from Wilding and Ranganath, 2012, (p.379). Positive is plotted upwards.

Mid-frontal Old/New Effect / FN400

The mid-frontal old/new effect is typically observed between 300 and 500 ms after the onset of presentation of the test stimulus, and takes the form of a positivity which is largest at frontal midline sites. It has been theorised that this effect reflects familiarity (Rugg et al., 1998), often described as the sensation of knowing that something has been encountered before, but without any clear memory of the event or its context. Some demonstration of this is seen in the findings of a set of studies employing two types

of new test items; genuinely new items, and plurality reversed versions[3] of previously studied items. When these were presented to participants, intermixed with previously studied old items, a mid-frontal old/new effect was observed for the plurality reversed new items (Curran, 2000). This effect was seen in the absence of a LPONE, whereas for correctly recognised old items, both mid-frontal effects and LPONE were observed. The argument runs that false alarms to similar lures should be triggered by familiarity, and so the finding of mid-frontal old/new effects, in the absence of LPONE, suggests that the mid-frontal old/new effect should be assumed to index familiarity. Similar findings have been reported in a range of other studies (Curran & Dien, 2003; Nessler, Mecklinger, & Penney, 2001).

It has also been suggested that the mid-frontal old/new effect represents conceptual priming — facilitation by virtue of semantic processing (Paller, Voss, & Boehm, 2007; Voss & Paller, 2006). Seen in this light, the mid-frontal old/new effect is very closely related to the N400, occurring in the same time window although with a somewhat more anterior distribution than a classical N400 (Kutas & Hillyard, 1984), lending it the alternative name of the FN400.

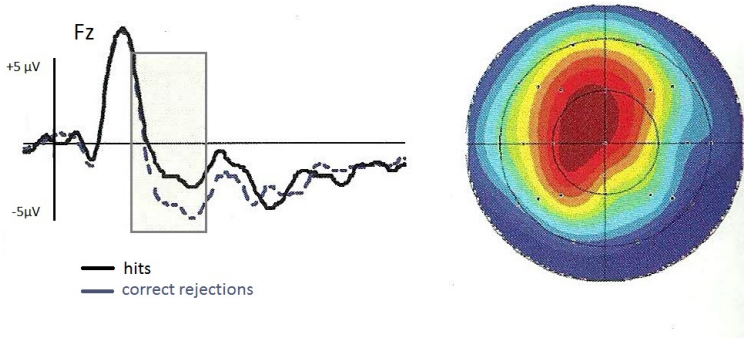

FIGURE 3.10: This figure shows waveforms demonstrating a Mid-Frontal Old/New Effect (also known as the FN400). Correctly identified old items elicit a relative positivity over frontal electrodes, compared to correctly identified new items. This relative positivity is greatest at 300-500ms (outlined). Diagram adapted from Wilding and Ranganath, 2012, (p.380). Positive is plotted upwards.

[3]Here, as in the cited paper, *plurality reversal* refers to the technique of presenting items which had been singular in a study phase as plural in a test phase, and vice versa.

3.6 Chapter Summary

ERPs allow the tracking of cognitive processes with very high temporal resolution, although the relationship between the observed pattern of voltage on the scalp and the underlying neural generators remains unclear. Carefully designed experiments which contrast experimental conditions can allow the comparison of waveforms obtained to make some inferences about the underlying neural processes engaged by those conditions. Having outlined the basics of ERP methodology, we will now move on to the specific paradigms and methodology employed in the experiments reported in this thesis.

Chapter 4

General Methods

4.1 Introduction

This chapter lays out the methodology of the experiments reported in Chapters 5 - 8 of this thesis, detailing the experimental paradigm, the development of the stimuli, criteria for participation and the experimental procedure. It also details the methods and equipment used to collect ERP data, and the analysis techniques employed to interpret that data.

4.2 Brief overview of the experimental paradigm

The aim of the experiments described in chapters 5 to 8 was to incrementally investigate whether delay might be the critical feature in disfluency. The paradigm employed in the experiments in this thesis is an extended version of the paradigm used by Corley et al. (2007), and described in more detail in MacGregor, (2008). Using this paradigm, Corley

et al. (2007) found ERP evidence of an effect of disfluency on immediate online language processing, even in the absence of a secondary task. In this thesis, this paradigm is expanded upon by adding delay-control conditions, to allow the comparison of effects elicited by disfluency and by delay only, within the same experimental setting and participant group. Immediate processing effects were assessed using ERPs, and longer-term consequences were assessed with a surprise recognition memory test.

The effect of fluency was assessed by means of a second order comparison using a manipulation of contextual word predictability. The impact of predictability was observed in fluent utterances and the resulting ERP effect compared to the predictability effects elicited in disfluent and interrupted conditions. A second order comparison was necessary to allow comparison to fluency conditions. A direct comparison of ERP responses to fluent, disfluent and interrupted sentences was not possible, as there are necessarily physical differences in stimlus sentences of these conditions, resulting in systematic differences in the baseline period (-100ms – 0ms before target word onset). Specifically, the pre-target baseline for fluent utterances was measured during the presentation of the pre-target word, whereas, for disfluent or interrupted utterances, the pre-target baseline was measured during the disfluent filler or interruption. As the stimulus input during the pre-target baseline was systematically different for the three fluency conditions, direct comparison would be invalid.

An additional benefit to the second-order comparison for fluency conditions was that it permitted a certain level of quality control. This design makes it possible to check that standard predictability effects are present for fluent utterances, particularly with regard to ERPs and recognition probability. Failure to obtain standard effects for fluent utterances would cast doubt on any interpretation of effects for disfluent or interrupted utterances.

I think a lot of people go to university because they like the student	[—/*er*/**]	**lifestyle.**
If you take your student card they'll give you a student	[—/*er*/**]	**discount.**
I think a lot of people go to university because they like the student	[—/*er*/**]	**discount.**
If you take your student card they'll give you a student	[—/*er*/**]	**lifestyle.**

TABLE 4.1: An example pair of stimulus sentences as used in the experiments described in this thesis. Each sentence can end predictably or unpredictably, and sentences are paired so that the predictable ending for one sentence constitutes a plausible but unpredictable ending for the other sentence in the pair. To create the various fluency conditions, disfluent fillers (er) or other interruptions (coughs, beeps) occur directly before the sentence final target word.

In each of the experimental chapters, stimuli consisted of pairs of sentence frames and pairs of utterance-final target words. Each target word was the most predictable ending for one sentence (mean cloze probability 0.82), and also constituted a plausible but highly unpredictable ending for the other sentence in the pair (cloze probability 0). Cloze probability was determined by means of a cloze test, described below. For an example of a pair of stimulus sentences as used in these experiments, see Table 4.1. One third of the sentences were presented fluently, one third contained a filler type disfluency (*er, erm*) before the target word, and one third contained a time matched interruption before the target word. Analysis focused on ERP and behavioural responses to the utterance final target words only.

In Experiments 1 and 2, described in Chapters 5 and 6 a 2x3 design was employed, with factors of predictability [*predictable, unpredictable*] and fluency [*fluent, disfluent, interruption*]. In Experiment 3 (Chapter 8), a dimension of context was added to the basic experimental paradigm described here. As context is pertinent only to that experiment, it is not discussed futher here, but is more fully explained in Chapter 8.

Presentation was balanced across participants so that no participant heard any sentence frame or target word more than once, and across participants, all target words and sentence frames contributed equally to the conditions obtained by crossing predictability with disfluency.

4.3 Stimuli

Generating Stimuli

Stimuli were selected from a set drawing on those used by Corley et al. (2007), and extended by the author. Stimulus sentences were created in pairs, so that the highly predictable ending for one sentence could serve as a plausible but unpredictable ending for the other. To validate the predictability of sentence endings, all prospective stimuli (including those previously used by Corley and colleagues) were submitted to cloze probability testing. In an online test, written sentence frames were presented with a blank in place of the final word, and participants were asked to give the most probable word to complete the sentence.

To prevent cross-contamination between sentences, each participant in the cloze test was exposed to a list containing only one sentence from each pair. Participants for the cloze test were recruited from this Stirling University Psychology Participant Pool, and reported no neurological or speech-language defects. All participants spoke English as their first language. Participants received course credit as compensation for participating in the cloze probability test, which took approximately 30 minutes to complete. Participants for the cloze probability test were drawn from the same population as participants were experiments 1 to 3, but individuals who had participated in the cloze probability test were excluded from participating in the subsequent experiments. Ethical approval for cloze testing was granted by the University of Stirling Psychology Ethics Committee. At least 18 participants responded per sentence. Utterances selected for inclusion in the ERP experiments had a mean high cloze probability of 0.82 (range 0.56 - 1), and a mean low cloze probability of 0.

Recording and Editing Auditory Stimuli

It is well documented that listeners are able to use phonetic cues to predict upcoming words well before their onset. In order to prevent these cues confounding listeners' predictions, all sentence frames were recorded with the same pseudo-target; *pen*. This was chosen as the unvoiced plosive phoneme at the start makes it easily identifiable on a spectrogram, and hence easy to edit accurately. For each sentence frame, two versions were recorded, one fluent, and one with an *er* before the pseudo-target, along with any vowel lengthening or changes in pace natural for the speaker. Targets were recorded in the utterance final position of a carrier sentence. Following recording, stimuli were edited using Adobe Audition CS5.5 (www.adobe.com/audition) to remove the pseudo-target, *pen*[1] from sentence frames, and to excise the carrier sentence from the recordings of target words.

For the experiment described in Chapter 5, a sine wave beep was generated (frequency 277Hz). The frequency was chosen based on the formant frequencies of the speaker's voice, so that the tone was well matched to the speech. In the interrupted condition, this was played after the fluent sentence frame, for the duration of the disfluency in the equivalent disfluent recording.

For the experiment detailed in Chapter 6, a large selection of mid-speech coughs were recorded in carrier sentences. To generate the stimuli interrupted by a cough, coughs were appended to fluent sentence frames. This meant that there were no prosodic cues (such as a change in pace or vowel lengthening) that would indicate the speaker was having difficulty. As the primary aim was to study the effect of delay introduced by disfluency, it was important to ensure that the coughs and disfluencies introduced the

[1] Audio files were cut at the onset of the pseudo-target, *pen*, as determined from the spectrogram, and so any pauses around disfluencies or prolongations naturally produced by the speaker remained intact.

same amount of delay. For each sentence frame, the duration of the *er* and surrounding pause was measured, and a cough was selected of the same duration (+/- 7ms), to be appended to the fluent recording of the same sentence. The distribution of the length of the naturally produced *ers* can be seen in Figure 4.1.

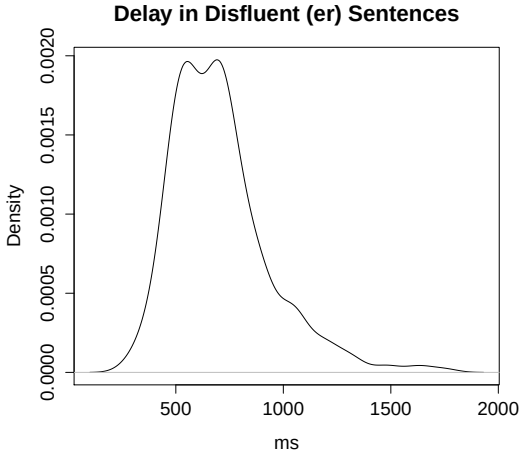

FIGURE 4.1: Density plot showing the distribution of the duration of the *ers* in the stimulus sentences selected for inclusion in the experiments reported in this thesis. Disfluent *ers* were produced within the stimulus sentences, and the speaker was allowed to produce the disfluencies in a manner that she felt to be natural for the sentence, and her speech pattern. As can be seen in the figure, the vast majority of disfluencies had a duration of betwen 400ms and 800ms.

Sentence frames, targets and beeps were not spliced together, but kept as separate audio files to allow flexibility of combinations. To ensure that there was no delay between sentence frames and targets, all audio files were pre-loaded at the start of the experiment, and the experimental script and audio speakers were subjected to extensive timing tests to ensure that any delay between the offset of the sentence and onset of the target word was no greater than 3ms.

Recording took place in the University of Edinburgh Department of Psychology recording studio with the assistance of sound technician Ziggy Campbell. Utterances were digitally recorded (16-bit, 48kHz) by a single female native English speaker. During editing, amplitudes were normalised to ensure that apparent volume approximately matched across

sentence frames and targets, and across stimuli. Finally, all stimluli were resampled at 16-bit/22050Hz and converted to *.wav* files.

Filler utterances were included in the experiment to mask potentially salient features of the experimental stimuli. These were 80 utterances of varying constraint, with similar topics and grammatical structure to the experimental stimuli. Thirty were fluent, and fifty contained some form of disfluency, in various locations. Predictability was also varied across filler utterances.

4.4 Participants

Participants were recruited using the University of Stirling Psychology Experiment Recruitment System, and word of mouth. All participants were right handed native English speakers aged 18-35, with normal or corrected to normal vision, and reported no speech/language or hearing difficulties. Participants were compensated £7.50/hour and psychology students had the option of receiving part payment in course credits. Informed consent was obtained prior to participation, and all participants were fully debriefed after participating. Ethical approval was obtained from the University of Stirling Psychology Ethics Committee prior to testing.

4.5 Software

Experiments were conducted using E-Prime 1.2[2] software to present stimuli and record responses. To ensure that E-Prime was recording the timings of stimulus presentation and responses accurately, each experiment was subjected to extensive timing testing

[2]www.pstnet.com

using Black Box ToolKit[3]. Timing testing was carried out extensively for both auditory and visual stimuli, and the E-Prime experiments were adjusted to ensure optimal accuracy. Particularly important for the auditory stimuli was to pre-load and pre-play all auditory stimuli so that they were held in the computer's cache. Failure to do this resulted in serious timing discrepancies through the course of the experiment. EEG data was recorded using Neuroscan 4.3 Aquire, Neuromedical Supplies[4].

4.6 Procedure

There were two parts to the experiments detailed in Chapters 5 and 6. The first part was designed to investigate the effect of disfluency and delay on language processing. The second part was designed to investigate the effects of delay and disfluency on longer term memory representation, and comprised a surprise recognition memory test for target items from the first part of the experiment. During the listening phase, participants were unaware that their memory would subsequently be tested.

4.6.1 Listening Task

Participants were told they would be participating in an experiment about language processing and comprehension. They were advised that they would hear a series of extracts from natural conversation, which had been re-recorded in a studio. They were told that as these were out of context, some would make more sense than others, and they should listen naturally for understanding, as in conversation. Forty of the eighty filler utterances were followed by a simple yes/no comprehension question relating to the filler utterance, and participants were told that they should answer these as quickly and

[3]www.blackboxtoolkit.com
[4]www.compumedicsneuroscan.com/

accurately as possible, although there was no time limit for their response. They were not told how many comprehension questions there would be, but advised that these would occur at random, so they should ensure that they would be able to answer a question about any sentence.

Before the experiment began, the importance of sitting still, relaxing and avoiding eye-movement was emphasised to participants. To this end, they were shown their continuous EEG, and allowed to experiment with facial and eye movements to see the effect of these on the EEG. In addition, before beginning part 1 of the experiment, participants completed a practice block, in which they heard five filler utterances, one of which was followed by a comprehension question. This allowed participants to familiarise themselves with the procedure, and ensure that the speaker volume was set to an appropriate level. The practice block was repeated, as necessary, until both the experimenter and the participants were confident that instructions had been understood and the EEG equipment was working properly.

The listening phase of the experiment was broken into four blocks of approximately 12 minutes each, interspersed with breaks of a few minutes. During the experimental blocks, the experimenter had the option of introducing extra breaks if necessary, for example if the continuous EEG indicated a problem with an electrode, or that the participant appeared to be getting very tired. At the beginning of each block, there was a visual reminder encouraging participants to relax and to keep their eyes fixed in the middle of the computer screen. The background of the computer screen was dark blue. Choosing a dark background avoided causing participants eye-strain or encouraging them to squint in the darkened room. The onset of each utterance was marked by a pink fixation cross, which appeared for 1000ms, which changed to green as the utterance began. The cross

remained on the screen for the duration of the utterance, after which the screen was blanked for 1500ms.

4.6.2 Recognition Memory

After completing all four listening blocks, participants were told that the second part of the experiment would not be more listening, but a memory test, investigating how well they had remembered individual words from the sentences they had just heard.

The stimuli for the memory test consisted of the 324 single words which had featured as targets in the previous listening task, interspersed with 324 new words, which had not occurred anywhere in the listening block; neither as targets, nor as parts of sentences or fillers. The participants' task was to discriminate between old and new words, using a response box. The response box had five buttons, and participants were instructed to use the two outermost buttons, one to indicate 'old' and one to indicate 'new'. Which button was used for 'old' and 'new' was counterbalanced across participants. Participants responded using their index fingers.

Stimuli were presented visually, in white text against a dark blue background. At the beginning of each trial, the screen was blanked for 250ms, followed by a green fixation cross for 400ms. The stimulus was then presented for 1500ms, followed by a blank screen for 500ms. Participants were free to respond at any time during the stimulus display or during the blank screen. If participants responded, this was followed by a certainty judgement in which participants were given up to 5000ms to indicate their certainty about the response they had just made using the buttons 1-5 on the response box. If no initial old/new judgement was made, the certainty screen was skipped. The trial ended with a screen inviting participants to press any button to begin the next trial.

4.7 EEG Collection

EEG data were collected from 62 Ag/AgCl electrodes embedded in an elasticated cap (Quikcap system[5]).The arrangement of the electrodes in the cap was based on an extended version of the international 10-20 system (Jasper & Carmichael, 1958). Data were recorded referenced to a ground electrode located between Cz and CPz. EOGs were recorded from electrodes located on the outer canthi of both eyes to monitor lateral eye movements, and above and below the left eye to monitor eye blinks. Electrode impedances were kept below 5kΩ. The EEG recordings were amplified (band pass filter 0.01—40Hz) and continuously digitised (16 bit) at a sampling frequency of 250Hz.

4.8 ERP Processing

To minimise the effect of eye-blink artifacts on the data, their contribution to the ERP waveforms was estimated and corrected using a regression procedure (Neuroscan Ocular Artifact Reduction). An average blink for each subject was calculated from a minimum of 32 blinks, before the contribution of the blink was removed from all other channels on a point-by-point basis. The data were then divided into epochs timelocked to the onset of target words in the stimulus utterances. These epochs began 100ms before the onset of the target, and continued for 2000ms after the onset (total 2100ms). Epochs were baseline corrected using the 100ms pre-stimulus baseline period and re-referenced to the average of the left and right mastoid electrodes. Any epochs with a baseline drift greater than $75\mu V$ were rejected, along with epochs where the amplitude on any channel (excluding VEOG) exceeded $75\mu V$. For the experiment reported in chapter 6.4.4 and for EEG collected during the surprise memory task (sections 8.3.2 and 8.4.1),

[5]www.neuroscan.com/quick_caps.cfm

this artifact rejection procedure used a parameter of 3 standard deviations from the mean amplitude, rather than $75\mu V$ absolute amplitude. This parameter was selected for pragmatic reasons — these experiments suffered from low trial numbers, and so more lenient filters were required in order to allow analysis of the data. Following artifact rejection, data were smoothed over 5 points, so that each point in the resultant ERP represented the mean of the two previous and two subsequent points. Finally, ERPs were averaged over each presentation condition, and then over multiple participants, to form grand average ERPs.

4.9 Analyses

This section describes the standard ERP and behavioural analyses employed in Chapters 5-8. Mixed effects modelling is introduced in Chapter 9, and the methods used for this will be described at that point. All analyses were carried out using the R statistical computing software environment[6] (R Core Team, 2013). Within R, ANOVA analyses were performed using the "ez" package (Lawrence, 2013), to allow the integration of Greehouse-Geisser corrections for non-sphericity.

4.9.1 Listening Task — ERP Analyses

ERPs were quantified by averaging their amplitude (relative to the pre-stimulus baseline) over time-windows of interest. Data were initially analysed using Analysis of Variance (ANOVA). As the ANOVA model assumes sphericity (homogeneity of variance among levels of each factor), degrees of freedom are adjusted using the Greenhouse-Geisser

[6] www.r-project.org

correction, and corrected F-ratios and p values are reported where appropriate. Only significant outcomes relevant to the experimental manipulation are reported.

Analyses primarily focussed on the N400 effect. However, where there was evidence for predictability effects early in the epoch, or in the later 600-900ms time window, these were also submitted to analysis.

Amplitude Analyses

As noted previously, utterances in the various fluency conditions *(fluent, disfluent, interrupted)* could not be directly compared, as the physical differences in the stimuli in the pre-stimulus baseline period would necessarily lead to systematic differences. Thus comparisons between fluency conditions are achieved by comparing the predictability effect in each fluency condition.

To establish whether significant predictability effects were produced in each fluency condition, ERPs produced in predictable and unpredictable conditions were compared using two ANOVAs. First, a global analysis was used, incorporating factors of predictability *(predictable, unpredictable)*, location *(F, FC, C, CP, P)*, hemisphere *(left, right)* and site *(superior: electrodes 1 & 2, medial: electrodes 3 & 4, inferior: electrodes 5 & 6)*. This ANOVA assessed for the presence of significant predictability effects within each fluency condition, as well as the presence of any hemispheric or site differences between predictability conditions. The electrodes incorporated in this analysis can be seen in Figure 4.2 (left). It was not possible to conduct an ANOVA with data from all 64 electrode sites and incorporating a factor of electrode. This failed on the grounds that the electrode factor had too many degrees of freedom for the number of subjects submitted for analysis. Conducting analysis with data from all 64 electrodes but without a factor

of electrode would have violated the ANOVA structure, as not all levels of all factors would have been balanced across levels of other factors[7]. For this reason, the subset of electrodes shown in Figure 4.2 (left) were selected for the global analyses.

Where no hemispheric differences between predictability conditions were revealed, or where the global analysis suggested a larger effect towards midline sites, a midline analysis was performed on data from midline sites only. This ANOVA incorporated factors of predictability *(predictable, unpredictable)* and location *(F, FC, C, CP, P, PO)*.

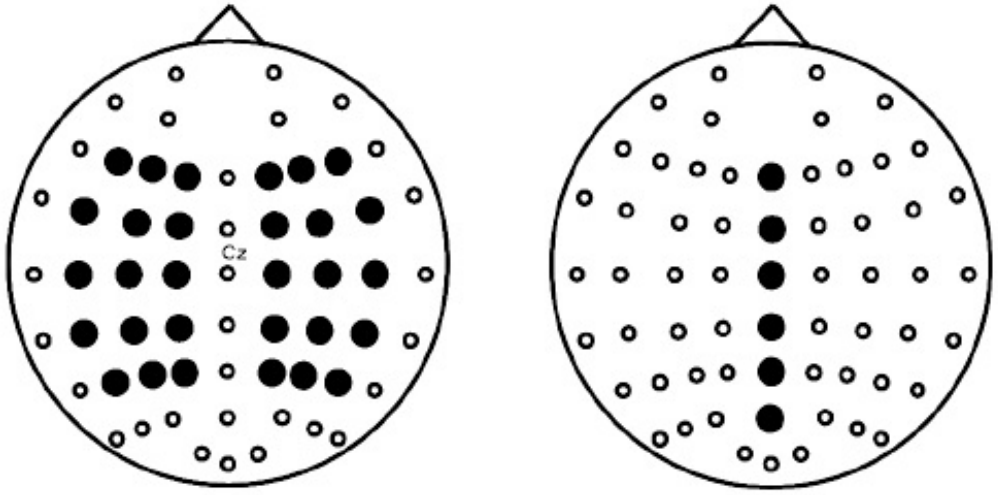

FIGURE 4.2: Map of electrodes reported in analyses. To the left, electrodes incorporated in global hemisphere analysis. Where an effect was found which was not hemisphere biased and strongest towards the midline, this hemisphere analysis was followed up with an analysis based on six midline sites (right).

Topographic Analyses

To make comparisons between fluency conditions, it was necessary to assess for interactions between predictability and fluency across conditions. For this comparison to be meaningful, it first had to be established that there were no distributional differences

[7]The failure of global ERPs on account of having too few subjects for the degrees of freedom being tested is specific to the R environment. Using SPSS (as has been used for the majority of the work which precedes this thesis, this ANOVA would have been permitted.

between the predictability effects obtained in the various fluency conditions. This was assessed by means of a mulitilevel ANOVA, using data from the subtraction waveforms representing the difference between predictable and unpredictable targets. This ANOVA employed factors of fluency *(fluent, disfluent, interruption)*, location *(F, FC, C, CP, P)*, hemisphere *(left, right)* and site *(superior, medial, inferior)* and incorporated data from the electrodes used in the global analysis (see Figure 4.2, left). Before topographic analysis was carried out, data were rescaled using the McCarthy and Wood max-min method (McCarthy & Wood, 1985), in order to remove magnitude differences between effects.

Quantitative Comparison

Where significant predictability effects emerged for multiple fluency conditions, and there was no evidence for distributional differences of these effects, further analyses were performed to assess for quantitative differences in the magnitude of the predictability effects between fluency conditions. To this end, a factor of fluency was incorporated into the hemispheric and midline analyses. To avoid amplifying noise in the data, this analysis was not performed on a subtraction waveform (unpredictable - predictable), but on the ERPs from all conditions, and assessed for interactions between predictability and other factors.

4.9.2 Recognition Memory — Behavioural Analyses

Memory performance in each condition was quantified as the probability of previously heard words being later recognised. During the memory test, new words which had not appeared anywhere in any of the stimulus utterances were also presented, intermixed with the 'old' words. This allowed the calculation of a false alarm rate per subject, which

made it possible to determine to what extent each participant was relying on accurate memory, rather than guessing, but it is not possible to calculate a false alarm rate for each conditionas the new words had not been previously presented, and so could not be assigned to any presentation condition.

An ANOVA with factors of predictability *(predictable, unpredictable)* and fluency *(fluent, disfluent, interruption)*, and using stimulus as a random factor, was used to assess differences in the memorability of predictable and unpredictable words across fluency conditions.

4.9.3 Recognition Memory — ERP Analyses

Analysis of the ERPS for the recognition memory task focussed on two time windows; 300-500ms and 500-800ms. These were selected a-priori based on the recognition memory ERP literature. Analysis followed a similar strategy to that described for the listening task. Only data from correctly identified targets was taken into account.

Amplitude Analyses

The waveforms for correctly recognised old targets in each fluency and predictability condition were compared to the waveform for correctly identified new words using a global ANOVA with factors of condition *(old, new)*, location *(F, FC, C, CP, P)*, hemisphere *(left, right)*, and site *(superior: electrodes 1 & 2, medial: electrodes 3 & 4, inferior: electrodes 5 & 6)*. The ANOVA drew on data from the electrodes indicated in Figure 4.2 (left). As was the case for the listening task ERPs, this ANOVA assessed for the presence of any effects of condition, as well as characterising the distribution of these differences across hemispheres, locations and sites. Only main effects of condition, or

interactions with the effect of condition are of interest to the experimental manipulation, and so only these are reported.

Topographic Analyses

Where recognition memory effects emerged in two or more predictability and fluency conditions, their topographic distributions were compared, using the strategy described in Section 4.9.1.

Quantitative Comparison

In the event of significant recognition memory effects emerging for multiple fluency and predictability conditions, and in the absence of differences in topographic distribution, effects were quantitatively compared. This was carried out using difference waveforms. As there were no new items in each fluency and predictability condition, ANOVA structure would have been violated had quantitative comparison been attempted using all waveforms and incorporating a factor of condition. Thus, the comparison employed an ANOVA with factors of fluency *(fluent, er, interruption)*, location *(F, FC, C, CP, P)*, hemisphere *(left, right)*, and site *(superior: electrodes 1 & 2, medial: electrodes 3 & 4, inferior: electrodes 5 & 6)*.

Chapter 5

Disfluency as Delay: Comparing the Immediate Effects of Fillers and Beeps

5.1 Introduction

Filled and silent pauses are widely distributed throughout spontaneous speech, affecting approximately six percent of words (Fox Tree, 1995). These disfluencies have been demonstrated to affect both online language processing, changing listeners' expectations for upcoming speech (Arnold et al., 2004) and changing the listeners' attention to the speech stream (Collard et al., 2008). To what extent these effects are driven by the form of the filler, and to what extent they depend on the delay the filler introduces into the utterance remains unclear. In this chapter I describe an ERP experiment in which fillers are compared with beeps, in order to establish whether the same effects can be elicited by words preceded by non-lexical beeps as by disfluent fillers.

Fillers have been demonstrated to affect listeners' processing of speech, and subsequent memory for speech. On hearing a filler, listeners are likely to predict that the following item will be one which is new to the discourse, as evidenced in mouse tracking (Barr, 2001) and eye-tracking paradigms (Arnold et al., 2004). Eye-tracking also suggests that disfluency leads listeners to expect referents which are difficult to name as a function of their nature, or which they expect to be problematic for the speaker (Arnold et al., 2007). Taken together these studies suggest that following a disfluency, speakers turn their expectation towards referents which are most problematic for the speaker. ERP evidence suggests that this shift in expectation can also occur even when listeners are not presented with a limited set of possible referents, as in mouse- and eye-tracking studies (Corley et al., 2007). Focusing on the well documented N400 component which is larger for unpredictable than predictable words, Corley et al. (2007) found a reduction in the N400 effect when auditory target words were preceded by a filler. This reduction demonstrates that predictable and unpredictable words were processed more similarly when they were preceded by a filler than when they were presented in fluent sentences.

In addition to immediate language processing, disfluency also impacts subsequent memory. A surprise memory test administered to participants at the end of Corley et al.,'s (2007) experiment found that participants were more successful at recognising items which had been disfluent at presentation compared to those which had been fluent. Some explanation of the mechanism of this phenomenon may be found in the work of Collard et al. (2008). This study found that the P300 — an ERP component thought to index attention — was not present for acoustically deviant target words which had been preceded by a disfluent filler, although these items still elicited a Mis-Match Negativity (MMN), indicating that participants detected the change in acoustic quality. Collard et al. (2008) interpreted this lack of P300 as indicating that participants attention to the

speech stream had already been heightened by the disfluent filler, and so although participants were aware of the acoustic change when they heard the target word, attention could not be raised any further, explaining the absence of the P300. This heightened attention to material following disfluency may explain the improved recognition memory demonstrated by Corley et al. (2007), and replicated by Collard et al. (2008).

However, it remains unclear whether these predictive and attentional effects are driven by the form of the disfluency, or by the delay that disfluency introduces to the utterance. A number of studies have addressed this question, using a variety of methods. Brennan and Schober (2001) compared the effects of fillers with time matched silent pauses in a task requiring participants to identify a single target among a set of simple geometrical shapes. Both speed and accuracy were higher in response to utterances containing fillers or silent pauses, and no difference in effects was found between the disfluent conditions. Similarly, in a task requiring participants to press a button identifying which of a pair of visually presented shapes was being described, Watanabe, Hirose, Den and Minematsu (2008) reported no difference between the effects of silent pauses and fillers in biasing listener's expectations towards complex referents. Expectancy changes associated with silent pauses were also reported by MacGregor et al. (2010). MacGregor reported an attenuated N400 in response to unpredictable target words completing sentences, when these targets had been preceded by a pause, reflecting the outcome of Corley et al. (2007) who reported a similar effect using fillers.

In an experiment comparing the effects of fillers with silent pauses and artificial beeps on participants reaction times, Corley and Hartsuiker (2011) found no difference between these three types of delay. Participants followed instructions to identify targets from pairs of images displayed on a screen, and responded faster when the target word in the

instructions was preceded by a delay than when the instruction was fluent. This effect was found for both easy and difficult to name targets.

Noisy external interruptions have also been contrasted with disfluent fillers in a task requiring participants to disambiguate sentences containing a temporary syntactic ambiguity. Participants performed more accurately when there was a delay before the head noun of the ambiguous phrase, whether the delay constituted a filler *(uh)*, or a noisy interruption, such as doorbells ringing or dogs barking (Bailey & Ferreira, 2003).

There is also some evidence suggesting that the surface form of a disfluency influences listeners' interpretations. In particular, Fox Tree (2001) found participants to be faster at identifying target words when they were preceded by filler than when they occurred in fluent utterances. Interestingly, this benefit was only found for the filler *uh* and not the filler *um*. In this experiment, however, not only the phonetic form of the filler changed, but the length of the attendant pauses varied too. The *ums* heralded longer pauses than did the *uhs*. Both the mean lengths of the fillers themselves (327ms for *uh*; 384ms for *um*), and the periods of silence around the fillers varied systematically, meaning that the hesitations marked by *uhm* (1080ms) were longer than those marked by *uh* (1031ms). The author concluded that participants were affected by the phonetic forms of fillers, and that the brief delay signalled by *uh* heightened listeners attention, whereas listeners may not attempt to maintain the same heightened state of attention for the duration of a longer pause, as signalled by *uhm*. Although this study provides an interesting indication that the phonetic form of a filler may influence how it is interpreted by the listener, the combination of delay and filler does not allow an orthogonal contrast of delay with phonetic form.

Further evidence for the role of form in driving listeners' responses to disfluency is

provided by the findings of MacGregor, (2008). MacGregor repeated the experiment reported in Corley et al. (2007), in which an attenuated N400 for unpredictable words was observed following a filler. Using the same stimuli, but with the fillers replaced with silent pauses, no such attenuation was found.

Fraundorf and Watson (2011) also reported a specific effect of disfluent fillers, which did not generalise to delay, in a storytelling task. Participants were asked to retell previously heard passages from the novel, Alice in Wonderland. The authors recorded the probability of each plot point being recalled, according to whether the point have been fluent, interruped by a filler, or interrupted by a cough in its original presentation. They reported a memory benefit conferred by fillers, but not by coughs.

Finding a Control Condition for Disfluent Fillers

When setting out to investigate the role of delay in driving disfluency effects, the form of the delay deserves some careful thought. Some studies, including those described above (Watanabe, Hirose, Den, & Minematsu, 2008; Brennan & Schober, 2001; MacGregor et al., 2010; MacGregor, 2008), have used a silent pause as a comparison condition. In these cases, silence is assumed to add delay to an utterance without contributing any supplementary information about the speaker's state. However, silent pauses may also be interpreted as disfluent (Maclay & Osgood, 1959), and so cannot safely be assumed to be a neutral delay, free of connotation. In an attempt to address this problem, Bailey and Ferreira (2003) used environmental noises, such as animal calls and doorbells to add delay to utterances without adding propositional content. However, for these interruptions to be naturalistic and, most importantly, believable to the listener, it must be plausible that the speaker would stop speaking because of the interruption. It is not clear that this is necessarily always the case — anecdotal evidence suggests that speakers

may continue to talk over such an interruption unless they were unable to continue for some other reason, and so the cessation of speech may be interpreted by the listener as a disfluency.

One approach is to add delay to speech without suggesting that the original speech was broken, by using an interruption clearly edited into the stimulus post-recording, such as the beep used in Corley and Hartsuiker (2011). A beep which has been retrospectively edited into the recording is unlikely to be interpreted as a natural disfluency, and as such may be considered the purest form of non-meaningful delay it is possible to use in auditory sentences. For this to be effective, participants must be explicity informed that the beep has been edited in during processing, and that it is not obscuring any words. Evidence to suggest that this type of high level knowledge affects the way listeners respond to interruptions in speech can be found in Arnold et al. (2007), who found that listeners' bias to expect unfamiliar items following disfluency disappeared when participants were told that the speaker had object agnosia (difficulty describing familiar items), demonstrating that expectations and responses to disrupted speech can be modulated by participant instruction.

As an artificial tone is not speaker-generated or controlled, it is unlikely to be interpreted as indicating speaker difficulty, it presents a more easily controlled, although less ecologically valid alternative to silent pauses for introducing delay to an utterance, where connotations of speaker difficulty are not desirable. This is the approach adopted in this experiment.

Summary

This experiment is designed to investigate the role of delay in driving the disfluency effects previously described. The experiment will directly contrast non-linguistic delay, filled with a beep, against disfluent fillers and fluent control sentences, and take as a dependent variable the size of the N400 effect between predictable and unpredictable targets. Results of previous studies suggest that targets preceded by a disfluency will exhibit a reduced N400 effect compared to targets in fluent sentences. If this reduction is a product simply of the delay introduced by the filler, then the same reduction should be elicited by the beep condition. If, however, the form of the filler is critical for reducing the size of the N400 effect, then the effect seen following a beep should instead pattern with the effect for targets in fluent conditions, and an attenuated N400 should be seen only where a disfluent filler interrupts the utterance.

5.2 Methods

Stimuli

Stimuli consisted of 324 highly constrained utterances ending in predictable or unpredictable target words. One third of the utterances were fluent, one third contained a filler (*er, um*) before the target word, and one third contained a beep before the target word. This beep was carefully time matched to the disfluency in the equivalent disfluent sentence (tolerance +/-7ms). Importantly, the beep was spliced into the fluent recording of the sentence. This was to ensure there were no prosodic cues that would cause listeners to expect disfluency. For a full description of stimuli, see Section 4.3. An example

stimulus set can be seen in table 4.1. Listeners also heard 80 non-experimental filler utterances, of similar length and concerning similar topics to the experimental utterances, but containing a slightly wider range of disfluencies, distributed throughout the sentence in order to mask some of the salient features of the experimental utterances.

Participants

Twenty five[1] right-handed native English speakers (5 male, mean age 22.0 years; range 18-35 years) took part in the experiment. For full details of recruitment, see section 4.4.

Procedure

The experiment consisted of two halves; the first half focused on online processing, while the second was designed to assess the longer term effect of delay and disfluency on memory. During the first block, participants heard the stimulus sentences, and were instructed to listen naturally for understanding. Forty non-experimental filler utterances were followed by a simple on-screen *yes/no* comprehension question which participants were required to answer with a button press. Stimuli were randomised and presented in four blocks lasting approximately 12 minutes each, seperated by a short break.

Following the listening section of the experiment, participants completed a surprise recognition memory test, featuring the utterance-final target words. These were interspersed with frequency matched 'new' words, which had not been encountered anywhere in the previous block. Words were visually presented, and participants were asked to

[1]The original intention was to collect data from 24 subjects. Data from one extra subject were collected as a replacement for a dataset which on visual inspection at time of collection appeared to be noisy. In processing, however, this dataset provided enough trials for inclusion. Visual inspection of the grand average ERPs reveals them to be less noisy when all 25 subjects are included. As statistical analysis reveals no meaningful difference in the pattern of results whether 24 or 25 subjects are included, the data reported here make use of all 25 datasets.

	fluent		er		beep	
	predictable	unpredictable	predictable	unpredictable	predictable	unpredictable
minimum	25	28	25	25	26	26
maximum	53	54	54	54	54	54
mode	50	51	39	42	50	50
mean	41.84	42.68	42.24	42.84	41.92	41.88

TABLE 5.1: Numbers of trials included in ERP analysis for each condition (n=25).

discriminate as quickly and accurately as possible between old and new words using their index fingers to press two response keys on a button box. Response keys were counterbalanced across participants. As this chapter is focussed primarily on the immediate effects of disfluency, results of the memory test are discussed in section 8.3.1.

Throughout the experiment, EEG was recorded from the scalp using the Neuroscan Quickcap system (see Section 4.7).

Data analyses

ERPs for each condition were formed by averaging 2000ms epochs time locked to the onset of the target word, with a 100ms pre-stimulus baseline. For a participant's data to be incorporated into the analysis, a minimum of sixteen useable trials per condition was required. See Table 5.1 for details of the number of trials incorporated into the analysis for each condition. ERPs were quantified by measuring the mean amplitude over two time windows of interest, 300-500ms and 600-900ms, based on previous research on the effect of disfluency on language processing (MacGregor et al., 2009; MacGregor, 2008; Corley et al., 2007). There was no behavioural measure in this part of the experiment, although there was a filler task; answering yes/no comprehension questions about filler utterances. All participants scored at least 97% on these questions, and no further analysis of this filler task will be reported. A fuller explanation of the processing of the EEG and analysis of ERPs can be found in Section 4.8.

5.3 ERP results

For fluent utterances, ERPs to unpredictable target words show a negativity compared to predictable target words (see Figure 5.1). This negativity onsets relatively early, around 40ms after stimulus-onset, and is long lasting, continuing throughout the epoch. The difference is maximal around 350ms after stimulus onset and at parietal midline sites. This relative negativity elicited by unpredictable words lasts until 800ms at parietal sites, and throughout the epoch at frontal and fronto-central sites.

ERPs to unpredictable targets in disfluent utterances (see Figure 5.2) were more negative than to predictable targets, with the negativity onsetting around 180ms post-stimulus-onset and maximal at 400ms, and greatest at centro-parietal midline sites. The relative negativity for unpredictable words lasts until 500ms at frontal sites, and 700ms at parietal sites. Following the negativity, unpredictable words show a relative positivity compared to predictable words over frontal midline sites and over left parietal sites (see Figure 5.2).

Following an artificially inserted beep (see Figure 5.3), both predictable and unpredictable targets elicited an early positivity at 60ms, followed by a negative peak at 100ms and a second positive peak at 200ms. This may be interpreted as an N1 – P2 complex elicited by the change of acoustic quality from an artificial sine tone to natural speech. ERPs to unpredictable words are more negative than ERPs to predictable words. This negativity onsets at around 135ms, and continues until around 700ms.

To assess the significance of the effects observed in each time window of interest, each fluency condition was analysed separately in ANOVAs incorporating factors of predictability *(predictable, unpredictable)*, location *(F, FC, C, CP, P)*, hemisphere *(left, right)* and site *(mid, superior, medial, inferior)*. Where this broad ANOVA revealed significant

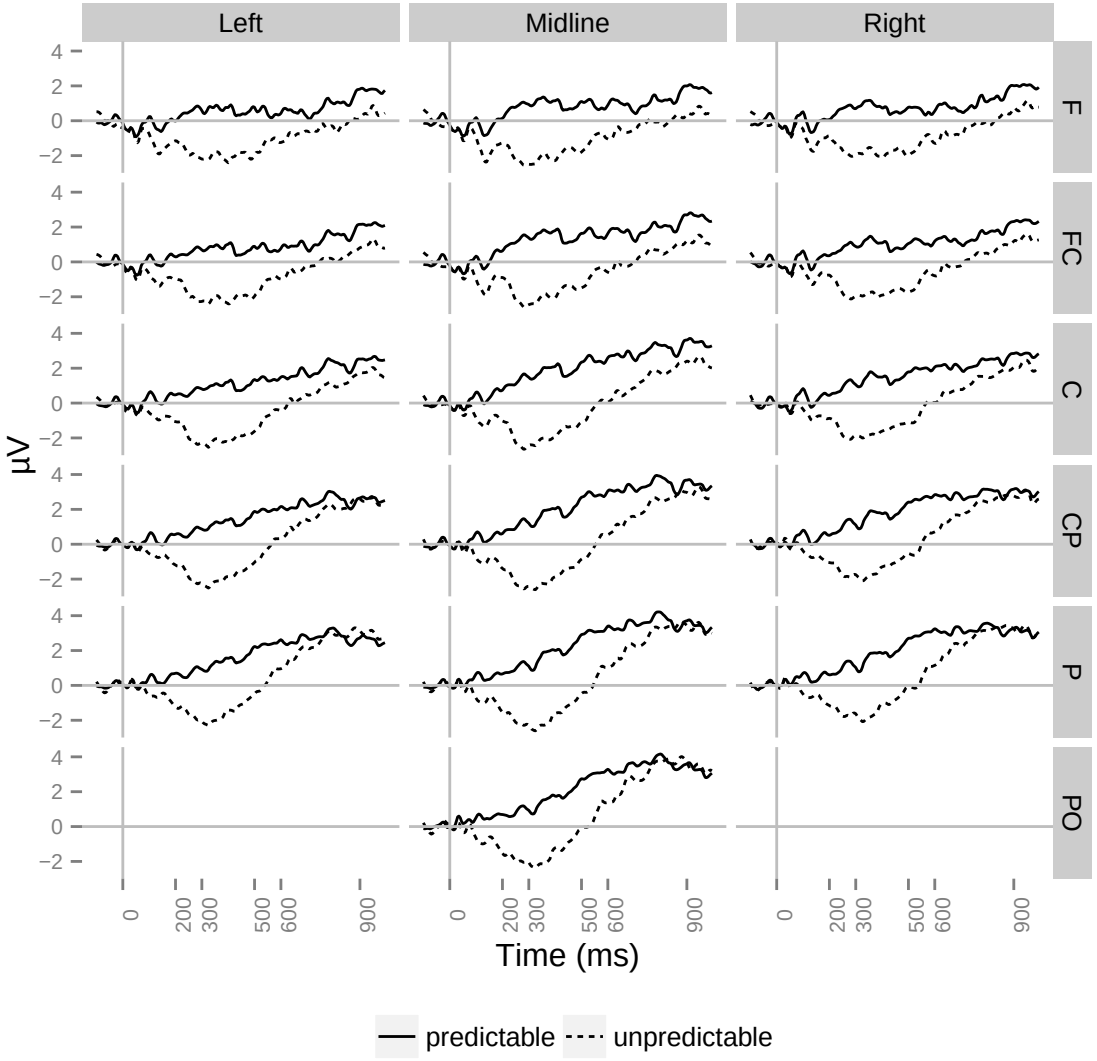

FIGURE 5.1: Grand average ERPs (n=25) for final words in *fluent* utterances. Shown here are ERPs as measured at frontal (F), fronto-central (FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline. Unpredictable words elicit a broadly distributed negativity relative to predictable words. This negativity onsets around 40ms after stimulus-onset, and is long-lasting, continuing throughout the epoch at frontal sites, but fading at parietal sites around 800ms. The relative negativity for unpredictable words is largest in the standard N400 time window.

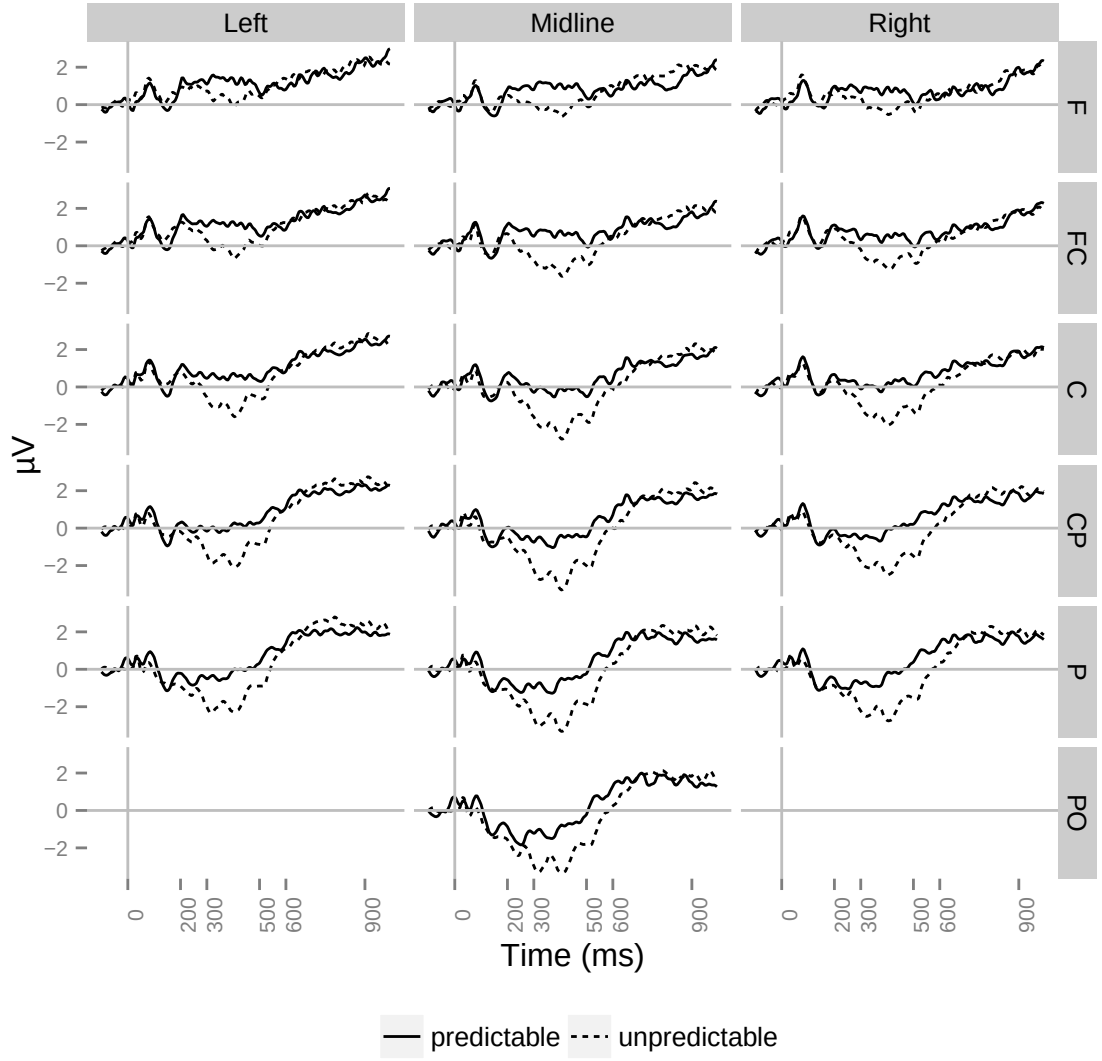

FIGURE 5.2: Grand average ERPs (n=25) for final words in *disfluent (er)* utterances. Shown here are ERPs as measured at frontal (F), fronto-central (FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline. Unpredictable words show a relative negativity compared to predictable words, which onsets around 180ms, and continues until 500ms at frontal sites, and 700ms at parietal sites, before giving way to a relative positvity at frontal and left parietal electrode sites.

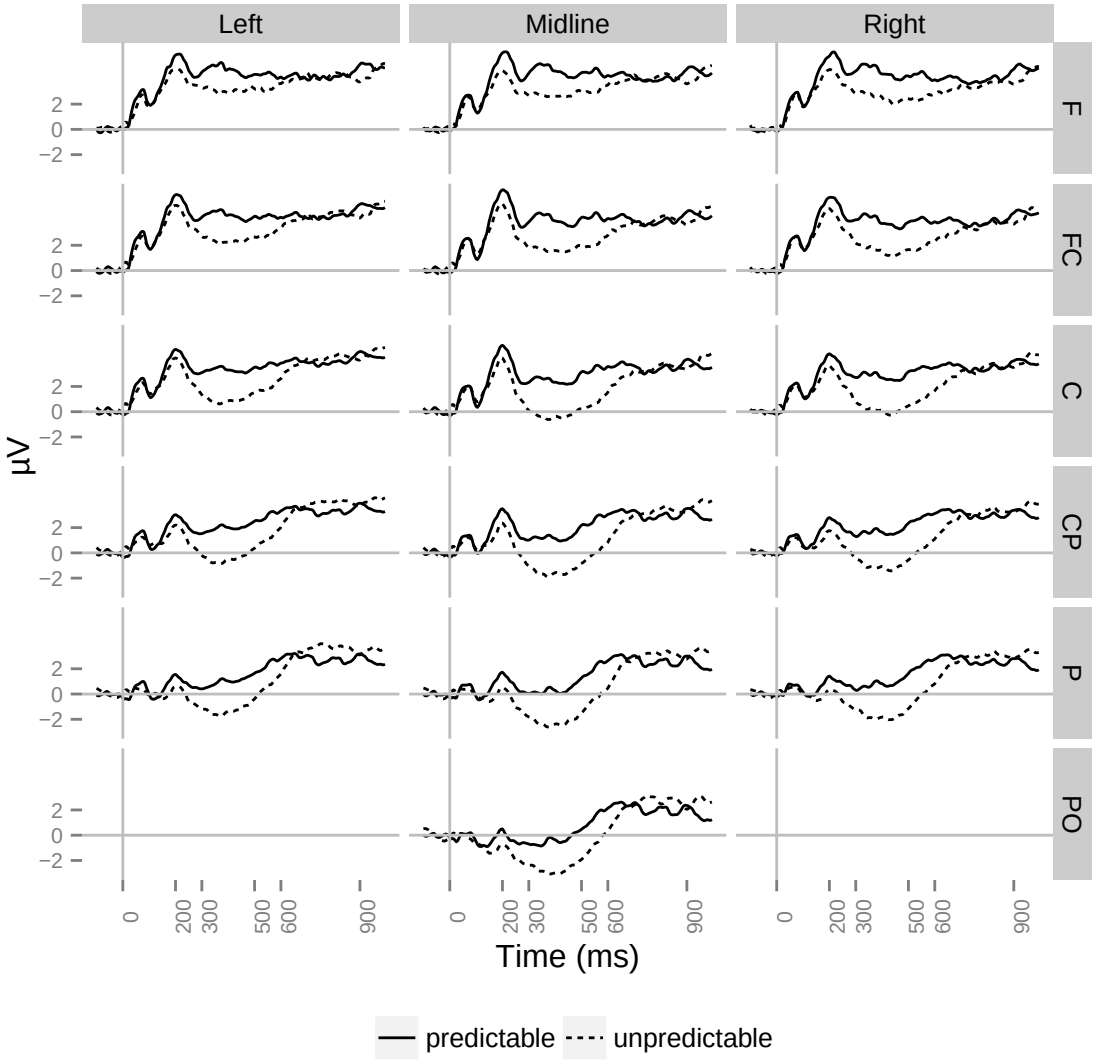

FIGURE 5.3: Grand average ERPs (n=25) for final words in *interrupted (beep)* utterances. Shown here are ERPs as measured at frontal (F), fronto-central (FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline. Unlike in the ERPs elicited by fluent and disfluent utterances (see Figs. 5.1 and 5.2), the beginning of the epoch is characterised by a positive-negative-positive complex for both unpredictable and predictable words. The topography and the timing of this activity are consistent with an N1-P2 complex, elicited by the change in acoustic quality from an artificial beep to natural speech. Following this early activity, unpredictable words elicit a more negative ERP than predictable words, with the relative negativitiy onsetting around 135ms, and continuing until 700ms. This negativity for unpredictable words is broadly distributed across the scalp, but maximal at centro-parietal midline electrode sites.

effects of interest, this was followed up with a second ANOVA, incorporating data from the midline, and factors of predictability *(predictable, unpredictable)* and location *(F, FC, C, CP, P)*.

5.3.1 0-200ms

As the relative negativity elicited by unpredictable words onset early in all three fluency conditions (40ms, 180ms, 135ms respectively), the early (0-200ms) time window is analysed in order to establish whether these effects are significant, and to determine whether these should be considered distinct from any subsequent N400 effect, or an early onset of the same. Although there is some variability in the onset of the early negativity across fluency conditions, the 0-200ms time window was selected in line with the literature (MacGregor, 2008).

Amplitude analysis — 0-200ms

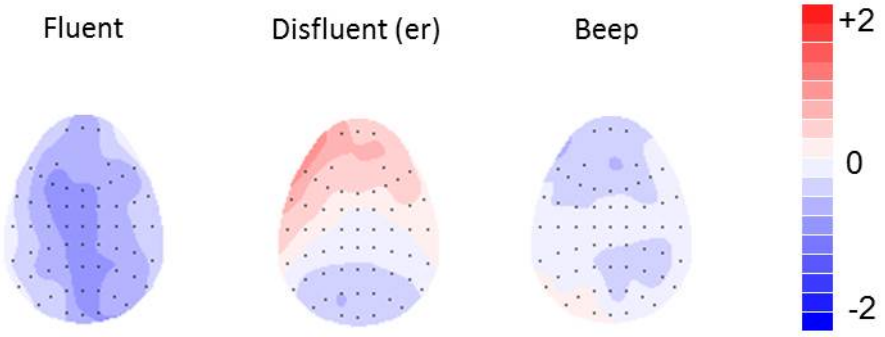

FIGURE 5.4: Scalp topographies (n=25) showing predictability effects in the 0-200ms time window for targets in utterances which were fluent (left), disfluent (middle) and interrupted by a beep (right).

A multilevel global ANOVA revealed a main effect of predictability in the 0-200ms time window for targets in fluent utterances [$F(1,24) = 4.36$, $\eta^2_G = 0.052$, $p < 0.05$]. No interactions involving predictability reached significance. A midline ANOVA also revealed a

main effect of predictability, but no interaction with location. This reflects a negativity for unpredictable words which is evenly spread over the scalp. For disfluent and interrupted (beep) utterances, neither the global nor the midline ANOVAs revealed any main effect of predictability, or significant interactions involving predictability, confirming the impression given by Figure 5.4. As only fluent utterances elicited a significant effect in the 0-200ms time window, there was no need to carry out a topographic or quantitative comparisons between conditions.

5.3.2 300 – 500ms

Amplitude analysis — 300-500ms

To determine the reliability of the N400 effects observed in the waveforms, a multilevel ANOVA was performed on the data from each fluency condition separately to assess for any effects of predictability *(predictable, unpredictable)*, and interactions of predictability with location *(F, FC, C, CP, P,)*, hemisphere *(left, right)* and site *(mid, superior, medial, inferior)*.

For fluent utterances, a multilevel ANOVA revealed a main effect of predictability [$F(1,24) = 53.98$, $\eta^2_G = 0.414$, $p < 0.0001$], reflecting a relative negativity elicited by unpredictable than predictable words. It also revealed an interaction between predictability and site [$F(1.09,26.15) = 28.90$, $\eta^2_G = 0.012$, $p = 0.0001$], reflecting a greater negativity towards the midline, and an interaction between predictability, location and site [$F(2.37,56.91) = 3.52$, $\eta^2_G = 0.0004$, $p < 0.05$] reflecting the centro-parietally maximal nature of the effect as seen in Figure 5.5, consistent with an N400 effect. A follow-up ANOVA incorporating data from midline electrodes *(Fz, FCz, Cz, CPz, Pz, POz)* revealed a significant main effect of predictability [$F(1,24) = 53.37$, $\eta^2_G = 0.453$, $p < 0.0001$],

although the interaction between predictability and location did not reach significance. This confirms the impressing given by Figure 5.5 of a relative negativity for unpredictable words which is broadly spread across the scalp.

For disfluent utterances, the global ANOVA revealed a main effect of predictability [F(1,24) = 9.89, η^2_G = 0.073, p <0.005], again reflecting a greater negativity in response to unpredictable words compared to predictable ones. There was also an interaction between predictability and site [F(1.03,24.68) = 4.57, η^2_G = 0.002, p <0.05], reflecting the fact that the effect was larger towards midline sites, as can be seen in Figure 5.5. A midline ANOVA revealed a main effect of predictability [F(1,24) = 9.39, η^2_G = 0.083, p <0.01], but no significant interaction between predictability and location.

For targets preceded by a beep, the global ANOVA also revealed a main effect of predictability [F(1,24) = 27.64, η^2_G = 0.191, p <0.00005], and an interaction between predictability and site [F(1.09,4.35) = 4.90, η^2_G = 0.002, p <0.05]. There was also a significant interaction between predictability, hemisphere and site [F(1.12,26.89) = 4.08, η^2_G = 0.0004, p <0.05], confirming the impression given in Figure 5.5 (right) that following a beep, the ERPs to targets exhibit greater negativity for unpredictable words than predicable words. This effect is larger towards the midline but with a slightly greater spread in the right than left hemisphere. As there was an interaction of predictability with hemisphere and site, suggesting a lateralisation of the predictability effect, the midline analysis is not reported.

Topographic analysis — 300-500ms

To assess whether the observed effects in the 300-500ms time window differed in topographic distribution, a multilevel global ANOVA was performed on the mean voltage

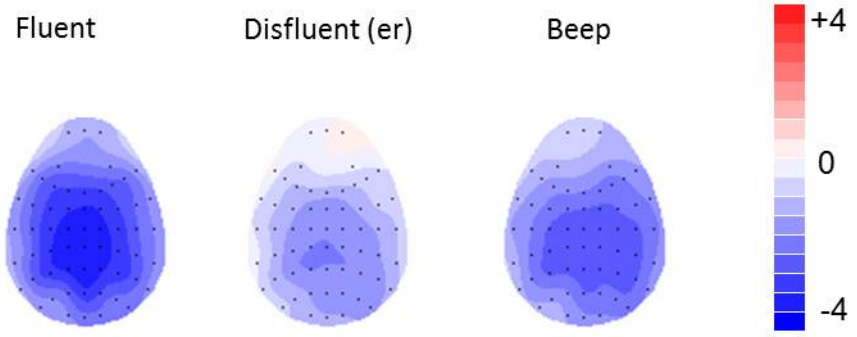

FIGURE 5.5: Scalp topographies (n=25) showing the predictability effects in the 300-500ms time window for targets in utterances which were fluent (left), disfluent (middle) and interrupted by a beep. (right). All three fluency conditions elicit a relative negativity for unpredictable words, which is broadly distributed over the scalp and larger at centro-parietal midline electrode sites.

differences between ERPs for predictable and unpredictable targets, which had been re-scaled using the McCarthy and Wood max-min method (McCarthy & Wood, 1985). The ANOVA employed factors of fluency *(fluent, er, cough)*, location *(F, FC, C, CP, P)*, hemisphere *(left,right)* and site *(superior, medial, inferior)* The ANOVA revealed a main effect of fluency [F(1.98, 47.59) = 3.86, $\eta^2_G = 0.068$, p <0.05], but no interactions between fluency and any other factors. As the observed effects appear stronger towards the midline, a follow up ANOVA examined data from six midline electrodes with factors of fluency *(fluent, er, cough)* and location *(F, FC, C, CP, P, PO)*. This midline ANOVA revealed no main effects or interactions involving fluency. Thus, on the basis of global and midline analyses, there is no evidence for any distributional difference between the effects, and there is no reason to assume that different neural generators underlie the effects observed in fluent, disfluent and interrupted conditions.

Quantitative Comparison — 300-500ms

Having established that all three fluency conditions produce ERPs consistent with the presence of an N400 effect, and that these effects do not vary topographically, a factor of

fluency *(fluent, er, beep)* was added to the global ANOVA to allow a quantitative comparison of effects between fluency conditions. As there was evidence of a hemispheric bias for the beep condition, this analysis drew on data from the electrode array used in the global analysis, rather than the midline. This analysis was performed with the intention of investigating the way predictability effects vary with fluency. Performing a subtraction on the ERPs would have increased the noise in the data, and so both levels of predictability are used instead. As such, only significant interactions implicating a factors of predictability and fluency are of interest to the experimental manipulation, and it is only these that are reported here. The ANOVA revelealed a main effects of predictability [$F(1,24) = 65.69$, $\eta^2_G = 0.167$, $p < 0.0001$] and fluency [$F(1.69,40.49) = 21.58$, $\eta^2_G = 0.199$, $p < 0.0001$], as well as an interaction between fluency and predictability [$F(1.98,47.59) = 3.63$, $\eta^2_G = 0.020$, $p < 0.05$], indicating that the size of the N400 effect was significantly affected by the fluency condition of utterances.

Finally, visual inspection of the data (absolute magnitudes at each electrode) confirmed that CPz represented the effect maxima in all three fluency conditions. A follow-up comparison concentrated on data from this electrode. This comparison confirmed the impression given by Figure 5.6, that there is a marginally significant difference between the size of the N400 effect elicited by fluent (mean = -3.66, sd = 2.59) and disfluent (mean = -1.89, sd = 3.53) utterances [$t(24) = -2.03$, $p = 0.054$], but no significant difference between fluent and interrupted (mean = -2.74, sd = 2.66) [$t(24) = -1.23$, $p > 0.2$], or disfluent and interrupted conditions [$t(24) = 1.02$, $p > 0.3$].

5.3.3 600 – 900ms

Inspection of data in the 600-900ms time window reveals a relative positivity for unpredictable compared to predictable words at left posterior electrodes. As can be seen

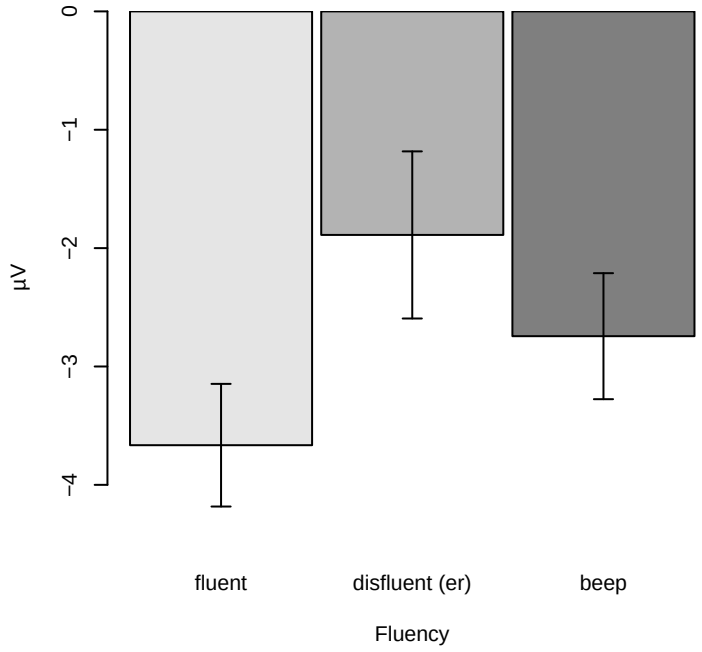

FIGURE 5.6: Mean voltage difference between unpredictable and predictable targets at the CPz electrode in the 300-500ms time window (n=25). Error bars represent one standard error of the mean. Unpredictable target words in fluent utterances elicited the most negative ERP relative to predictable words. The N400 effect for targets in disfluent conditions was reduced compared to fluent utterances. For target words competing utterances interrupted by a beep, the absolute magnitude of the N400 falls between that of fluent and disfluent utterances, although it does not differ significantly from either.

in Figure 5.7, for disfluent and interrupted utterances, this positivity appears to extend forward to frontal midline electrodes.

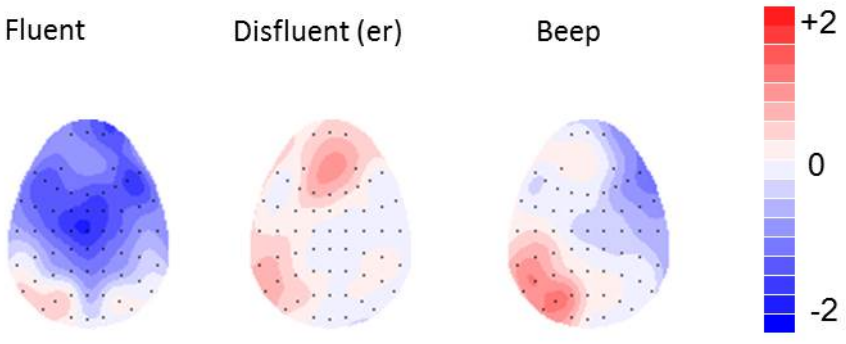

FIGURE 5.7: Scalp topographies (n=25) showing predictability effects in the 600-900ms time window for targets in utterances which were fluent (left), disfluent (middle) and interrupted by a beep (right).

Amplitude analysis — 600-900ms

To establish whether the visually observed effects in the 600-900ms time window are reliable, each fluency condition was analysed separately in an ANOVA incorporating factors of predictability *(predictable, unpredictable)*, location *(F, FC, C, CP, P)*, hemisphere *(left, right)* and site *(mid, superior medial, inferior)*.

A multilevel ANOVA incorporating data from fluent utterances revealed no main effect of predictability, but did reveal significant interactions of predictability with site [$F(1.07, 25.69) = 4.36$, $\eta^2_G = 0.002$, $p = 0.044$], and predictability with location and site [$F(2.30,55.13) = 11.10$, $\eta^2_G = 0.001$, $p < 0.0001$]. This reflects a negativity for unpredictable words compared to predictable words, which is greater towards the midline than lateral sites, and that this focus over the midline is greater towards fronto-central sites than parietal sites.

For data from disfluent utterances, a multilevel ANOVA revealed no main effect of predictability. There was, however, an interaction of predictability with location and site [$F(2.63,63.20) = 7.38$, $\eta^2_G = 0.0006$, $p < 0.0005$], and an interaction of predictability with location, hemisphere and site [$F(2.71, 65.08) = 4.14$, $\eta^2_G = 0.0002$, $p < 0.05$]. These results reflect the presence of a relative positivity for unpredictable compared to predictable words at left hemispherere and midline sites. At frontal sites the effect is larger towards the midline, whereas at posterior sites, the effect is larger over left than right hemisphere sites (as can be seen in Figure 5.7).

A multilevel ANOVA incorporating data from utterances interrupted by a beep revealed no main effect of predictability, but did reveal interactions of predictability with hemisphere [$F(1,24) = 5.14$, $\eta^2_G = 0.003$, $p < 0.05$], predictability with location and site [$F(3.09,74.10) = 6.34$, $\eta^2_G = 0.0005$, $p < 0.001$], and predictability with hemisphere and

site [F(0.87, 20.86) = 4.42, η^2_G = 0.0005, p <0.05]. This comfirms the impression given in Figure 5.7 of a left parietal positivity at inferior electrodes and at frontal midline sites for unpredictable compared to predictable words.

As each of the previous sets of analyses have revealed significant effects of hemisphere, follow-up midline analyses were not performed.

Topographic analysis — 600-900ms

To assess for topographic differences between any effects found in the 600-900ms time window, a multilevel global ANOVA was carried out on the rescaled mean voltage differences between ERPs for predictable and unpredictable targets. The ANOVA employed factors of fluency *(fluent, er, cough)*, location *(F, FC, C, CP, P)*, hemisphere *(left,right)* and site *(superior, medial, inferior)* (see Figure 4.2). This global ANOVA revealed no main effects or interactions involving fluency. A follow-up multilevel ANOVA examining data from six midline electrodes *(F, FC, C, CP, P, PO)*, and incorporating factors of fluency and location, also revealed no main effects or interactions involving fluency. As there is no evidence for any distributional difference between the effects, there is no reason to assume that different neural generators underlie the effects observed in the fluent, disfluent and interrupted conditions, and so the conditions were quantitatively compared.

Quantitative Comparison — 600-900ms

As all three fluency conditions revealed significant predictability effects in the 600-900ms time window, these were quantatively compared using a multilevel ANOVA with factors of fluency *(fluent, disfluent, beep)*, predictability *(predictable, unpredictable)*, location *(F,*

FC, C, CP, P), hemisphere *(left, right)* and site *(F, FC, C, CP, P, PO)*. This revealed a main effect of fluency [F(1.49, 35.86) = 18.29, η^2_G = 0.132, p <0.0001], reflecting the fact that words following a beep elicited more positive ERPs than words which had been fluent or disfluent. There was also a significant interaction of fluency with location [F(2.26,54.15) = 30.73, η^2_G = 0.048, p <0.0001], which reflects the fact that the positivity seen in the 600-900ms time window is parietally maximal for fluent utterances, broadly spread for disfluent utterances, and fronto-centrally maximal for targets in utterances interrupted by a beep. The ANOVA also revelealed an interaction of predictability with location [F(1.25, 30.04) = 5.05, η^2_G = 0.002, p <0.05]. This reflects a gradient across the scalp when considering predictability effects collapsed across fluency conditions, with frontal sites exhibiting a relative negativity for unpredictable words, and posterior sites being more positive. Importantly, there were no significant interactions between fluency and predictability, meaning that there is no reason to assume any differences between predictability effects reported in fluent, disfluent and interrupted conditions. Finally, as there was not a clear effect maxima, and there was not a theoretical reason to expect one, no single electrode focussed analysis is reported.

5.3.4 Effects over time

In order to establish whether the effects reported in the 0-200ms, 300-500ms and 600-900ms time windows should be considered to be separate and distinct, or continuations of one another, the data were analysed for differences in topography and magnitude over time.

	0-200			300-500			600-900		
Global ANOVA	fluent	er	beep	fluent	er	beep	fluent	er	beep
Predictability	*			*	*	*			
Predictability : Location									
Predictability : Hemisphere									*
Predictability : Site				*	*	*	*		
Predictability : Location : Hemisphere									
Predictability : Location : Site				*			*	*	*
Predictability : Hemisphere : Site						*			*
Predictability : Location : Hemisphere : Site								*	
Midline ANOVA									
Predictability	*			*	*	N/A	N/A	N/A	N/A
Predictability : Location						N/A	N/A	N/A	N/A

TABLE 5.2: Summary of statisitcally significant ($p < 0.05$) results from global and midline ANOVA analyses ERPs to targets in fluent, disfluent and interrupted utterances.

300-500ms quantitative analysis	
Fluency	*
Predictability	*
Fluency : Predictability	*
Fluency : Predictability : Location	
Fluency : Predictability : Hemisphere	
Fluency : Predictability : Site	
Fluency : Predictability : Location : Hemisphere	
Fluency : Predictability : Location : Site	
Fluency : Predictability : Hemisphere : Site	
Fluency : Predictability : Location : Hemisphere : Site	

TABLE 5.3: Summary table of signifcant main effects and interactions ($p < 0.05$) from a quantitative analysis of ERPs to targets in fluent, disfluent and interrupted utterances in the 300-500ms time window.

0-200ms — 300-500ms

For fluent utterances, a global ANOVA was run with factors of epoch *(0-200ms, 300-500ms)*, location *(F, FC, C, CP, P)*, hemisphere *(left, right)* and site *(superior, medial, inferior)* and drawing on rescaled predictability effects data. This ANOVA revealed a main effect of epoch [$F(1, 24) = 34.10$, $\eta^2_G = 0.123$, $p < 0.0001$), as well as an interaction between epoch and site [$F(1.07, 25.64) = 12.17$, $\eta^2_G = 0.003$, $p < 0.005$] reflecting the fact that the predictability effects are strongly focussed on the midline in the 0-200ms time window, whereas in the 300-500ms window, the effects are less tightly midline-focussed. A follow-up ANOVA incorporating data from midline sites revealed a significant main effect of epoch [$F(1, 24) = 29.50$, $\eta^2_G = 0.125$, $p < 0.0001$], and a marginally significant interaction of epoch with location [$F(1.33, 31.84) = 3.55$, $\eta^2_G = 0.003$, $p = 0.1$]. These results reflect a more negative effect in the later epoch, and that this increase in negativity over time is greater at central and centro-parietal locations.

As topographic analysis has revealed differences in the distributions of predictability effects for fluent utterances at 0-200ms and 300-500ms, these are not quantitatively compared. Critically, because the distributions of the predictability effects differed between the early (0-200ms) and mid (300-500ms) epochs, the early effect cannot be identified as an early onset of the N400, but should instead be considered a separate effect, which may be identified as a Phonological Mis-match Negativity (PMN). By contrast, analysis of the data for disfluent utterances and utterances interrupted by a beep did not elicit any significant effects in the 0-200ms time window, and so these are not compared to the 300-500ms time window.

300-500ms — 600-900ms

A global ANOVA on ERPs in fluent utterances revealed no main effect of epoch between the 300-500ms and 600-900ms time windows. There were, however, significant interactions of epoch with location [$F(1.25, 30.09) = 14.44$, $\eta^2_G = 0.0003$, $p < 0.0005$], epoch with site [$F(1.04, 24.90) = 8.95$ $\eta^2_G = 0.006$, $p < 0.01$] and a three way interaction of epoch with location and site [$F(2.24, 53.69) = 4.24$, $\eta^2_G = 0.013$, $p < 0.05$]. These significant results reflect the fact that the fluency effect is focussed at fronto-central midline sites in the later time window, but broadly spread over the scalp in the N400 time window. A midline ANOVA also revealed a significant interaction of epoch with location [$F(1.74, 41.75) = 6.23$, $\eta^2_G = 0.005$, $p < 0.01$]. As the predictability effects elicited by fluent utterances in the 300-500ms time window and the 600-900ms time window differ topographically, they are not quantitatively compared.

For disfluent utterances, a global ANOVA revealed no main effect of epoch, but did show an interaction of epoch with site [$F(1.06, 25.54) = 7.20$, $\eta^2_G = 0.002$, $p < 0.05$], along with an interaction of epoch with location and site [$F(1.87, 44.93) = 6.74$, $\eta^2_G = 0.0004$, $p < 0.005$]. This reflects the fact that in the 300-500ms time window, the predictability effect is most negative over midline sites at centro-parietal locations, whereas in the later time window, the effect has a positive focus at left parietal sites and at midline frontal sites. A midline ANOVA also revealed an interaction of epoch with location [$F(1.35, 32.30) = 6.23$, $\eta^2_G = 0.005$, $p < 0.01$] again reflecting the shift from negative to positive between the N400 epoch and the later epoch, and that this shift is greater at frontal and central sites. As the effects in the 300-500ms time window differ topographically from the effects in the 600-900ms window, no quantitative comparison was carried out.

For utterances interrupted by a beep, a global ANOVA revealed interactions of epoch

with location [F(1.15, 27.66) = 8.38, $\eta^2_G = 0.009$, p <0.01], epoch with site [F(1.10, 26.44) = 5.47, $\eta^2_G = 0.001$, p <0.05] and epoch with location and site [F(2.69, 64.59) = 6.77, $\eta^2_G = 0.0003$, p <0.001]. There was no main effect of epoch. This reflects the fact that the effect in the 600-900ms time window was greater at inferior sites at posterior locations and midline sites at frontal locations, wheras in the earlier 300-500ms time window, there was a negativity focussed around the midline at central and centro-parietal electrodes. A follow-up midline ANOVA also revelealed an interaction of epoch with location [F(1.74, 41.75) = 6.23, $\eta^2_G = 0.005$, p <0.01], reflecting the positive shift occurring between epochs, and that this shift was greater at posterior locations. As there is a topographic difference between the effects seen in the 300-500ms and 600-900ms epochs for utterances interrupted by a beep, these time windows are not quantatively compared.

In all three fluency conditions, the N400 effects differ topographically from the predictability effects seen in the later (600-900ms) time window. Consequently, these effects should not be considered a continuation of the N400, but rather separate and distinct activity.

5.4 Summary and Discussion

The experiment described in this chapter sought to investigate the role of delay in accounting for ERP effects which have previously been reported following disfluency. Corley et al. (2007) found that when highly constrained sentences ending in predictable and unpredictable target words contained a filler disfluency before the target, N400 amplitudes for unpredictable compared to predictable words were reduced. In an attempt to systematically investigate the trigger for this effect, we built on Corley's (2007) design, adding a third fluency condition. This condition comprised an artificial beep, edited in

to the recording of each fluent sentence. Each beep was time matched to the length of the naturally produced filler *(er)* in the corresponding disfluent sentence. This allowed a clear contrast between the phonetic form of fillers and the delay they introduce into an utterance. An artificial beep edited into recordings cannot be taken to indicate speaker difficulty, and so any effects conferred by beeps must be assumed to be a function of the delay added to the speech signal, rather than a sympathetic response to speaker difficulty.

5.4.1 N400

The findings of Corley et al. (2007) are replicated in the current experiment. Unpredictable words elicit an N400 effect compared to predictable words. Moreover, the magnitude of this effect was reduced following a disfluency. The third condition, investigating the effect of an artificial beep on the N400 effect, produced an N400 whose magnitude lay between that elicited by fluent and disfluent utterances, but did not differ significantly from either.

These findings are broadly in line with previous research, which has reported reduced N400 effects for words following disfluency. Compared with fluent utterances, words following an artificial beep tone appear to elicit a slightly reduced N400 effect, although this reduction does not appear to be statistically significant. This reduction may suggest that the delay introduced by the filler plays some role in facilitating the effects observed following disfluency but may not be the only factor governing the process. It is possible that the attenuation in the N400 seen for disfluent utterances relies not only on either the delay or the phonetic form of the filler, but a combination of the two. If the delay allows comprehension processes to unfold more fully, this does not preclude listeners simlutaneously using their experience of disfluency distribution to begin modifying their

expectations, upon hearing a speaker become disfluent. As such, it may be that the N400 attenuation, reported by Corley et al. (2007) and replicated here following disfluencies, is a result of two or more cumulative processes occurring as part of the speech comprehension process.

Alternatively, the finding that the magnitude of the N400 elicited by predictable and unpredictable words following an artificial beep does not align with either the effect produced by fluent or disfluent sentences may reflect some ambivalence for listeners about how to process this stimulus. It seems plausible that humans are by nature hard-wired to expect naturalistic language input, and an artificial beep mid-speech may be the source of some confusion in processing terms.

5.4.2 600-900ms

In the later 600-900ms time window, unpredictable words elicit a left parietal positivity compared to predictable words. The timing and distribution of this effect are consistent with the late positivity reported by MacGregor (2008) in a similar experiment. It is possible that this late positivity represents a weak Late Positive Complex (LPC), thought to reflect memory retrieval and control processes (Federmeier et al., 2007; Van Petten et al., 1991). If this is the case then it may be associated with holding the unpredictable word in working memory while the sentence context is re-activated, to resume comprehension of the message. This late positivity was significantly influenced by fluency, with fluent utterances eliciting much smaller positivity than disfluent (*er*) or interrupted (beep) utterances.

If this effect is to be interpreted as an LPC, assumed to reflect memory control and retrieval processes, then the larger effect seen following *ers* or beeps may indicate that

memory is being engaged to a greater extent when there has been temporal disruption in the utterance. In these cases, it may require more from the memory system to retrieve the original sentence context than when there has been no delay, as in the fluent condition.

5.4.3 0-200ms

Comparison of the ERP waveforms for predictable and unpredictable words revealed an apparent early onsetting negativity. This was analysed over the 0-200ms time window, and found to be significant for fluent utterances only. Comparison of the effect in the early (0—200ms) and the N400 (300—500ms) time windows revealed signficant differences in both magnitude and topography, with the effect being stronger and more widespread in the later window.

This may be interpreted as a Phonological Mis-match Negativity (PMN), as participants very quickly realise that the phoneme at the beginning of the target word is not the initial phoneme of their expected target word. The fact that this effect is not present for disfluent and interrupted utterances may reflect the fact that the delay interposed by the *er* or the beep allows phonological expectations to subside. This would mean that it would take listeners longer to detect that a target word is not the most expected.

5.4.4 Discussion

The effects detected in the early and late time windows are consistent with participants being sensitive to delay introduced to an utterance, at least at a perceptual and functional level. Of primary interest to the aims of this study, however, is how delay affects prediction as indexed by the N400 effect. Consequently, the remainder of this discussion

will focus primarily on how the N400 effects observed in this experiment inform our understanding of the effects of disfluency processing.

In this study, an artificially inserted beep was used to add delay to the utterance without connotations of speaker difficulty. One point worth considering is that the lack of a disfluent filler explicitly indicating speaker difficulty does not necessarily mean that listeners do not assume speaker difficulty on detecting a delay. It is possible that a disruption to the rhythm of speech does more than simply allowing the unfolding of comprehension processes, but that listeners begin changing their expectancy for upcoming material as soon as a rhythmic disruption is detected. This would represent something of an efficient short-cut for the listener. That said, there is evidence to suggest that speakers do not indiscriminately change their expectancy following disfluency without reference to knowledge about speaker state; no effect of disfluency was found when participants in a visual world paradigm were told that the speaker suffered from object agnosia (Arnold et al., 2007). However, it is unclear how far Arnold's findings generalise to more natural speech contexts, without closed referent sets or task demands.

Further, results showed that the magnitude of the N400 effect for interrupted utterances does not appear to align with either fluent or disfluent utterances. Statistically, the magnitude of N400 effect for interrupted utterances did not differ from either fluent or disfluent utterances. This may indicate that the data collected are simply too noisy to detect a potentially subtle fluency effect between two conditions. Alternatively, given the relatively small difference between the mean sizes of the N400 effects for fluent and disfluent utterances ($1.77\mu V$) and the noise associated with auditory data, it may not be feasible to collect data with little enough variance to detect differences between fluent and interrupted, or disfluent and interrupted utterances, particularly if the true

mean voltage for interrupted utterances does lie between that for fluent and disfluent utterances.

Another point worth considering is the extent to which beeps may affect attention. A prominent N1 — P2 complex indicates that participants were sensitive to the change in the acoustic properties of the stimulus — from artificial beep to speech. It is possible that this artificial beep raised participants attention by nature of it's acoustic salience compared to the rest of the stimuli. If a beep raised or oriented attention to a greater degree than a disfluent filler, then any reduction of the N400 seen following a beep may have been attained using a different mechanism than that achieved by disfluent fillers. If the N400 is attenuated when attention is not oriented to the stimulus (Otten, Rugg, & Doyle, 1993), then it might be logical to extrapolate that with increased attention, N400 amplitude should also increase. If this is the case, and both disfluent fillers *(er)* and beeps raise attention, but by nature of their acoustic salience, beeps orient attention to a greater extent than fillers, then we would expect the size of N400 seen following a beep to be larger than that seen following a filler, even if the underlying effects which had first caused a reduction from the N400 size elicited in fluent utterances had been the same.

This theory would suggest that while a beep may have added the same amount of delay to an utterances as the equivalent filler, it may have had the effect of raising attention to a greater extent than did fillers. Given that the aim of the experiment was to establish which features of fillers were responsible for triggering the attenuation in N400 reported by Corley et al. (2007), and replicated here, the possibility of differences in attention levels following fillers and beeps is problematic. This would make it impossible to say whether any difference in N400 between disfluent and interrupted utterances was a function of different processes initiated by the phonetic form of the filler, which inform

listeners of speaker difficulty, or whether the underlying processes were the same between the two conditions, but differing attention levels had caused an increase in N400 effects for interrupted utterances.

To test this, it would be logical to attempt to add delay to an utterance in a way which does not have the same auditory salience as an artificially created sine-wave beep, and which has more ecological validity, whilst not suggesting speaker difficulty. It is this challenge which will be addressed in the following chapter.

Chapter 6

Disfluency as Delay: Comparing the Immediate Effects of Fillers and Coughs

6.1 Introduction

This chapter describes an experiment which investigates the contribution of delay to disfluency effects by directly contrasting traditional filler disfluencies (*er, um*) with non-disfluent interruptions. These non-disfluent interruptions take the form of speaker-generated coughs, which add delay to an utterance by plausibly interrupting the speech stream, but which are not assumed to indicate speech planning difficulties.

In the previous chapter, an experiment contrasting disfluent fillers against artificial sine wave beeps revealed a reduced N400 for words following fillers compared to words completing fluent sentences. Where sentence final words were preceded by a beep, the

magnitude of the elicited N400 effect showed a small numeric reduction, compared to fluent target words in fluent sentences, although this failed to reach significance. If we were to further investigate this slight numeric reduction, and find it to be a real effect, then this would appear to implicate the delay introduced into an utterance by a filler in eliciting the N400 attenuation reported by Corley et al. (2007). One critical difference, however, between an artificially inserted beep and a filler, is that a filler is speaker generated. Like beeps, coughs add delay to an utterance, but unlike beeps, coughs are speaker generated and so may be considered to be somewhat closer to disfluent fillers.

Why a cough?

As was discussed in the previous chapter (Section 5.1), one of the biggest challenges to addressing the question of delay and phonetic form is the difficulty in identifying a suitable control condition. Attempting to replicate the phonetic form of fillers without introducing delay is necessarily problematic, and so most researchers wishing to address this issue have instead focused on finding a way to introduce delay without the phonetic form of the filler.

Studies using a silent pause as a control delay condition produce generally mixed results, and are all subject to the criticism that silence does not necessarily represent a delay free of connotations of speaker difficulty. Silences can be disfluent (Maclay & Osgood, 1959), and have specifically been associated with difficulty at the grammatical and articulatory level of speech production (Fraundorf & Watson, 2008). If listeners hear a delay in speech than it is natural to assume that the speaker has stopped for a reason. Generally, in dialogue, speakers like to hold the floor until they have finished their declaration, and since the cessation of the speech stream risks losing the floor, speakers are unlikely to stop

the utterance, unless they are prevented from continuing; either by conceptualisation, formation or articulation difficulties, or by external factors.

Noisy interruptions, such as doorbells and barking dogs were used as a delay control condition by Bailey and Ferreira (2003) in a syntactic judgement task. However, with specific regard to the predictive effects of disfluency, these types of noisy external interruptions may not have enough ecological validity to be useful. If, in reality, speakers may prefer to talk over such interruptions, listeners may not accept these as external reasons for the cessation of speech. They may instead implicitly assume that if the speaker has stopped, then they must be experiencing difficulties in conceptualising planning or articulating the utterance. This was also the implication of Arnold et al. (2007), who reported no change in listeners responses to disfluency when fillers were preceded by noisy distractions. Presumably, in this experiment, participants did not believe the speaker would stop simply because of the noise, and so interpreted that they must be experiencing production difficulties. This would allow listeners to begin updating their expectancy based on an assumption about speaker state.

A third option for introducing delays is to consider what else may plausibly prevent speech from being produced. Whilst silences imply formation difficulties and noisy interruptions imply willingness by speakers to pause utterances, something outside of the speaker's control, which prevents the physical production of speech, may present a reasonable option. A cough is an involuntary reflex action to clear the airway of irritants and foreign objects. Coughs often originate in the upper respiratory tract, particularly when a person has a viral infection such a cold, as swelling in the throat is misinterpreted by the sympathetic nervous system as irritation from a foreign object (Irwin, Rosen, & Braman, 1977; NHS, 2013). As such, a cough can be considered a reflex action beyond a speaker's control, produced in the respiratory tract and so preventing

the physical production of speech. Coughs may therefore be considered a plausible reason for cessation of an utterance without implicating difficulty with the conceptualisation or formation of speech.

Coughs have previously been used as a control noise condition to investigate discourse level memory (Fraundorf & Watson, 2011), and listeners on-line comprehension of spoken instructions (Barr, 2001; Barr & Seyfeddinipur, 2010). Barr and Seyfeddinipur (2010) contrasted disfluent fillers (*ums*) with time-matched coughs and sniffles. Mouse movement was tracked as listeners selected one of two images on a computer screen, as they were described by pre-recorded spoken instructions, produced by one of two speakers. All of the targets were abstract irregular shapes, and spoken instructions were either fluent, or contained an *um* or cough at the beginning of the description of the image. In addition, to build up items' status as either given or discourse-new, trials were presented in blocks, with the final pair of images constituting the experimental trial. Of the pair of images presented in the final trial, one had previously been described by that speaker, one had not. Results revealed that listeners began moving the mouse towards the item that was discourse-new for the speaker when they heard the speaker say *um* at the beginning of the utterance, but not when they heard a cough or sniffle. This finding suggests that listeners use specific understanding about what causes disfluency in speakers to predict what they will say.

This experiment differs from many others in that the disfluencies are introduced right at the beginning of an utterance. Although coughs and fillers were time-matched, such that in these two conditions, speech always onset at 4871ms, the delay condition (cough) is not clearly comparable to the mid-sentence fillers typical of most disfluency studies. If the function of disfluency is to (at least partially) allow the unfolding of comprehension processes during the delay it introduces, then this would require that comprehension

has already begun for this effect to be evident. Whilst an utterance intitial *um* gives some context in terms of alerting the listener to speaker difficulty, comprehension of the speech has not yet ensued and so no benefit from delay could be derived, either in the cough or the *um* condition.

In the experiment described in this chapter, coughs serve as a delay condition with which to compare disfluent fillers. As in the previous chapter, this experiment is an extension of the paradigm used in Corley et al. (2007) and so disfluencies and coughs are presented mid sentence, directly before target words. Participants listened to utterances for understanding, and ERPs were measured in the absence of any secondary task.

Summary

One of the main difficulties in attempting to isolate the effects of form and delay in disfluency studies is finding an appropriate control condition. If the listener is to believe that the speaker stopped speaking then any noisy interruption must be plausible. Coughs are a reflex action governed mainly by the sympathetic nervous system and so outside of a speaker's control. Importantly, they originate in the vocal tract, and take priority over speech articulation. Upon hearing a speaker cough, a listener should not be surprised to hear speech temporarily suspended, but should also not attribute this delay to problems with conceptualisation or formation of the ongoing utterance.

Coughs are used in this experiment by way of a mid-utterance delay control condition, contrasting with disfluent fillers and fluent utterances. This study follows the format of the previous chapter in terms of experimental design. A reduced N400 effect is expected for disfluent compared to fluent utterances. If this reduction is driven by the delay a disfluency introduces to the utterance, then the same reduction is expected for targets

preceded by coughs. If, on the other hand, the delay is a function of the form of the filler, explicitly indicating to participants that the speaker is in difficulty, then no reduction is expected for targets preceded by coughs, given that coughs cannot be interpreted as indicating speaker difficulty.

6.2 Methods

Stimuli

Stimuli consisted of the same 324 highly constrained predictable and unpredictable utterances as in the previously described experiment. Again, one third of the utterances were fluent and one third contained a filler (*er, um*) before the target word. The remaining sentences contained a cough before the target word before the target. Coughs were produced by the speaker mid-way through carrier sentences, and then edited in to the fluent recording of each experimental utterance. Appending coughs to the fluent sentences rather than editing out fillers from disfluent utterances ensured that there were no prosodic cues which would indicate disfluency' to the listener. Coughs for each utterance were carefully selected to be time matched to the filler in the equivalent disfluent sentence (tolerance +/-7ms). For a full description of the stimuli, and how they were generated, along with some example stimuli, see Chapter 4.3.

Participants

Twenty four right-handed native English speakers (5 male, mean age 20.5years; range 18-30 years) took part in the experiment.

Procedure

Testing followed the procedure of the previous experiment. The first half of the experiment focussed on online processing, while the second focussed on subsequent recognition memory. During the first half, participants were instructed to listen naturally for understanding as they heard the stimulus sentences. They were asked to respond to simple yes/no comprehension questions which followed forty of the eighty filler utterances. The stimuli were presented in four blocks of approximately 12 minutes each, interspersed with short breaks.

The second half of the experiment consisted of a surprise memory test, in which utterance final target words from the stimulus sentences were displayed on the screen, interspersed with frequency matched 'new' words, which had not appeared anywhere in the aural stimuli. Participants were asked to respond as quickly and accurately as possible to discriminate between old and new words using two buttons on the button box. Buttons were counterbalanced across participants. EEG was recorded from the scalp using the Neuroscan Quickcap system throughout the experiment. See Section 4.7 for a fuller explanation.

Data analysis

To form ERPs for each condition, epochs were formed which were time-locked to the onset of the target word. Epochs began 100ms before onset, and continued for 2000ms from the onset of the target words. Analyses were based on the ERPs of 24 participants who provided a minimum of 16 trials per condition. For details of the number of trials incorporated into the analysis for each condition, see Table 6.1. ERPs were quantified by measuring the mean amplitude over time windows of interest. There was no behavioural

measure in the listening section of the experiment, although to confirm that participants were attending to the stimuli, their responses to the yes/no comprehension questions were checked. All participants responded to these questions with at least 97% accuracy.

6.3 ERP results

ERPs to target words occurring in fluent utterances showed a relative negativity for unpredictable words. The negativity onset around 200ms after the onset of the target word and was maximal at 400ms, lasting until 600ms at left hemisphere sites and a little longer at right hemisphere sites. This negativity was greater towards the midline and at parietal sites, and more pronounced in the right than left hemisphere. Around 600ms, the negativity for unpredictable words was overtaken by a relative positivity focussed around left hemisphere parietal electrodes. This positivity lasted until 950ms.

Both predictable and unpredictable words preceded by a disfluent filler exhibited an early complex with a positive peak at 90ms, followed by a negative peak at around 150ms. The positivity had a broad fronto-central distribution, whereas the ensuing negativity was larger over the midline and at fronto-central, central and centro-parietal locations. Unpredictable words presented in disfluent sentences elicited a relative negativity compared to fluent words. The negativity onset at 250ms, and reached its maximum at

	fluent		er		beep	
	predictable	unpredictable	predictable	unpredictable	predictable	unpredictable
minimum	24	21	20	25	26	23
maximum	54	54	54	54	54	53
mode	54	53	52	33	53	53
mean	42.09	41.22	42.00	41.52	40.70	41.35

TABLE 6.1: Numbers of trials included in ERP analysis for each condition (n=24).

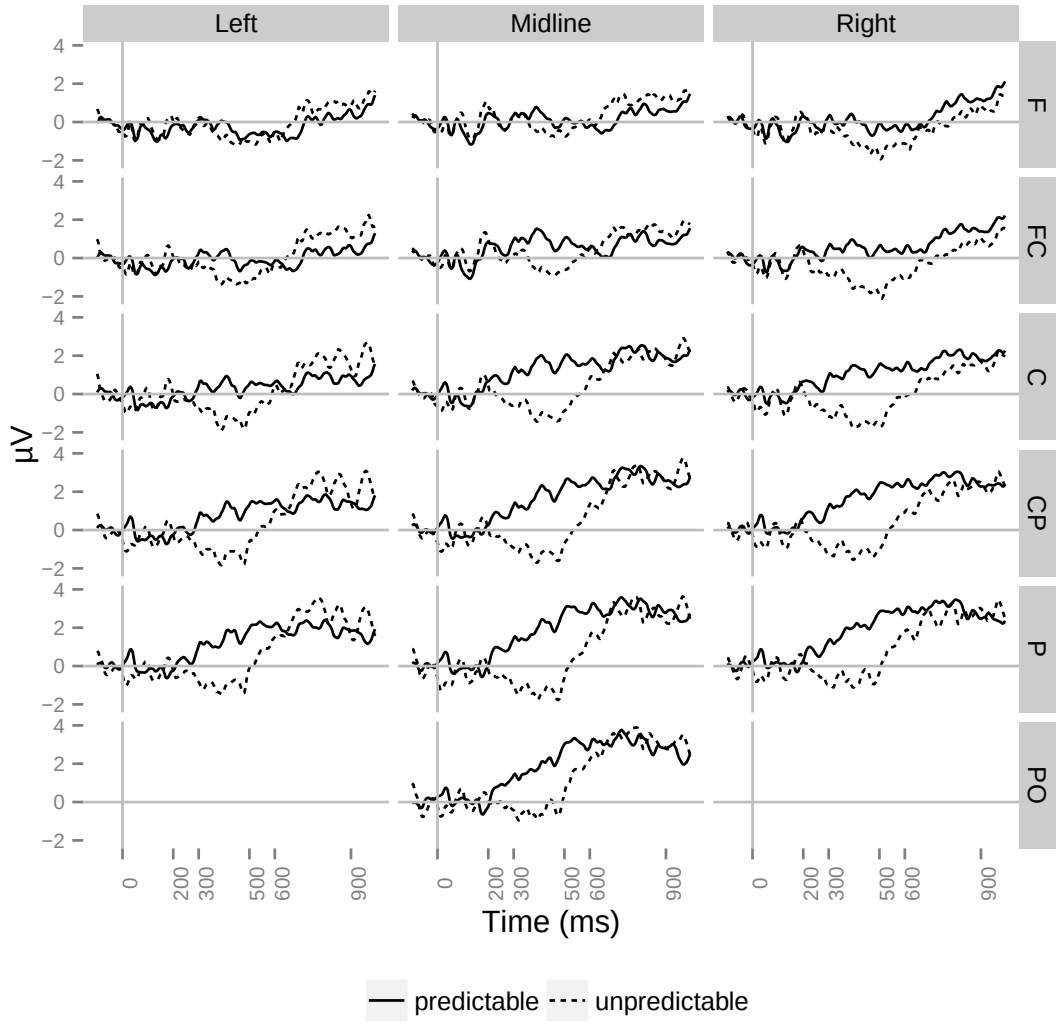

FIGURE 6.1: Grand average ERPs (n=24) for final words in *fluent* utterances. Shown here are ERPs as measured at frontal (F), fronto-central(FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline. Unpredictable words elicit a relative negativity, which onsets around 250ms and reaches a maximum at 400ms. The negativity is larger at centro-parietal and parietal sites towards the midline, and slightly larger in the right than left hemisphere. Around 600ms after stimulus onset, unpredictable words show a more positive ERP than predictable words in the left hemisphere. This relative positivity lasts until 950ms.

450ms. This effect was greater at midline sites, and over centro-parietal electrode locations. The effect was slightly longer lasting at the front than rear of the scalp, where it was replaced by a relative positivity for unpredictable words. This positivity was seen between 650ms and 920ms, and was greater over the left than right hemisphere.

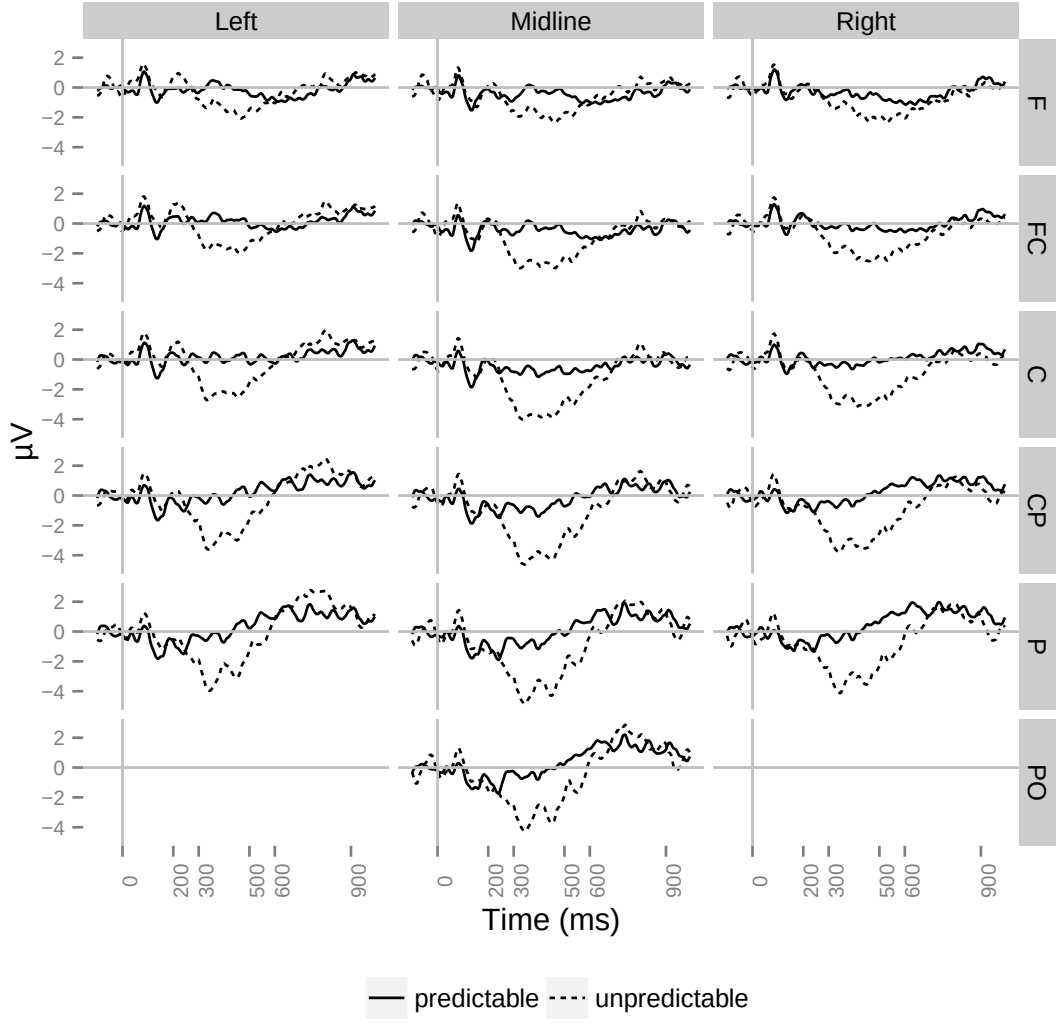

FIGURE 6.2: Grand average ERPs (n=24) for final words in *disfluent* utterances. Shown here are ERPs as measured at frontal (F), fronto-central(FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline. Both unpredictable and predictable words show a centro-frotnally maximal positive peak at 90ms followed by a negative peak at 150ms, which is greater over the midline. Around 250ms, there begins a relative negativity for unpredictable words. The negativity is maximal at 450ms and broadly distributed, although it is greater towards parietal sites. Towards the rear of the scalp, and particularly in the left hemisphere, this negativity is followed by a relative positivity for unpredictable words, lasting from 650ms to 920ms.

Following a cough, both unpredictable and predictable target words elicited a negative peak at 150ms which was larger for predictable than unpredictable words. A relative negativity for unpredictable words onset at 250ms, and reached a maximum at 470ms. This negativity was greater at parietal and cento-parietal sites, and towards the midline. Around 600ms, this negativity gave way to a positivitity for unpredictable words, onsetting first at frontal midline electrode sites, and spreading into the left hemisphere, developing two distinct maxima around 650ms — one over fronto-central sites and the other over left-parietal sites. Around 900ms, this positivity became weaker with a broader whole-head topographic distribution, lasting throughout the remainder of the epoch.

In all fluency conditions, the predictability effects shown appear to be consistent with the timing and topography of an N400 effect. As this is key to the experimental manipulation, analysis is focussed on the 300-500ms time window. Analysis is also performed on data from the 200-300ms time window, as predictability effects appear to onset during this time, and on data from the 600-900ms time window, where all fluency conditions revealed a relative positivity for unpredictable words.

To assess the significance of the effects observed in each time window of interest, each fluency condition was analysed separately in ANOVAs with factors of predictability *(predictable, unpredictable)*, location *(F, FC, C, CP, P)*, hemisphere *(left, right)* and site *(mid, superior, medial, inferior)*. Where this broad ANOVA revealed significant effects of interest, and no lateralisation of effects, this was followed up with a second ANOVA, incorporating data from the midline, and factors of predictability *(predictable, unpredictable)* and location *(F, FC, C, CP, P)*.

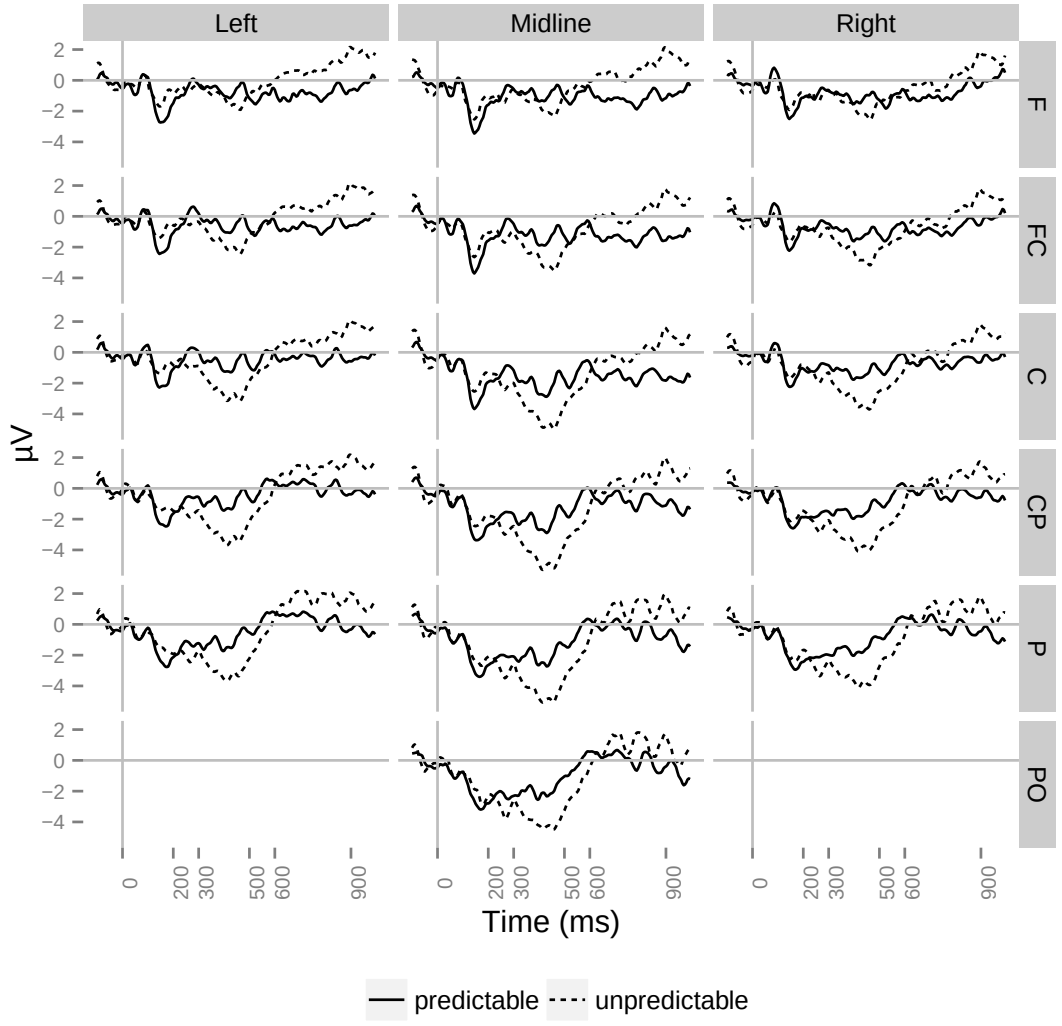

FIGURE 6.3: Grand average ERPs (n=24) for final words in utterances interrupted by a *cough*. Shown here are ERPs as measured at frontal (F), fronto-central(FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline. Predictable and unpredictable words both show a negative peak 150ms after stimulus onset, and this is larger for predictable than unpredictable words. Unpredictable words elicit a relative negativity beginning at 250ms, and continuing until 600ms, with a maxima at 470ms. The negativity is greater at parietal and centro-parietal sites, and towards the midline. At 600ms, there onsets a relative positivity for unpredictable words. This appears first at frontal midline electrode sites, before spreading into the left hemisphere and developing two distinct maxima at 650ms — one over fronto-central sites and one over left parietal sites. This pattern persists until 900ms, when it loses its identity, resolving into a broad whole-head weak positivity, persisting throughout the remainder of the epoch (—2000ms).

6.3.1 200-300ms

The relative negativity for unpredictable words appears to onset at or soon after 200ms for all three fluency conditions tested (200ms, 250ms, 250ms respectively). Thus predictability effects in the 200-300ms time window are analysed in order to establish whether this early negativity should be considered significant, and to allow a comparison establishing whether this negativity should be considered an early onset of the N400, or a separate and distinct effect.

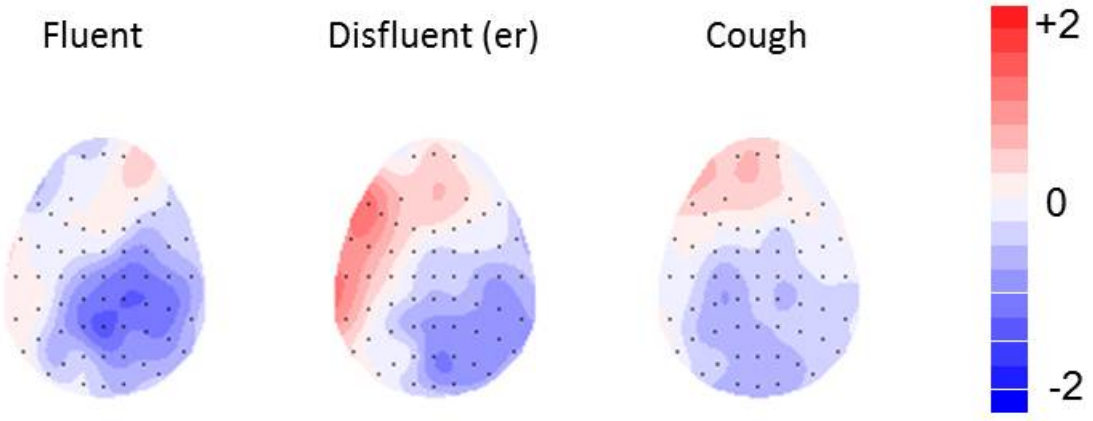

FIGURE 6.4: Scalp topographies (n=24) showing predictability effects over the 200-300ms time window, for targets in utterances which where fluent (left), disfluent (middle) and interrupted by a cough (right).

Amplitude analysis — 200-300ms

For ERPs to target words in fluent utterances, a multilevel global ANOVA with factors of predictability *(predictable, unpredictable)*, location *(F, FC, C, CP, P)*, hemisphere *(left, right)* and site *(superior, medial. inferior*) revealed no significant effects or interactions involving predictability. A follow-up midline ANOVA incorporating data from six midline locations *(F, FC, C, CP, P, PO)* and with factors of predictability *(predictable, unpredictable)* and location also revealed no significant effects or interactions.

For targets in disfluent utterances, the global ANOVA revealed no main effect of predictability, but did show an interaction of predictability with hemisphere [F(1,23) = 4.76, $\eta^2_G = 0.004$, p <0.05]. This reflects the fact that in the 200-300ms time window, there is no overall positive or negative effect associated with predictability; rather, there is a relative negativity in the right hemisphere and towards the rear or the scalp and a relative positivity in the left hemisphere and towards the front of the scalp. The ANOVA also revealed an interaction of predictability with hemisphere and site [F(1.12,25.78) = 8.76, $\eta^2_G = 0.001$, p <0.01], reflecting the fact that in the right hemisphere, the predictability effect is spread across sites at all levels (superior, medial and inferior), whereas in the left hemisphere, the effect is stronger at inferior sites than towards the midline. As this analysis implicated a factor of hemisphere, the midline analysis is not reported.

For targets in utterances interrupted by a cough, neither the global nor midline ANOVAs revealed any main effects or interactions involving predictability.

6.3.2 300–500ms

Amplitude analysis — 300-500ms

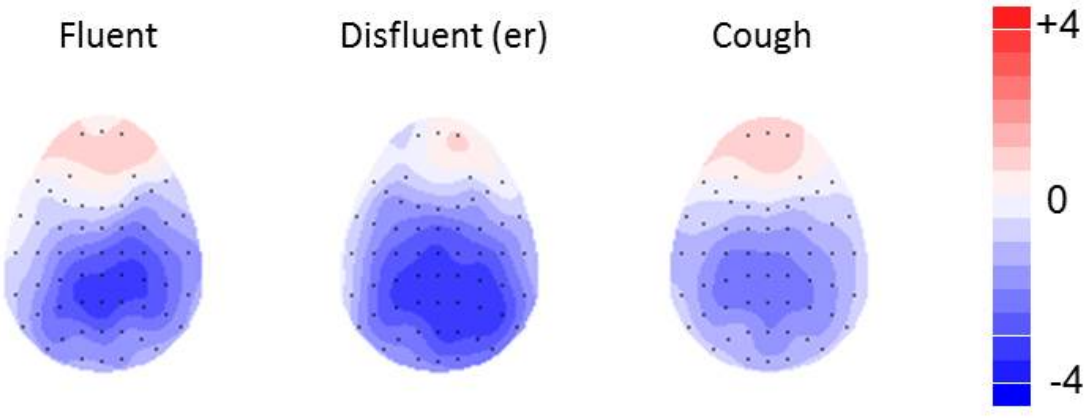

FIGURE 6.5: Scalp topographies (n=24) showing predictability effects over the 300–500ms time window, for targets in utterances which where fluent (left), disfluent (middle) and interrupted by a cough (right).

To characterise the pattern of ERP effects, each fluency condition was analysed separately using a global ANOVA incorporating factors of predictability *(predictable, unpredictable)*, location *(F, FC, C, CP, P)*, hemisphere *(left, right)* and site *(superior, medial, inferior)*.

For fluent utterances, a multilevel global ANOVA revealed a main effect of predictability [$F(1,23) = 23.06$, $\eta^2_G = 0.168$, $p < 0.0001$], reflecting a relative negativity elicited by unpredictable words. The ANOVA also revealed a significant interaction between predictability and location [$F(1.05,24.10) = 5.80$, $\eta^2_G = 0.030$, $p < 0.05$], reflecting a greater negativity towards the rear of the scalp, and an interaction between predictability and site [$F(0.32,7.30) = 12.28$, $\eta^2_G = 0.005$, $p < 0.005$], reflecting the centrally maximal nature of the effect. Additionally, there was a marginally significant interaction of predictability with hemisphere [$F(1,23) = 4.09$, $\eta^2_G = 0.006$, $p < 0.1$], reflecting a slight bias towards the right hemisphere.

A follow-up midline analysis incorporating factors of predictability *(predictable, unpredictable)* and location *(F, FC, C, CP, P, PO)* revealed a main effect of predictability [$F(1,23) = 27.74$, $\eta^2_G = 0.175$, $p < 0.00005$], and an interaction of predictability with location [$F(1.19,27.42) = 4.62$, $\eta^2_G = 0.039$, $p < 0.05$], reflecting a predictbility effect which is greater at centro-parietal and parietal electrode locations. As can be seen in Figure 6.5 (left) fluent utterances elicit a characteristic N400 effect: a centro-parietal midline maximum negativity.

For disfluent utterances, there was a main effect of predictability [$F(1,23) = 20.99$, $\eta^2_G = 0.190$, $p < 0.0005$], along with significant two-way interactions between predictability and location [$F(1.16,24.4) = 11.57$, $\eta^2_G = 0.017$, $p = 0.005$], as well as predictability and site [$F(1.11,25.76) = 9.80$, $\eta^2_G = 0.004$, $p < 0.005$]. As for fluent utterances, the predictability

effect is greatest towards the midline and towards the rear of the scalp. In addition, there was a significant interaction between predictability, location, hemisphere and site [F(2.95,67.74) = 4.06, η^2_G = 0.0002, p = 0.05], reflecting the fact that the predictability effect is more spread across levels of site in the right than left hemisphere, and that this gradient difference is only significant at locations where the effect is strongest — at centro-parietal and parietal locations. As the predictability effect is strongest towards the midline, the global ANOVA was followed up with a midline analysis. This midline analysis revealed a main effect of predictability [F(1,23) = 23.43, η^2_G = 0.212, p <0001], and a significant interaction of predictability with location [F(1.48,34.01) = 4.51, η^2_G = 0.011, p <0.05]. As can be seen in Figure 6.5 (middle), the ERP effect for disfluent utterances also exhibits a typical N400 pattern, being larger at centro-pariatal electrodes, and increasing in size towards the midline at posterior, but not anterior locations.

For utterances interrupted by a cough, the global ANOVA revealed a main effect of predictability [F(1,23) = 10.98, η^2_G = 0.079, p <0.005], and interactions of predictability with location [F(1.09,24.34) = 7.21, η^2_G = 0.011, p <0.05], as well as predictability and site [F(1.06,22.3) = 8.00, η^2_G = 0.002, p <0.05]. As for fluent and disfluent utterances, ERPs to unpredictable words are more negative than those to predictable words, and this effect is larger at posterior locations, as well as being larger towards the midline. A follow-up midline analysis also revealed a main effect of predictability [F(1,23) = 10.88, η^2_G = 0.083, p <0.005], and a marginally significant interaction of predictability with location [F(1.27,29.12) = 3.63, η^2_G = 0.009, p <0.01]. As Figure 6.5 (right) confirms, words preceded by coughs also elicited a clear N400 effect.

Topographic analysis — 300-500ms

Before comparing the predictability effects elicited in the three reported fluency conditions, it is necessary to establish that there is no topographic difference between their distributions. A multilevel global ANOVA was performed on the rescaled mean voltage differences between ERPs for predictable and unpredictable targets employing factors of fluency *(fluent, er, cough)*, location *(F, FC, C, CP, P)*, hemisphere *(left,right)* and site *(superior, medial, inferior)* (see Figure 4.2). The analysis revealed no main effect of fluency, nor any interactions involving fluency (all F's >0.32). A follow up midline ANOVA examining data from six midline electrodes with factors of fluency *(fluent, er, cough)* and location *(F, FC, C, CP, P, PO)* also revealed no main effect of fluency, nor any interaction between fluency and location (all F's >0.30). Taken together, these analyses therefore provide no evidence for significant differences in the distribution of ERP effects elicited across the fluent, disfluent and interrupted conditions, confirming the impression given by Figure 6.5.

Quantitative Comparison — 300-500ms

The preceding analyses demonstrate that N400 effects are present for fluent, disfluent and interrupted conditions, and furthermore, that these effects do not differ in topography. An additional analysis was carried out to establish whether the size of the N400 effects differed across conditions, adding a factor of fluency *(fluent, er, cough*) to the global and midline ANOVAs. As in the previous chapter, this made use of the original ERP waveforms rather than subtraction waveforms to avoid amplifying noise in the data; thus the outcomes of interest are those implicating predictability. The quantitative comparison of effects between the three conditions using a multilevel global ANOVA revealed

main effects of fluency [F(1.80,41.47) = 6.83, η^2_G = 0.077, p <0.005] and predictability [F(1,23) = 34.79, η^2_G = 0.141, p <0.00001]. Importantly, no interactions involving predictability and fluency reached significance, indicating that this experiment has not revealed evidence that the manipulation of fluency had any effect on the magnitude of predictability effects.

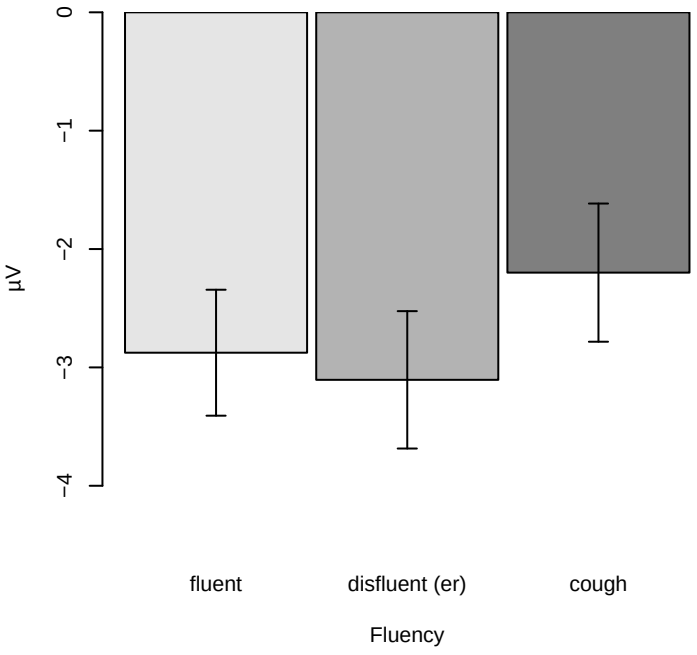

FIGURE 6.6: Mean voltage difference between unpredictable and predictable targets at the CPz electrode in the 300-500ms time window (n=24). Error bars represent one standard error of the mean. No significant differences in predictability effects were found across fluency conditions.

Finally, pairwise comparisons were carried out on data from electrode CPz (capturing the maxima of the N400 effect); the mean magnitude is illustrated in Figure 6.6. The analysis confirmed the impression given by the figure that there are no significant differences in amplitude between ERPs to utterances which are fluent (mean = -2.88, sd = 2.61), disfluent (mean = -3.11, sd = 2.84) and interrupted by a cough (mean = -2.20, sd = 2.86).

6.3.3 600 - 900ms

Inspection of the data in the 600-900ms time window reveals a relative positivity for unpredictable words. This appears to be distributed over left hemisphere parietal electrodes and fronto-central midline sites.

Amplitude analysis — 600 - 900ms

To establish whether these apparent predictability effects in the 600-900ms time window are reliable, each fluency condition was analysed separately in an ANOVA incorporating factors of predictability *(predictable, unpredictable)*, location *(F, FC, C, CP, P)*, hemisphere *(left, right)* and site *(mid, superior medial, inferior)*.

For ERPs to target words in fluent utterances, a multilevel ANOVA revealed no main effect of predictability, but did show an interaction of predictability with hemisphere [$F(2,23) = 13.08$, $\eta^2_G = 0.024$, $p < 0.005$], reflecting the fact that the predictability effect took the form of a relative positivity in the left hemisphere, and a relative negativity in the right. There were also interactions of predictability with location and site [$F(1.26,28.89) = 7.05$, $\eta^2_G = 0.004$, $p < 0.01$] and predictability with hemisphere and site [$F(1.12,25.72) = 12.38$, $\eta^2_G = 0.002$, $p < 0.005$], reflecting the fact that the predictability effect is more evenly spread across sites towards posterior locations than at frontal locations, and more evenly spread across sites in the left than right hemisphere.

Similarly, for disfluent utterances, a multilevel ANOVA revealed no main effect of predictability, but did reveal an interaction of predictability with hemisphere [$F(1,23) = 10.83$, $\eta^2_G = 0.012$, $p < 0.005$]. This again reflects the fact that the predictability effect renders a predominant relative positivity in the left hemisphere, and a relative negativity in the right. There were also significant interactions of predictability with location and

site [F(1.97,45.26) = 12.25, η^2_G = 0.001, p <0.0005] and predictability with hemisphere and site [F(1.08,24.81) = 6.56, η^2_G = 0.002, p <0.05]. These results reflect a predictability effect which is spread more evenly across sites in the left than right hemisphere, and spread more evenly across sites at frontal and posterior locations than at central locations.

A multilevel global ANOVA for ERPs to targets in utterances interrupted by a cough revealed a main effect of predictability [F(1,23) = 10.17, η^2_G = 0.045, p <0.005], and an interaction of predictability with hemisphere [F(1,23) = 5.55, η^2_G = 0.003, p <0.05], reflecting the fact that unpredictable words elicited a more positive ERP in the 600-900ms time window than predictable words, and that this positivity was stronger in the left hemisphere. There were also significant interactions of predictability with site [F(1.11,25.56) = 5.08, η^2_G = 0.002, p <0.05] and predictability with hemisphere and site [F(1.51,34.74) = 5.63, η^2_G = 0.0006, p <0.05], reflecting the fact that the predictability effect was greater towards the midline, and that this gradient towards the midline was larger in the right hemisphere. Additionally, there was a marginally significant interaction of predictability with location and site [F(1.33,30.53) = 3.23, η^2_G = 0.0009, p <0.1], reflecting an effect which was more midline-based at frontal locations than towards the rear of the scalp.

As the factor of hemisphere was implicated in all three fluency conditions, midline analyses are not reported.

Topographic analysis — 600-900ms

In order to establish whether there were any topographic differences between the predictability effects found for utterances which were fluent, disfluent and interrupted by

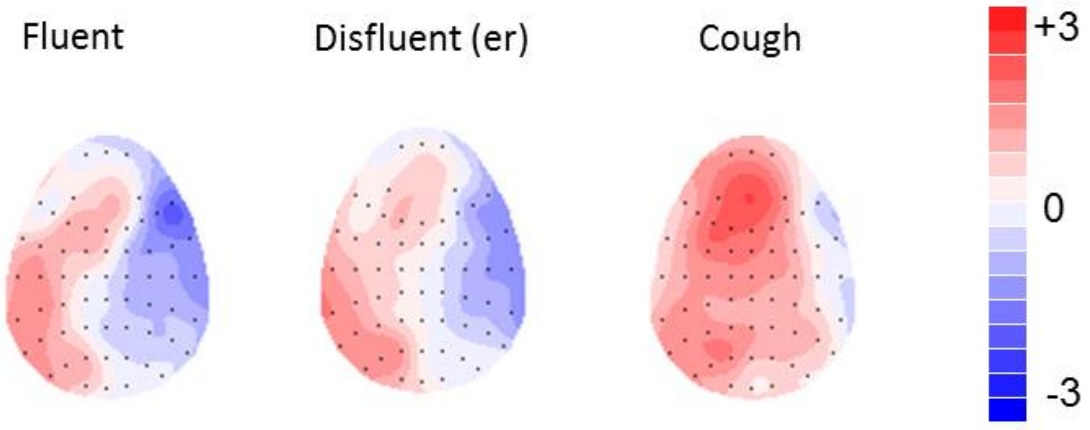

FIGURE 6.7: Scalp topographies showing predictability effects over the 600-900ms time window, for fluent, disfluent and interrupted utterances.

a cough, the rescaled difference waveforms were submitted to a multilevel ANOVA employing factors of fluency *(fluent, er, cough)*, location *(F, FC, C, CP, P)*, hemisphere *(left,right)* and site *(superior, medial, inferior)*. This ANOVA revealed no main effect of fluency, nor any interactions involving fluency. A follow-up midline ANOVA, with factors of fluency *(fluent, er, cough)* and location *(F, FC, C, CP, P, PO)*, also revealed neither a main effect of fluency nor any effect involving fluency. Thus, together, these analyses provide no evidence for a distributional difference in the predictability effects across fluency conditions; consequently, these data are quantitatively compared below.

Quantitative Comparison — 600-900ms

Having established that all three fluency conditions elicit significant predictability effects in the 600-900ms time window, and that these do not vary topographically, a quantitative comparison was performed with the aim of discovering whether these predictability effects vary across fluency conditions. The data were submitted to a mulitilevel ANOVA with factors of fluency *(fluent, disfluent, beep)*, predictability *(predictable, unpredictable)*, location *(F, FC, C, CP, P)*, hemisphere *(left, right)* and site *(superior, medial, inferior)*. As in previous quantitative comparisons, only significant outcomes

implicating predictability are of interest. Although the ANOVA revealed a main effect of fluency [F(1.60,36.86) = 5.04, $\eta^2_G = 0.043$, p <0.05], there were no significant interactions implicating fluency and predictability. Thus there is no evidence to suggest that the predictability effects reported here vary across fluency conditions.

6.3.4 Effects over time

In order to establish whether the effects reported in the 200-300ms, 300-500ms and 600-900ms time windows should be considered to be separate and distinct, or continuations of one another, the data were analysed for differences in topography and magnitude over time.

200-300ms — 300-500ms

Targets in fluent utterances did not elicit any significant predictability effects in the early window when analysed by either the global or midline ANOVAs, and so for fluent utterances, the early (200-300ms) and mid (300-500ms) epochs are not compared.

To establish whether the predictability effects for disfluent targets varied topographically across epochs, rescaled data were submitted to an ANOVA with factors of epoch *(0-200ms, 300-500ms)*, location *(F, FC, C, CP, P)*, hemisphere *(left, right)* and site *(superior, medial, inferior)*. This global ANOVA revealed a main effect of epoch [F(1,23) = 62.37, $\eta^2_G = 0.328$, p <0.00001], as well as significant interactions of epoch with hemisphere [F(1,23) 6.42, $\eta^2_G = 0.003$, p <0.05], epoch with hemisphere and site [F(1.13,26.03) = 9.87, $\eta^2_G = 0.0006$, p <0.005] and epoch with location, hemisphere and site [F(2.66,61.22) = 4.13, $\eta^2_G = 0.0001$, p <0.05]. These interactions confirm the impression given by comparing Figures 6.4 (middle) and 6.5 (middle), of a predictability effect which is more

	200-300			300-500			600-900		
Global ANOVA	fluent	er	cough	fluent	er	cough	fluent	er	cough
Predictability				*	*	*			*
Predictability : Location				*	*	*			
Predictability : Hemisphere		*		(*)			*	*	*
Predictability : Site				*	*	*			*
Predictability : Location : Hemisphere									
Predictability : Location : Site							*	*	(*)
Predictability : Hemisphere : Site		*					*	*	*
Predictability : Location : Hemisphere : Site					*				
Midline ANOVA									
Predictability		N/A	*	*	*	*	N/A	N/A	N/A
Predictability : Location		N/A		*	*	*	N/A	N/A	N/A

TABLE 6.2: Summary of significant [*] (p <0.05) and marginally significant [(*)] (p <0.1) results from global and midline ANOVA amplitude analyses of ERPs to targets in fluent, disfluent and interrupted utterances.

300-500ms Quantitative Comparison	
Fluency	*
Predictability	*
Fluency : Predictability	
Fluency : Predictability : Location	
Fluency : Predictability : Hemisphere	
Fluency : Predictability : Site	
Fluency : Predictability : Location : Hemisphere	
Fluency : Predictability : Location : Site	
Fluency : Predictability : Hemisphere : Site	
Fluency : Predictability : Location : Hemisphere : Site	

TABLE 6.3: Summary table of significant main effects and interactions (p <0.05) from a Quantitative Comparison of ERPs to targets in fluent, disfluent and interrupted utterances in the 300-500ms time window.

600-900ms Quantitative Comparison	
Fluency	*
Predictability	
Fluency : Predictability	
Fluency : Predictability : Location	
Fluency : Predictability : Hemisphere	
Fluency : Predictability : Site	
Fluency : Predictability : Location : Hemisphere	
Fluency : Predictability : Location : Site	
Fluency : Predictability : Hemisphere : Site	
Fluency : Predictability : Location : Hemisphere : Site	

TABLE 6.4: Summary table of signifcant main effects and interactions (p <0.05) from a Quantitative Comparison of ERPs to targets in fluent, disfluent and interrupted utterances in the 600-900ms time window.

negative in the right hemisphere in the earlier epoch, but not in the later (300-500ms) epoch. This hemisphere bias also interacts with site, with the gradient across sites being reversed in each hemisphere in the early epoch only, and this interaction is more prominent at central and centro-parietal sites. A follow-up midline ANOVA also revealed a main effect of epoch [$F(1,23) = 39.50$, $\eta^2_G = 0.274$, $p < 0.00001$], but the interaction of epoch and location did not reach significance. As the global topographic analysis has revealed evidence for topographic differences between predictability effects across the two epochs, these predictability effects are not quantitatively compared. Because the early-epoch predictability effect differs significantly in topographic distribution from the later N400 effect, it cannot necessarily be identified as an early-onset of the N400, but may have to be considered to be a separate effect.

For targets in utterances interrupted by a cough, no significant predictability effects were revealed in the early time window, and so no comparison between the early (200-300ms) and mid (300-500ms) epochs is reported.

300-500ms — 600-900ms

Comparison of the topography of rescaled predictability effects for fluent utterances in the 300-500ms epoch with the 600-900ms epoch revealed significant differences in their topographic distributions. A multilevel global ANOVA revealed a main effect of epoch [$F(1,23) = 176.23$, $\eta^2_G = 0.457$, $p < 0.00001$], as well as significant interactions of epoch with hemisphere [$F(1,23) = 9.59$, $\eta^2_G = 0.005$, $p < 0.01$], epoch with location [$F(1.43,32.98) = 18.74$, $\eta^2_G = 0.019$, $p < 0.00005$], epoch with site [$F(1.16,26.79) = 16.64$, $\eta^2_G = 0.003$, $p < 0.0005$] and epoch with hemisphere and site [$F(0.62,28.32) = 7.34$, $\eta^2_G = 0.0004$, $p < 0.01$]. These significant effects reflect the fact that the later epoch is generally more positive, and that this difference is greater in the left hemisphere; the

difference between ERPs in the mid and later epochs is larger at locations towards the rear of the scalp; the later epoch has an overall shallower gradient across sites, and this gradient is reversed across sites in each hemisphere in the 600-900ms epoch only. Given the different topographies of the predictability effects in the two time windows, no quantitative comparison between the two is reported.

The topography of predictability effects elicited by words in disfluent utterances also varied across the 300-500ms and 600-900ms time windows, confirming the impression given by Figures 6.5 (centre) and 6.7 (centre). A mulitlevel global ANOVA on the rescaled difference data revealed a main effect of epoch [$F(1,23) = 69.99$, $\eta^2_G = 0.370$, $p < 0.00001$], as well as interactions of epoch with hemisphere [$F(1,23) = 15.08$, $\eta^2_G = 0.007$, $p < 0.001$], epoch with location [$F(1.22,28.05) = 15.29$, $\eta^2_G = 0.013$, $p < 0.0005$], epoch with site [$F(1.04,24.06) = 15.62$, $\eta^2_G = 0.004$, $p < 0.001$], epoch with location and site [$F(2.32,53.26) = 14.87$, $\eta^2_G = 0.0009$, $p < 0.00001$] and epoch with hemisphere and site [$F(1.23,28.35) = 9.93$, $\eta^2_G = 0.0006$, $p < 0.005$]. These outcomes reflect the fact that the predictability effect takes the form of a broad centro-parietally maximal negativity in the earlier time window, whereas in the later epoch, this effect is generally more positive, particularly in the left hemisphere. In addition, the difference between epochs is more pronounced towards posterior locations and towards the midline. Given the distributional difference between effects in the two epochs, they are not quantitatively compared.

For targets in utterances interrupted by a cough, there were also distributional differences between the predictability effects elicited between the two epochs. A multilevel global ANOVA revealed a main effect of epoch [$F(1,23) = 65.62$, $\eta^2_G = 0.471$, $p < 0.00001$], as well as interactions of epoch with hemisphere [$F(1,23) = 6.15$, $\eta^2_G = 0.003$, $p < 0.05$], epoch with location [$F(1.32,30.27) = 7.47$, $\eta^2_G = 0.009$, $p < 0.01$], epoch with

site [F(1.07,24.53) = 20.43, η^2_G = 0.010, p <0.0005] and epoch with hemisphere and site [F(1.31,30.18) = 9.66, η^2_G = 0.0006, p <0.005]. This confirms the impression given by Figures 6.5 and 6.7 of a predictability effect with is broadly negative in the early epoch, and generally positive in the later epoch. In the later epoch this positivity is biased towards the left hemisphere, and the difference between epochs is larger at posterior sites. Once again, no quantitative comparison is performed, due to the differing topographic distributions across epochs.

In all three fluency conditions, the N400 effects differ topographically from the predictability effects seen in the later (600-900ms) time window. Consequently, these effects should not be considered a continuation of the N400, but rather separate and distinct components.

6.4 Summary and Discussion

This experiment sought to confirm and extend on the work reported in Chapter 5 by comparing fillers directly with speaker generated non-disfluent noise. The aim was to introduce delay in as natural a way as possible, causing the listener to believe that the speaker had paused without implicating difficulty in language production. Stimulus utterances were fluent, disfluent or interrupted by a cough. Analysis focused primarily on the N400 effect, comparing the ease of semantic integration of predictable and unpredictable words across three fluency conditions; fluent, disfluent, interrupted by a cough. We anticipated an attenuated N400 effect for disfluent compared to fluent utterances. Additionally, we hypothesised that if delay was key to triggering N400 attenuation, then

utterances interrupted by a cough should also exhibit an attenuated N400. If, however, the form of disfluency is crucial, then utterances interrupted by a cough would not exhibit N400 attenuation.

6.4.1 N400

In direct contrast to the findings of Chapter 5, and our expectations, no differences emerged between N400 affects across conditions.

A pairwise comparison revealed no difference between N400 effects elicited by fluent and disfluent stimuli. This represents a failure to replicate the findings of Corley et al. (2007), and Chapter 5 of this thesis. Given the lack of attenuation of the N400 effect for disfluent utterances, it is difficult to interpret the outcome for targets preceded by a cough. Similar to the results of Experiment 1 (Chapter 5), the N400 effect for targets preceded by a noisy interruption shows a numeric attenuation compared with fluent utterances, but this difference fails to reach statistical significance. This may suggest that the delay introduced by a cough somewhat influences ease of integration as indexed by the N400, but in the absence of significant effects, and the absence of an N400 attenuation for disfluency, this slight numeric reduction for coughs cannot be interpreted. The discussion will now turn to consider whether this null result should be considered to be reliable, and if so, what the implications of this finding are.

6.4.2 600—900ms

Analysis of the later 600-900ms epoch revealed a relative positivity for unpredictable words, following the N400. Comparison of the topographic distributions of this late positivity and the N400 effect confirmed the impression gained by visual comparison

of Figures 6.5 and 6.7 of a significant difference in polarity and distribution across the two time windows. Although visual inspection of the late positivity (*c.f.* Figure 6.7) appears to give the impression of this positivity being stronger and more widespread for utterances interrupted by a cough than fluent or disfluent utterances, this difference did not reach statistical significance, and no significant interactions between predictability and fluency were revealed.

The timing and distribution of this positivity are consistent with its identification as an LPC. Seen in this light, the failure to identify any significant differences in this effect across fluency conditions would lead to the interpretation that memory control processes are engaged equally whether utterances are fluent, disfluent, or interrupted by a cough, and thus that the resumption of the context of the sentence requires the same level of memory engagement, regardless of its fluency.

6.4.3 200—300ms

Visual inspection of the data revealed an early onsetting negativity for unpredictable words, and consequently predictability effects in the 200—300ms time window were examined. This analysis revealed a significant difference between hemispheres emerging for disfluent words, with the right hemisphere exhibiting more negativity than the left. No significant predictability effects are revealed for fluent utterances or those interrupted by a cough.

Where significant predictability effects were revealed, their topographic distribution was compared to the distribution of the effects in the 300-500ms time window. This comparison revealed significant differences between the early and later time windows.

The finding that the early time window predictability effect did not reach significance for fluent or interrupted utterances may not reflect an effect in disfluent utterances only, but rather a spreading negativity which onsets earlier for fluent utterances and interrupted utterances, becoming a whole-head negativity, whereas for disfluent utterances, the data captured in the 200—300ms time window still incorporates frontal and left hemisphere positivity which is being pushed out by the spreading negativity building towards an N400 effect. Although topographic comparison revealed differences between the negativity in the early time window and the N400 for disfluent utterances, it remains possible that the early negativity should be considered to be the early onset of the N400 effect. In this case, the earlier onsetting N400 for fluent utterances (as corroborated by visual inspection of the ERP waveforms, (*c.f.* Figures 6.1, 6.2, 6.3) would be consistent with participants detecting a phonological mismatch between the expected and realised target words faster in fluent utterances than in utterances where a delay has allowed phonological expectations to subside.

6.4.4 Discussion

The primary theoretical motivation for this study was to investigate how semantic expectation, as indexed by the N400, is modulated by delay, and so the discussion will focus primarily on the N400.

We anticipated an attenuated N400 effect for disfluent compared to fluent utterances. Additionally, we hypothesised that if delay was key to triggering N400 attenuation, then utterances interrupted by a cough should also exhibit N400 attenuation. If, however, the form of disfluency is crucial, then utterances interrupted by a cough would not exhibit N400 attenuation. The results of this study, however, revealed no difference in the N400 effect for disfluent compared to fluent words.

To get an indication of whether fluent utterances failed to produce a robust N400 effect, or the N400 effect for disfluent utterances was not attenuated, the magnitudes of the effects were compared between Chapter 5 and the current experiment. The N400 effect to fluent utterances is slightly smaller in the current experiment than in the results of the previous chapter, while N400 effects following disfluency were somewhat larger. This makes it difficult to establish what has caused the difference between the findings of the previous experiment, and the current one. Comparison with the N400 magnitudes reported in Corley et al. (2007) does, however, seem to suggest that a failure of the current experiment to find a difference between fluent and disfluent conditions may be put down to be disfluent *er* utterances failing to produce an attenuated N400, rather then fluent utterances failing to produce large, robust N400 effects.

Assuming that the null hypothesis is true; N400 effects are not reduced in disfluent compared to fluent sentences in this experiment; then it is important to question why the N400 effect might not be attenuated in this experiment. One possibility often raised in discussion around the subject is that the disfluency might not be believable, and somehow did not trigger a disfluency response in the listener. This would be interesting in that it would suggest delay is not the critical mechanism, but rather that listeners must believe in disfluency it to affect their processing. However, this experiment made use of the same recordings as Experiment 1, reported in Chapter 5. Given that these stimuli elicited a "disfluency effect" with an attenuated N400 for disfluent utterances in Chapter 5, it is not clear that lack of 'believable' stimuli should explain the lack of disfluency effects in the current experiment.

Another possibility is that the presence of coughs changed the way disfluency was interpreted. One could hypothesise that as coughs are speaker generated, they may be assumed to be partially under the speaker's control. Anecdotal evidence suggests that

speakers can partially withhold or delay a cough, and so it could be assumed that a cough will be withheld until it is least damaging to the utterance — for example, a point at which there would be a delay anyway, perhaps if the speaker sensed upcoming difficulty. Although these could explain the numeric (but not statistically significant) attenuation of the N400 effect seen for coughs, this does not explain the lack of attenuation for disfluent utterances. Hearing that a speaker is having physical production difficulties could potentially lead listeners to interpret all delay or disfluency as physical production difficulty. This may prevent listeners from updating expectations for material following disfluency, if they believe the *er* results from physical difficulty, rather than conceptualisation or formation difficulty. However, there was no evidence to suggest this effect of coughs neutralising *ums* in Barr and Seyfeddinipur (2010), so to accept this interpretation would be premature without further exploration.

It is possible that individual differences between listeners may affect the way they respond to disfluent fillers. Although participants for Experiments 1 (Chapter 5) and 2 (Chapter 6) were drawn from the same participation pool, using the same criteria, differences will always emerge between individuals selected for studies. If listeners were using perspective taking to update their predictions based on disfluency, and their responses to particular types of disfluency were in any way dependent on their preference for particular types of disfluency in their own speech (which may be influenced by their education/language/dialect background), then differences between individuals, and between groups of individuals may emerge. However, there is as yet no research to indicate that listeners' responses to disfluency are predicated on their own disfluency use.

It remains possible that there is a difference between the N400 effects for fluent and disfluent utterances, but that it is masked by noise in the signal. Inspection of the ERP waveforms (Figures 6.1 to 6.3) reveals the signal-to-noise ratio to be very low. This is

perhaps a function of the use of auditory stimuli, in which the auditory signal develops over time.

So far we have considered the options that the disfluent utterances may not elicit an attenuated N400 because disfluencies are unbelievable; that the presence of coughs may have changed the way disfluent stimuli are processed; that there may have been an effect of participant cohort; and that noise may have distorted the signal, preventing a difference from emerging. Whilst the suggestion that disfluencies may not be believable can be dismissed on the basis of the attenuated N400 effect in Experiment 1, which used the same recorded stimuli, the other three interpretations are at this point still open.

One further point to consider is that not all studies of the N400 and disfluency have reported an attenuation for disfluent targets. Corley et al. (2007) reported an attenuation effect on a small sample (n=12), but when this experiment was repeated with silences, the result is variable, with an N400 attenuation reported in MacGregor, Corley and Donaldson (2010), but not in MacGregor (2008). Disfluent repetitions also failed to produce an attenuation of the N400 (MacGregor et al., 2009). Given that the N400 attenuation in response to disfluency appears to be the exception rather than the rule, the potential of an interaction with other stimuli in the experiment, or the individual differences of participants should not be overlooked. Potential interactions between fluency and other factors are discussed and investigated in the following chapter.

Chapter 7

Beeps, Coughs and Fillers

7.1 Introduction

The experiments detailed in Chapters 5 and 6 have contrasted disfluent fillers with artificial beeps and speaker generated coughs respectively. Whilst the first experiment revealed a small effect of fluency on the size of the N400 effect, this was not found in the second experiment. If these results are an accurate depiction of effects, then it remains to explain why a "disfluency effect" was elicited in one experiment, but not the other. It seems possible that the difference in outcomes between the two experiments reported may be dependent on either context or cohort.

Background

Whilst both of the previously reported experiments displayed clear and robust N400 effects, only one revealed a reduction in the N400 effect following a disfluent filler *(er)*. This difference between the two experiments is particularly unexpected, given that the fluent and disfluent materials did not differ at all between the two experiments. The same

acoustic tokens were used for the fluent and the disfluent utterances in both experiments, as well as the same procedure and task.

A reduced N400 for incongruous words following disfluent hesitations has been reported by Corley et al. (2007) using a reduced version of the paradigm employed in this thesis. In Corley et al.,'s experiment, participants heard fluent and disfluent utterances with predictable and unpredictable final words, but there was no control delay condition. Disfluent utterances did not elicit significant N400 effects when ERPs to predictable and unpredictable words were considered. A similar result was also reported by MacGregor et al. (2010), who found no reliable N400 effect when comparing ERPs to predictable and unpredictable words following silent pauses.

It seems, however, that the effect of disfluency in attenuating the N400 effect is not always consistent. MacGregor (2008) found no difference between N400 amplitudes in fluent utterances and those interrupted by a silent pause. The contrast between MacGregor (2008) and MacGregor et al. (2010) strongly suggests that disfluency effects on the N400 are not entirely reliable.

A disfluency effect on the N400 also failed to appear following repetition disfluencies (MacGregor et al., 2009). Disfluent repetitions were created by copying pre-target words from fluent recordings and splicing them into the fluent utterances immediately before the target. This experimental manipulation produced found no indication of a difference in N400 amplitude between fluent and disfluent utterances. Taken together, these studies seem to indicate that the effect of disfluency on the N400 effect is somewhat complex, and occurs only for some forms of disfluency, or in some experimental contexts, but not others.

It is possible that the linguistic environment in which a disfluency is encountered will affect the way it is interpreted or used by the listener. As previously stated, Arnold et al. (2007) found that listeners' responses to disfluency were suppressed when they believed the speaker would produce disfluencies whether or not they were trying to name a difficult referent. This finding indicates that speakers are able to use some information from the global context of dialogue to make efficient use of cues gleaned from disfluency. If this is the case, it may be plausible to suggest that participants to some degree stopped interpreting disfluency as indicating speaker difficulty when they heard the speaker coughing frequently, as in Chapter 6. It is also possible that listeners may have interpreted the disfluencies as indicating speaker difficulty at the physical production stage, given that the speaker was apparently struggling with this, evidenced by the prevalence of coughs. An interpretation of disfluency as indicating production difficulty, rather than conceptualisation or formation difficulty may prevented listeners from changing their expectations in response to disfluency.

Whilst context effects are entirely plausible, it remains possible that a more mundane explanation would account for the differences across Experiments 1 and 2 (Chapters 5 and 6). By necessity, these two experiments employed different cohorts of participants. It is possible that individual participants vary in their responses to disfluency, which could lead to the emergence of cohort differences reflecting the idiosyncrasies of the individuals recruited for each experiment.

Experimental Design

In this experiment a third layer was added to the structure previously employed in Experiments 1 and 2, which had a 2x3 structure [predictability *(low, high)* x fluency *(fluent, er, interrupted)*]. As the previous experiments seemed to produce contrasting

results, they have been brought together in this experiment, which was separated into two blocks (counterbalanced for order). One block effectively contained half of the stimuli used in Experiment 1 (Chapter 5) (fluent, disfluent and beep-interrupted utterances), while the other block contained half of the material used in Experiment 2 (Chapter 6) (fluent, disfluent and cough-interrupted utterances). As these blocks altered the context of the stimuli (i.e. in one context, listeners were told the speaker had a cold, in the other they were listening to speech interrupted by artificial beeps), we refer to this factor in the experimental design as context. Hence the structure should be considered to be 2 by 3 by 2 [predictability *(low, high)* x fluency *(fluent, er, interrupted)* x context *(cough, beep)*].

Summary

Experiments 1 and 2 varied in their results, one finding an effect of disfluency on the N400, and one finding no such effect. Experiments 1 and 2 differed in two dimensions. Firstly, the global context in which stimuli were presented differed between experiments; in Experiment 1, artificial beeps were edited in to utterances to create delay before the target word; in Experiment 2, the speaker produced coughed and sniffed frequently, and participants were told that the speaker had a cold on the day of recording. Secondly, a different cohort of individuals were recruited for each study. It is possible that either of these factors may account for the differences in results across the two experiments.

In the experiment reported in this chapter, the effects of cohort and context are contrasted by presenting the stimuli in two blocks, one contrasting fillers with coughs, and one contrasting fillers with beeps. If individual differences account for the differing outcomes of the previous experiments, then the pattern of N400 effects between predictable and unpredictable words across fluent and disfluent utterances should not differ over the

two blocks. If, alternatively, the context in which the disfluencies are encountered affects their interpretation, then a difference in the N400 pattern between the two blocks would be expected.

7.2 Methods

Stimuli

Stimuli consisted of the same 324 highly constrained predictable and unpredictable utterances as in the previous two experiments. One third of the utterances were fluent, one third contained a disfluent filler *(er, um)* before the target word, and the remaining third contained a non-linguistic interruption directly before the target. Half of these interruptions took the form of coughs, and half were beeps. These were spliced into fluent sentence recordings, and so contained no prosodic or co-articulatory cues indicating upcoming disfluency. Both coughs and beeps had been carefully time-matched (+/-7ms) to the naturally produced filler in the disfluent recording of the same sentence. Stimuli were presented in two blocks; in one block, the beep-interruption was used, in the other, the cough-interruption. Stimuli were fully counterbalanced so that across participants, each target appeared in each of the twelve possible conditions.

Participants

Twenty four right handed native English speakers (13 male; mean age 21.1 years; age range 18-28 years) took part in the experiment. None of the participants had taken part in either of the previous experiments.

Procedure

Testing followed the procedure of the previous experiments. The first half of the experiment focussed on online processing, while the second focussed on subsequent recognition memory. During the first half, participants were instructed to listen naturally for understanding as they heard the stimulus sentences. They were asked to respond to simple yes/no comprehension questions which followed 40 of the 80 filler utterances. The stimuli were presented in short blocks, of approximately 12 minutes each. After the first two blocks, participants took a break of several minutes, in which they spent four minutes playing a simple non-language game on a handheld gaming device. The game involved moving a character around a maze, and had a repetitive synthesised background tune. This filler task was used to give participants some non-linguistic auditory input, with the aim of clearing echoic memory from the previously heard listening blocks. Following this break, participants listened to a further two blocks of auditory stimuli, in the opposite global context to the one they had heard before the game break.

The second half of the experiment consisted of a surprise memory test, in which utterance final target words from the stimulus sentences were displayed on the screen, interspersed with frequency matched 'new' words, which had not appeared anywhere in the aural stimuli. Participants were asked to respond as quickly and accurately as possible to discriminate between old and new words using two buttons on the button box. Buttons were counterbalanced across participants. Throughout the experiment, EEG was recorded from the scalp using the Neuroscan Quickcap system. See Section 4.7 for a fuller explanation.

	Beep Context						Cough Context					
	predictable			unpredictable			predictable			unpredictable		
	fluent	er	beep	fluent	er	beep	fluent	er	cough	fluent	er	cough
minimum	18	16	16	18	18	18	17	18	18	18	19	17
maximum	26	26	25	26	26	26	26	26	28	28	26	27
mode	22	21	23	20	22	21	22	22	23	25	24	22
mean	21.81	22.52	22.48	22.00	22.00	22.19	22.29	22.62	23.62	22.57	22.52	22.76

TABLE 7.1: Numbers of trials included in ERP analysis for each condition (n=21).

Data analyses

ERPs for each condition were formed by averaging 2000ms epochs which were time locked to the onset of the target word, using a 100ms pre-stimulus baseline. The raw EEG was processed with an artefact rejection amplitude parameter of three standard deviations from the mean, rather than $75\mu V$, as was used for the previous two experiments. Using a parameter of $75\mu V$, only twelve participants reached the minimum threshhold (16 trials per condition) for inclusion. Given that there were twelve conditions, rejection of half of the participants amounted to unacceptable data loss. This was particularly the case as each of the 324 individual stimuli appeared in each of the twelve conditions only twice across the experiment. Using the less conservative artifact rejection parameter of three standard deviations from the mean permitted some increase in noise to enter the ERP data at a single trial level, however this is insignificant compared to the improvement in signal to noise quality achieved by the addition of a further ten subjects into the data used for analysis. See Table 7.1 for details of the number of trials incorporated into the analysis for each condition.

7.3 ERP results

Comparison of the predictable and unpredictable waveforms in the fluent, disfluent and interrupted conditions in each of the context blocks revealed a broadly spread relative

negativity to unpredictable words, onsetting between 200ms and 300ms in all conditions, and dissipating between 530ms and 650ms. This negativity had a central or centro-parietal distribution in all but one condition (beep context, disfluent *er*), in which it had a fronto-central maxima. Broadly speaking, the timing, distribution and polarity of this effect are all consistent with an N400 effect. This was followed later in the epoch by a left hemisphere parietal positivity to unpredictable targets. This positivity is not seen for fluent targets in the cough context or words interrupted by a beep, but appears to a greater or lesser degree for the remaining four conditions (see Figure 7.9). It is unclear whether this should be interpreted as a Late Positive Complex (LPC) as described by Federmeier et al. (2007). Although Federmeier and colleagues described the LPC as having a mid-frontal distribution, they also noted that the positivity extended to posterior electrodes in the left hemisphere. Given the low trial numbers and noisiness of these data, it is possible that the late positivity may represent a LPC, and so the later 600-900ms time window is subjected to analysis below, in Section 7.3.2.

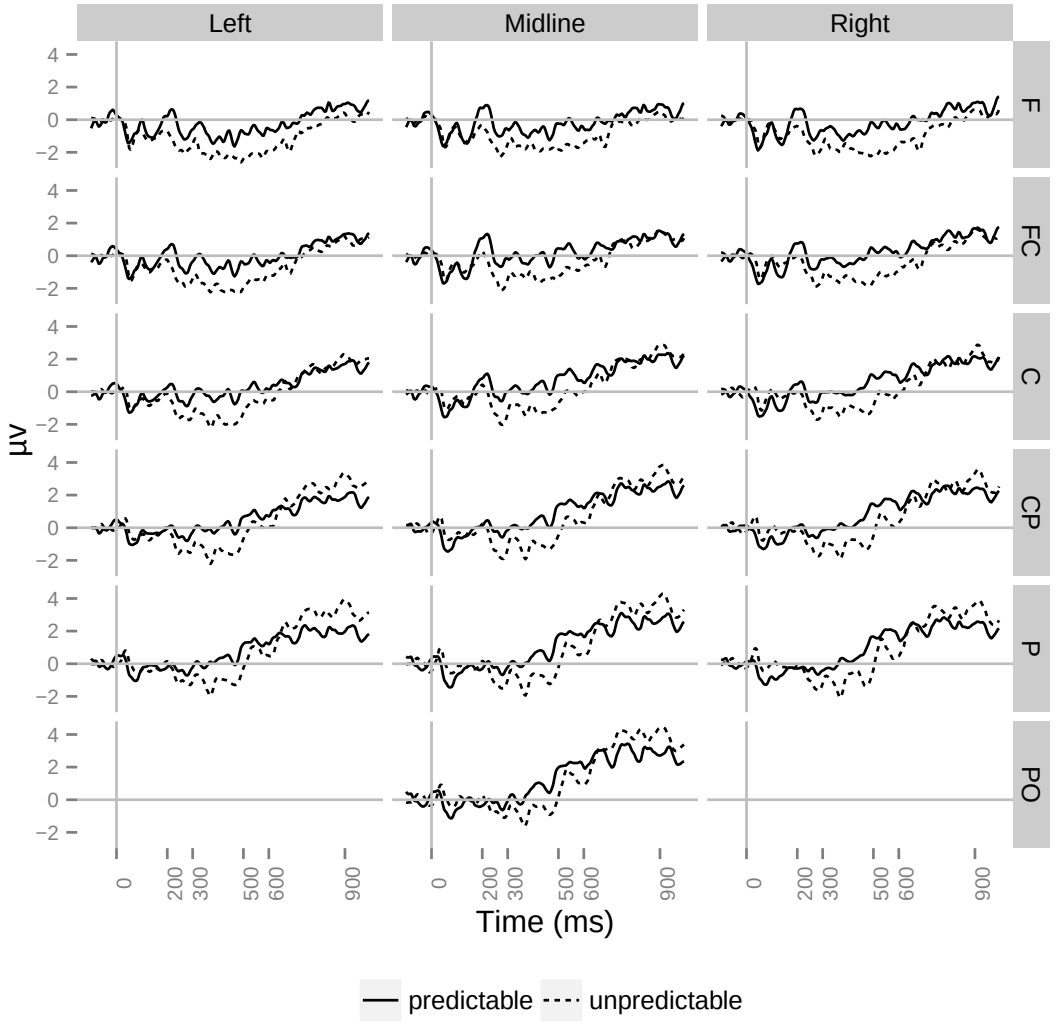

FIGURE 7.1: Grand average ERPs (n=21) for final words in ***fluent*** utterances in the **beep** context. Shown here are ERPs as measured at frontal (F), fronto-central (FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline. Unpredictable words elicit a broadly distributed negativity relative to predictable words. This negativity onsets around 350ms after stimulus-onset, and lasting until around 640ms. From 650ms, unpredictable words appear to elicit a relative positivity at posterior electrodes, which is greater in the left than right hemisphere.

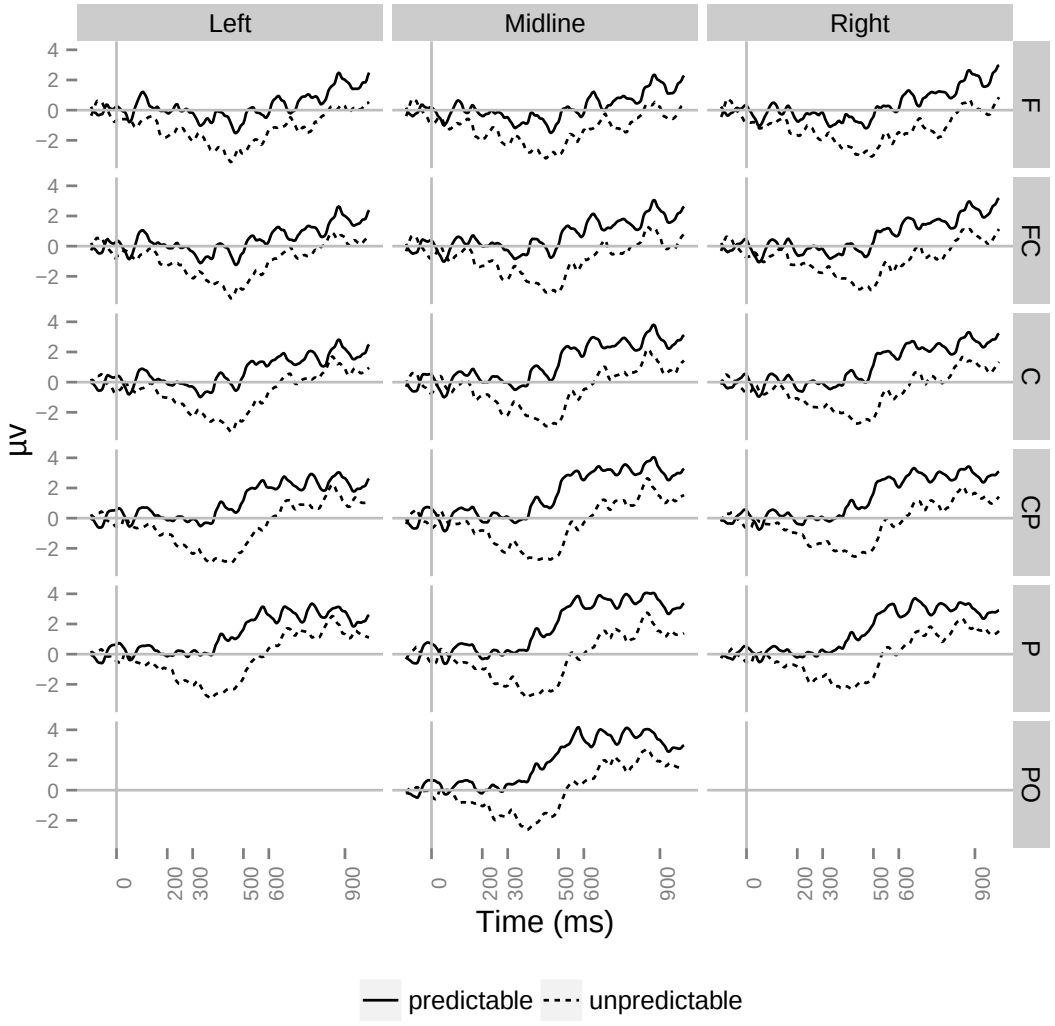

FIGURE 7.2: Grand average ERPs (n=21) for final words in ***fluent*** utterances in the **cough** context. Shown here are ERPs as measured at frontal (F), fronto-central (FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline. Unpredictable words elicit a broadly distributed negativity relative to predictable words. This negativity onsets around 70ms after stimulus-onset, and lasts until the end of the epoch.

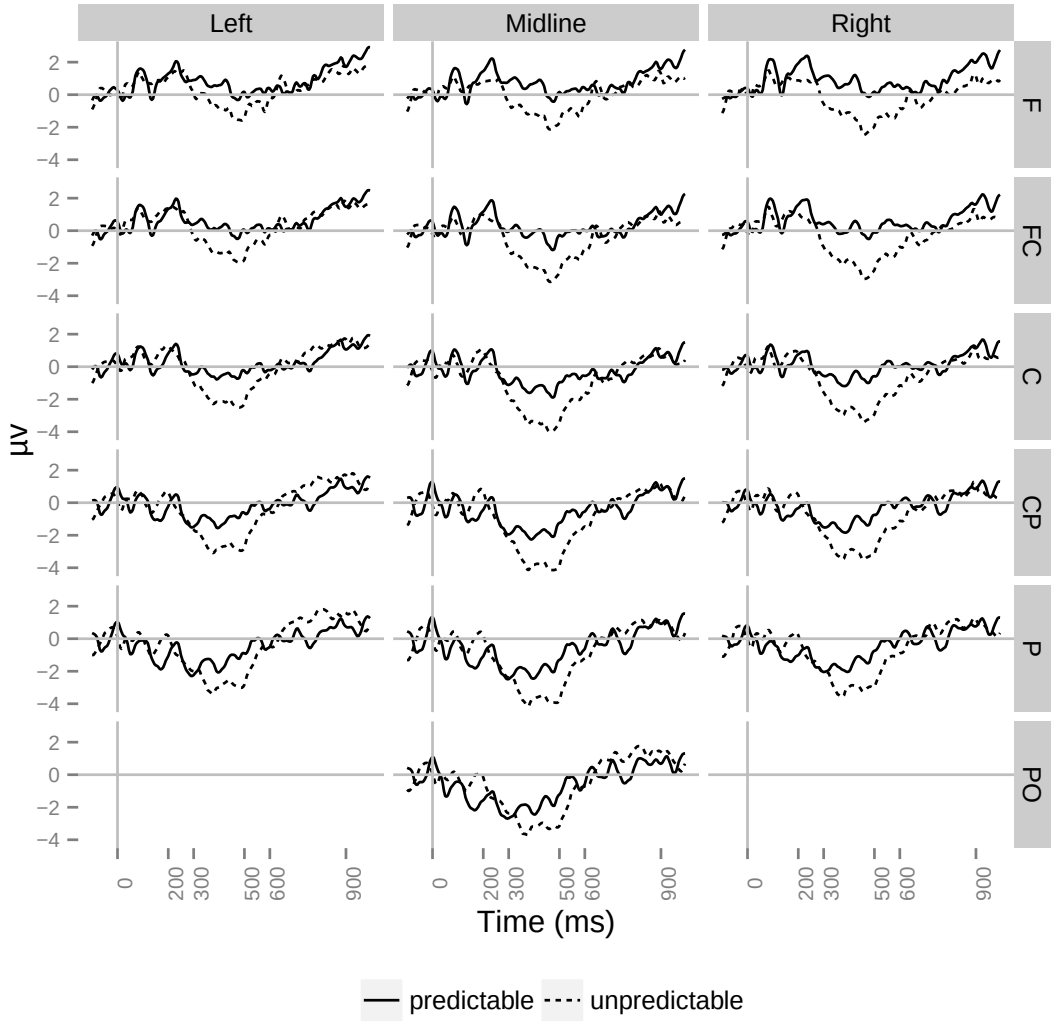

FIGURE 7.3: Grand average ERPs (n=21) for final words in ***disfluent (er)*** utterances in the **beep** context. Shown here are ERPs as measured at frontal (F), fronto-central (FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline. Unpredictable words elicit a broadly distributed negativity relative to predictable words. This negativity onsets around 312ms after stimulus-onset, and lasts until around 630ms.

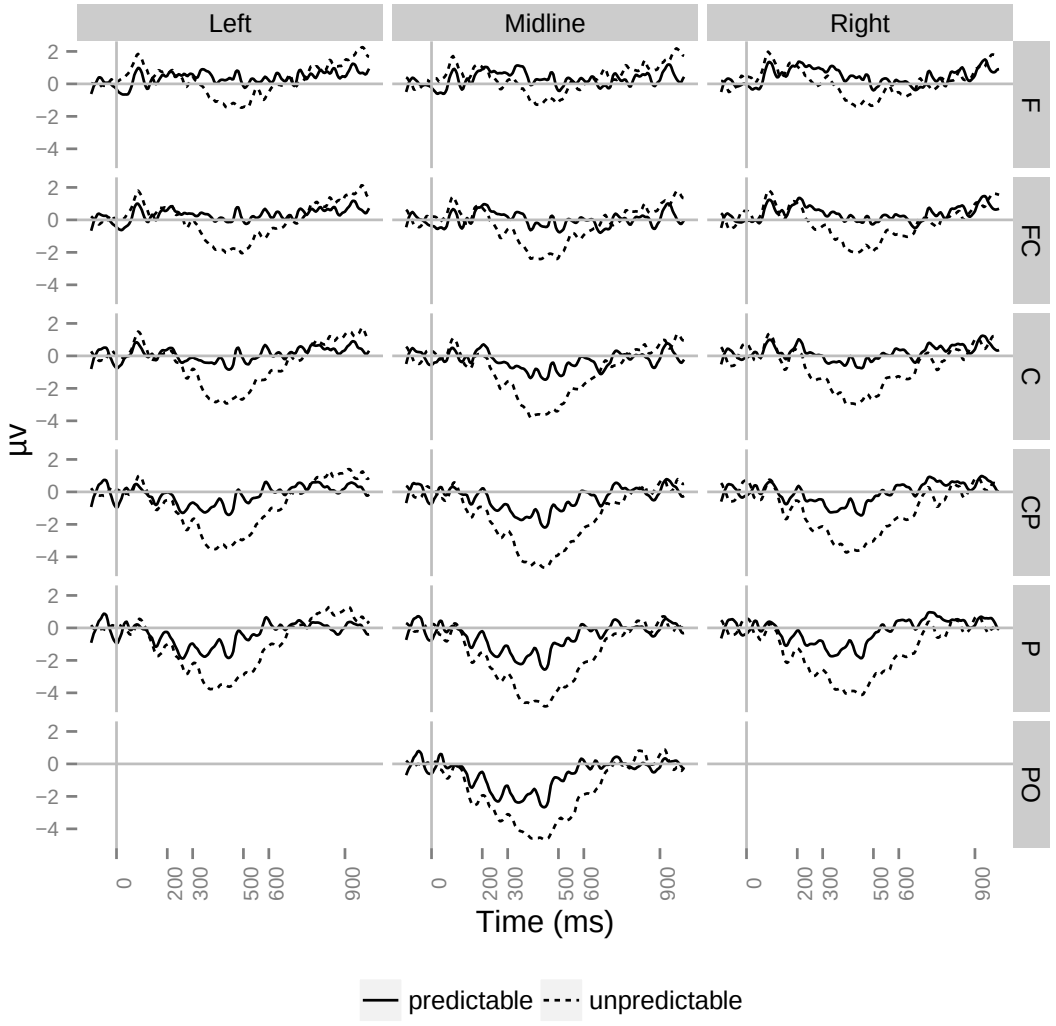

FIGURE 7.4: Grand average ERPs (n=21) for final words in ***disfluent (er)*** utterances in the **cough** context. Shown here are ERPs as measured at frontal (F), fronto-central (FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline. Unpredictable words elicit a broadly distributed negativity relative to predictable words. This negativity onsets around 160ms after stimulus-onset at posterior electrodes, and around 300ms at frontal electrodes. The relative negativity for unpredictable words lasts until around 700ms.

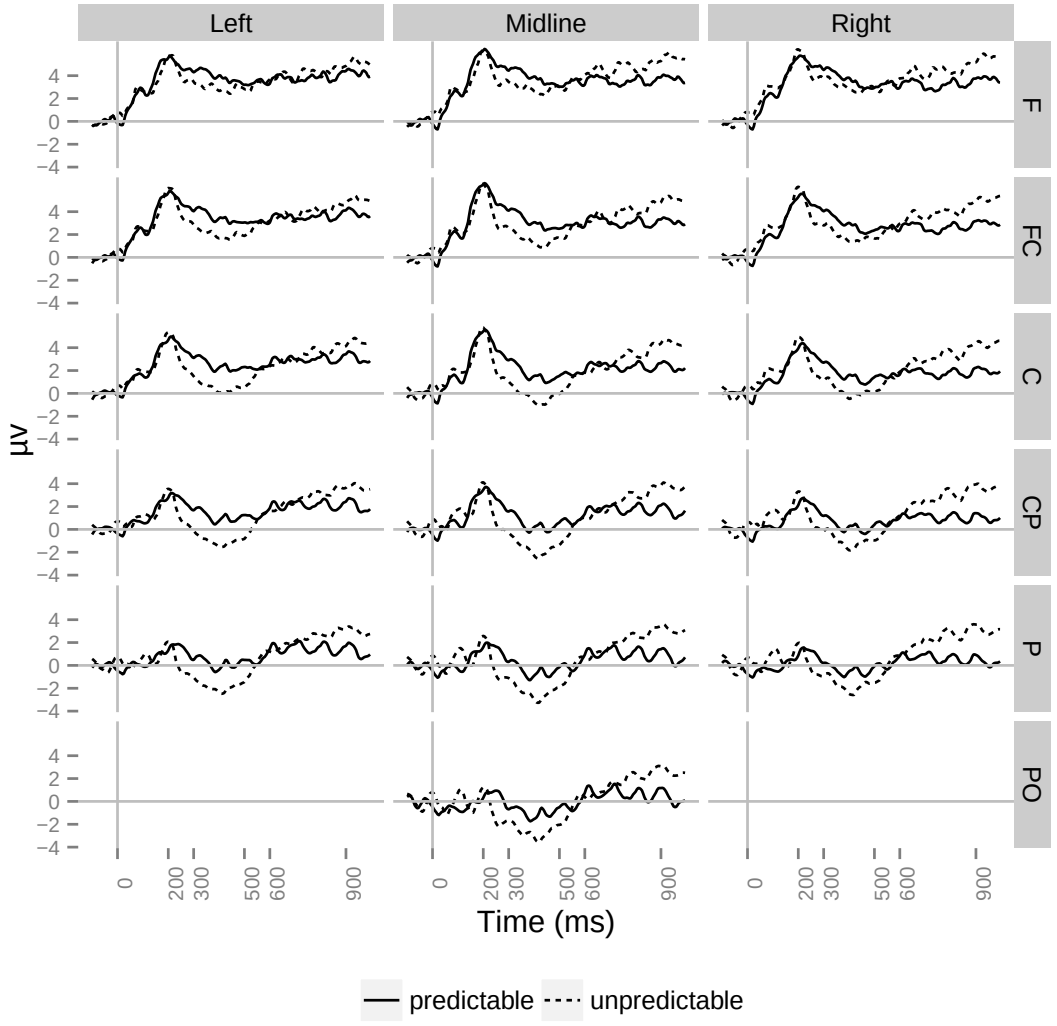

FIGURE 7.5: Grand average ERPs (n=21) for final words in ***interrupted (beep)*** utterances in the **beep** context. Shown here are ERPs as measured at frontal (F), fronto-central (FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline. Unpredictable words elicit a broadly distributed negativity relative to predictable words. This negativity onsets around 200ms after stimulus-onset, and lasts until around 550ms. Following this negativity, the ERPs display a relative positivity for unpredictable words, which onsets around 600ms in the right hemisphere, and around 700ms for mid-line and left hemisphere sites.

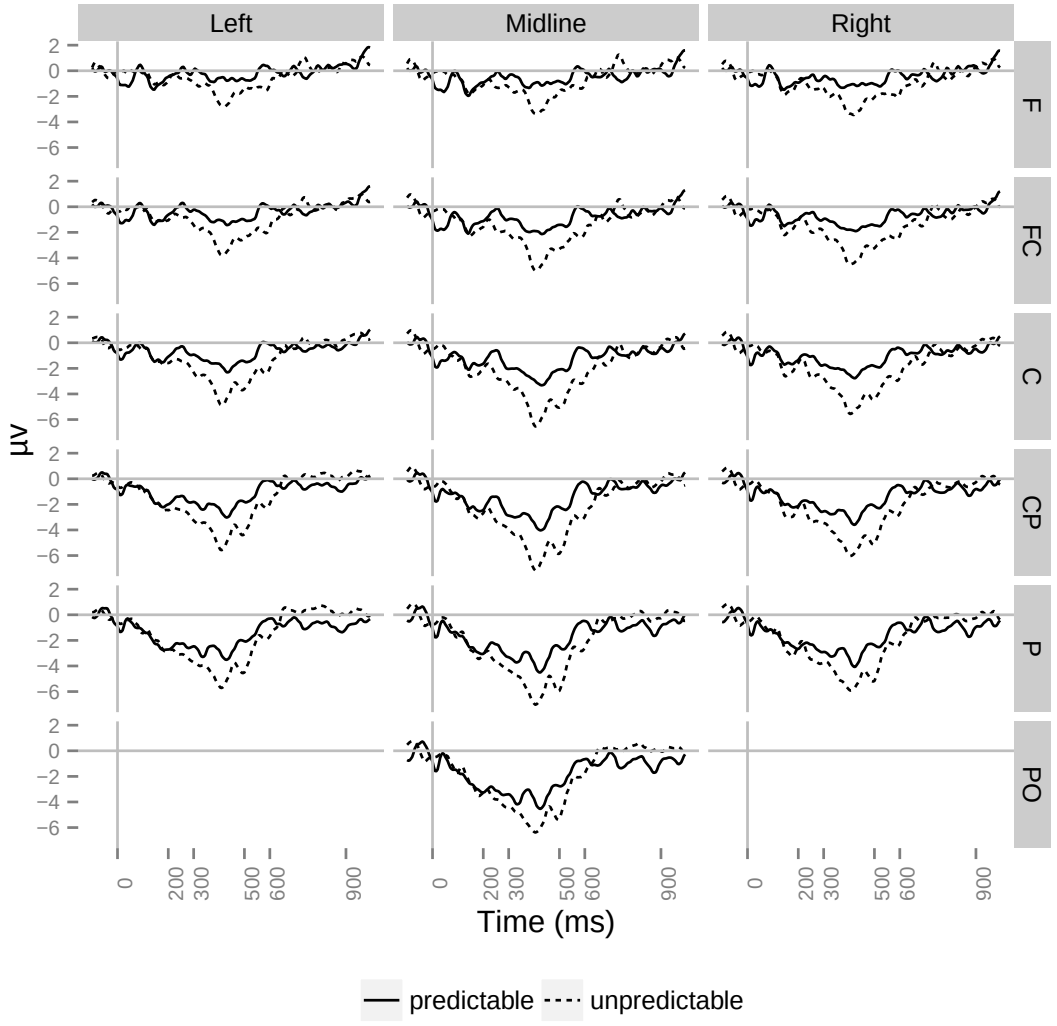

FIGURE 7.6: Grand average ERPs (n=21) for final words in ***interrupted (cough)*** utterances in the **cough** context. Shown here are ERPs as measured at frontal (F), fronto-central (FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline. Unpredictable words elicit a broadly distributed negativity relative to predictable words. This negativity onsets around 250ms after stimulus-onset, and lasts until around 630ms.

7.3.1 300-500ms

In fluent, disfluent and interrupted utterances in the cough-context data, ERPs to unpredictable target words are more negative than ERPs to predictable target words over the 300-500ms time window. This negativity appears maximal towards the midline and

towards centro-parietal electrodes, consistent with an N400 effect. In the beep-context data, ERPs to unpredictable targets in fluent utterances also produce a negativity compared to predictable targets over the 300-500ms time window, although the distribution of the effect is less consistent in this block than in the cough-context block. In the beep-context block, the N400 effect to fluent and interrupted targets is maximal in the left hemisphere towards the centro-parietal electrodes, whereas the effect to disfluent targets appears to be maximal in the right hemisphere, towards the fronto-central electrodes.

Amplitude analysis — 300-500ms

To determine whether the N400 effects observed in the waveforms were reliable, means of the data in the standard N400 time window (300-500ms) for each fluency condition were submitted to multilevel ANOVA with factors of predictability *(predictable, unpredictable)*, location *(F, FC, C, CP, P)*, hemisphere *(left, right)* and site *(mid, superior, medial, inferior)*, and incorporating the electrodes indicated in Figure 4.2.

For fluent stimuli in the beep context, a multilevel ANOVA revealed a marginally significant main effect of predictability [$F(1,20) = 4.07$, $\eta^2_G = 0.07$, $p < 0.1$], reflecting the slightly more negative ERP obtained for unpredictable words. There were no significant interactions involving predictability.

An ANOVA for fluent utterances in the cough context revealed a main effect of predictability [$F(1,20) = 15.35$, $\eta^2_G = 0.02$, $p < 0.001$]. There were also marginally significant interactions of predictability with location [$F(1.24,24.84) = 2.81$, $\eta^2_G = 0.009$, $p < 0.1$] and predictability with site [$F(1.1,21.8) = 3.25$, $\eta^2_G = 0.002$, $p < 0.1$]. This reflects a relative negativity for unpredictable words, which is slightly greater towards the midline and at posterior locations.

For disfluent utterances in the beep context, there was a main effect of predictability F(1,20) = 21.72, η^2_G = 0.122, p <0.0005]. There were also a significant interactions of predictability with location and hemisphere [F(1.6,16.1) = 6.05, η^2_G = 0.001, p <0.0005] and predictability with location, hemisphere and site [F(3.9,78.1) = 2.82, η^2_G = 0.0001, p <0.05]. As can be seen in Figure 7.7, the negativity for unpredictable compared to predictable words is greater towards the front of the head, and this effect of location is greater in the right hemisphere. Additionally, there are differences in voltage across levels of site in all quadrants except the front right.

An ANOVA incorporating data from disfluent utterances in the cough context revealed a main effect of predictability [F(1,20) = 14.08, η^2_G = 0.137, p <0.005], and an interaction of predictability with site [F(1.1,21.4) = 8.81, η^2_G = 0.002, p <0.0001]. There was also a marginally significant interaction of predictability with location and site [F(2.3,45.6) = 2.39, η^2_G = 0.0003, p <0.1]. This reflects an effect which is most negative towards midline sites, particularly over parietal locations.

For utterances interrupted by a beep, ANOVA revealed a main effect of predictability [F(1,20) = 8.30, η^2_G = 0.091, p <0.01]. There was also a marginally significant interaction of predictability with hemisphere [F(1,20) = 4.31, η^2_G = 0.004, p <0.1], as well as significant interactions of predictability with site [F(4.4,22.9) = 5.43, η^2_G = 0.002, p <0.05], and predictability with location, hemisphere and site [F(3.6,71.4) = 3.50, η^2_G = 0.0002, p <0.05]. This reflects the fact that at posterior locations, there is a stronger gradient of predictability across levels of site in the right hemisphere than the left, whereas at frontal locations, the gradient across site does not differ between the hemispheres.

An ANOVA incorporating data from utterances interrupted by a cough revealed a main

effect of predictability [F(1,20) = 10.86, η^2_G = 0.104, p <0.005], and a marginally significant interaction of predictability with site [F(1.1,22.4) = 3.88, η^2_G = 0.001, p <0.1]. This reflects a negativity for unpredictable compared to predictable words, which his greatest towards the midline.

For a summary of the significant effects of the amplitude analyses detailed in this section, turn to Table 7.2. Overall, all six conditions elicited a relative negativity for unpredictable words, and for five of the six conditions, this negativity was maximal over centro-parietal or parietal electrodes. The only exception was disfluent utterances in the beep context, for which the relative negativity for unpredictable words was greater towards the front of the head.

Topographic analysis — 300-500ms

To establish whether the predictability effects revealed above differed in topographic distribution, rescaled difference scores from predictable and unpredictable targets were submitted to a multilevel global ANOVA. The ANOVA had levels of fluency *(fluent, disfluent, interrupted)*, location *(F, FC, C, CP, P)*, hemisphere *(left, right)*, site *(superior, medial, inferior)* and context *(cough, beep)*, and incorporated data from the electrodes used in the global analysis (see Figure 4.2). This analysis revealed a significant interaction of fluency with context [F(1.9,37.2) = 1.09, η^2_G = 0.093, p <0.0005], but importantly, revealed no significant interactions of fluency or context with factors of site, location or hemisphere (all F's >0.05). As most of the observed effects appear stronger towards the midline, this was followed up with an ANOVA examining data from six midline electrodes (F, FC, C, CP, P, PO) incorporating factors of fluency and location. This also revealed no significant main effects of fluency or significant interactions involving fluency. Thus there is no evidence for any distributional differences between the effects elicited

observed in fluent, disfluent and interrupted utterances, whether in cough-contexts or beep-contexts. On the basis of this analysis there is no reason to assume that different neural generators underlie these effects, and so these effects are quantitatively compared below.

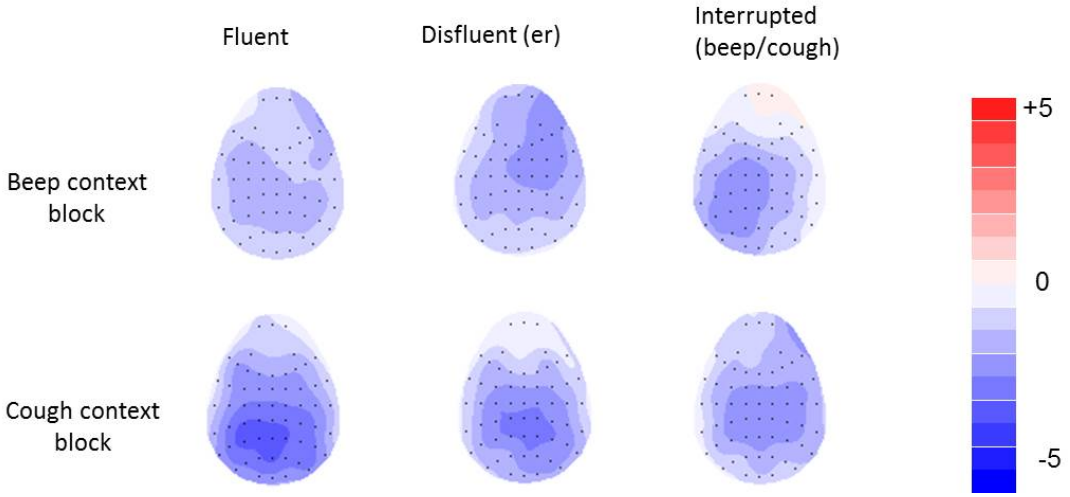

FIGURE 7.7: Scalp topographies (n=21) showing the predictability effects in the 300-500ms time window for targets in utterances which were fluent (left), disfluent (middle) and interrupted by a beep or a cough. (right). Scalp topographies from the beeps context are shown on the first line, and topographies from the coughs context are shown on the second. All fluency conditions elicit a relative negativity for unpredictable words, which is broadly distributed over the scalp. For disfluent *(er)* targets, this relative negativity has a mid-frontal focus; for the remaining conditions, the negativity is larger at centro-parietal midline electrode sites.

Quantitative analysis — 300-500ms

Having established that all six fluency and context conditions produced ERPs consistent with the presence of an N400 effect, and that these effects did not vary topographically, a factor of fluency *(fluent, er, interruption)* was added to the global and ANOVA to allow a quantitative comparison of effects between fluency conditions. This AVOVA revealed a significant interaction of predictability with location [$F(1.2,23.3) = 4.35$, $\eta^2_G = 0.003$, $p < 0.05$], but importantly, no significant interactions between predictability and fluency, or predictability and context. Thus there was no evidence that the magnitude of the N400 effect varied across fluency or context conditions. Mid-line analysis is not reported

as a factor of hemisphere was implicated for some conditions in the previously reported amplitude analysis.

Visual inspection of the data suggested that a combination of Cz, CPz and Pz represented a reasonable approximation of the effect maxima across the six fluency and context conditions, and so the absolute magnitude of the N400 effect averaged across these three electrodes is presented as a graph for visual inspection below (Figure 7.8). This adds to the impression given by comparison of the ERP waveforms (Figures 7.1 to 7.6), scalp topographies (Figure 7.7) and quantitative ANOVA of no significant differences in N400 effect magnitude across conditions.

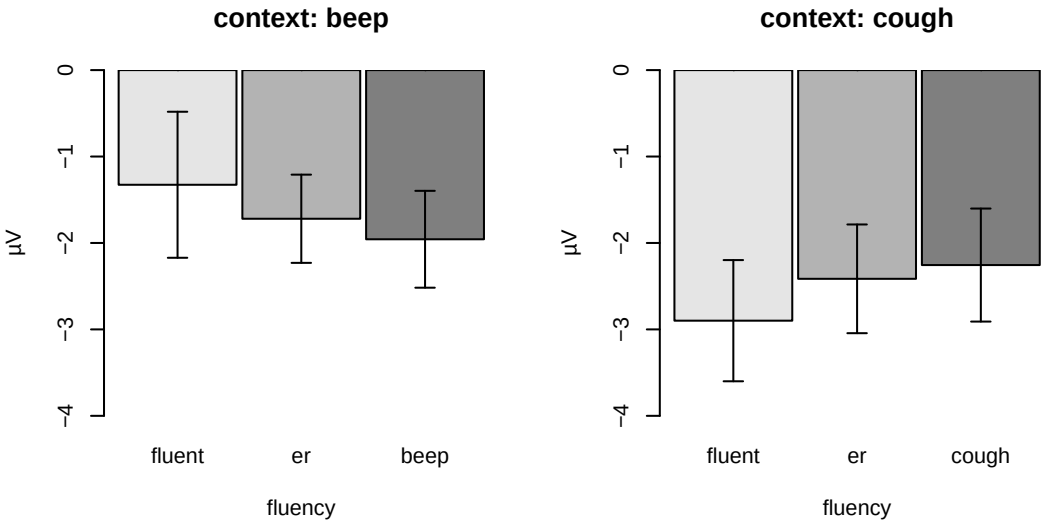

FIGURE 7.8: Mean voltage difference between unpredictable and predictable targets in the 300-500ms time window (n=21). Values are averaged across the Cz, CPz and Pz electrodes, which were considered to represent the maxima of the effect across conditions. Error bars represent one standard error of the mean. Unpredictable target words elicited more negative ERPs than predictable words in this epoch, but the size of this predictability effect did not vary significantly across fluency and context conditions.

7.3.2 600-900ms

Visual inspection of the ERPs suggested that the N400 was followed by a relative positivity for unpredictable words. This was also observed in the previous two experiments,

and so it is here that analysis now turns. In most of the fluency and context conditions, this positivity begins soon after the offset of the N400, around 600ms, and spreads from posterior locations in the left hemiphere to become broad ranging, focussing particularly over mid-frontal electrodes. The exceptions to this are interrupted words in the beep context, for which the relative positivity has a right- rather than left-hemisphere bias; and fluent words in the cough context, for which the period after 500ms is characterised by a centro-parietal negativity to unpredictable words.

Amplitude Analysis — 600-900ms

In order to establish whether there were meaningful predictability differences in the 600-900ms time window, subject means of the data in each condition were submitted to multilevel ANOVA, which incorporated data from the global analysis electrodes shown in Figure 4.2, and which had levels of predictability *(predictable, unpredictable)*, location *(F, FC, C, CP, P)*, hemisphere *(left, right)* and site *(superior, medial, inferior)*.

For fluent words in the beep context, ANOVA revealed an interaction of predictability with location [$F(1.2,24.8) = 8.38$, $p < 0.01$], and a marginally significant interaction of predictability with location and site [$F(2.0,39.4) = 2.54$, $\eta^2_G = 0.012$, $p < 0.1$]. This reflects an ERP effect which is more positive for unpredictable words at posterior locations, and in which a gradient is present over levels of site at frontal but not posterior electrode sites.

For fluent words in a cough context, ANOVA revealed a significant effect of predictability [$F(1,20) = 7.07$, $\eta^2_G = 0.086$, $p < 0.05$], but unlike in the beep context, this reflects an ERP effect which is more negative for unpredictable words. This negativity was more

prevalent over the right hemisphere, as evidenced by the marginally significant interaction of predictability with hemisphere [$F(1,20) = 4.07$, $\eta^2_G = 0.002$, $p < 0.1$]. The ANOVA also revealed significant interactions of predictability with site [$F(1.1,22.3) = 5.02$, $\eta^2_G = 0.002$, $p < 0.05$], predictability with location and site [$F(2.4,44.8) = 8.20$, $\eta^2_G = 0.002$, $p < 0.001$], and a marginally significant interaction of predictability with hemisphere and site [$F(1.2,24.3) = 3.02$, $\eta^2_G = 0.0003$, $p < 0.1$]. This confirms the impression given by Figure 7.9 (bottom left) of a relative negativity with is greater towards the midline, particularly at posterior sites, and which is more evenly distributed across levels of site in the right hemisphere. Importantly, there is no evidence to suggest that this is in any way similar to the mid-frontal and left-parietal positivity previously interpreted as a LPC.

Analysis of disfluent words from the beep context revealed no main effect of predictability, but did show an interaction of predictability with hemisphere [$F(1,20) = 4.85$, $\eta^2_G = 0.003$, $p < 0.05$], reflecting an effect with was more positive for unpredictable words in the left hemisphere. This effect was more evenly spread across sites at posterior than frontal locations, as evidenced by a significant interaction of predictability with location and site [$F(3.1,61.6) = 3.56$, $\eta^2_G = 0.0004$ $p < 0.05$]. The effect was also more evenly spread across sites in the left hemisphere, as evidenced by a marginally significant interaction of predictability with hemisphere and site [$F(1.1,22.7) = 3.85$, $\eta^2_G = 0.0004$, $p < 0.1$].

In the cough context, the ANOVA for the ERPs for disfluent words also produced no main effect of predictability, but did reveal a significant interaction of predictability with hemisphere [$F(1,20) = 4.64$, $\eta^2_G = 0.003$, $p < 0.05$]. Predictability and hemisphere interacted marginally significantly with site [$F(1.1,22.7) = 3.63$, $\eta^2_G = 0.0004$, $p < 0.1$], reflecting an effect which differed more across sites in the left hemisphere. Additionally, the ANOVA revealed a significant interaction of predictability with location and site

[F(1.9,37.7) = 6.74, η^2_G = 0.0008, p <0.005], reflecting the fact that the predictability effect elicits an ERP effect which is variable across sites at posterior electrodes.

For words which had been interrupted by a beep, ANOVA confirmed the impression given by Figure 7.9 of an ERP effect which is more positive for unpredictable words [F(1,20) = 4.42, η^2_G = 0.037, p <0.05], and that this relative positivity has a right hemisphere bias [F(1,20) = 8.33, η^2_G = 0.008, p <0.01]. There was a significant interaction of predictability with location and site [F(1.6,31.7) = 3.60, η^2_G = 0.0008, p <0.05], reflecting an effect which was greater toward inferior sites at posterior locations, as well as a marginally significant interaction of predictability with hemisphere and site [F(1.2,23.0) = 3.95, η^2_G = 0.0006, p <0.1], reflecting the fact that the ERP effect differed across site in the right hemisphere only.

Finally, ANOVA was performed on ERPs to words which had been interrupted by a cough. This revealed a marginally significant interaction of predictability with location and site [F(1.7,34.4) = 3.05, η^2_G = 0.0008, p <0.1], reflecting an effect with was more positive towards inferior sites at posterior locations only. No other main effect or interactions involving predictability reached significance.

Given that hemisphere was widely implicated in these amplitude analyses, midline analysis was not performed. For a summary of the significant results detailed in this section, see Table 7.3. For five of the six fluency and context conditions, analysis revealed significant positivity for unpredictable words. For fluent words in the cough context, however, the 600-900ms time window revealed a relative negativity for unpredictable words.

Topographic Analysis —600-900ms

Topographic analysis made use of the rescaled difference waveforms (predictable - unpredictable) for each fluency and context condition. ANOVA was performed with factors of fluency *(fluent, disfluent (er), interruption)*, context *(beep, cough)*, location *(F, FC, C, CP, P)*, hemisphere *(left, right)* and site *(superior, medial, inferior)*. This revealed significant interactions of context with hemisphere [$F(1,20) = 7.17$, $\eta^2_G = 0.002$, $p < 0.05$], and fluency with hemisphere [$F(3.4,34.2) = 4.38$, $\eta^2_G = 0.006$, $p < 0.05$], as well as a marginally significant interaction of fluency with context and hemisphere [$F(1.9,38.6) = 2.54$, $\eta^2_G = 0.003$, $p < 0.1$].Given these distributional differences between conditions, confirming the impression given by Figure 7.9 of a predictability effect which varies in topography across context and fluency conditions, no quantitative comparison is reported.

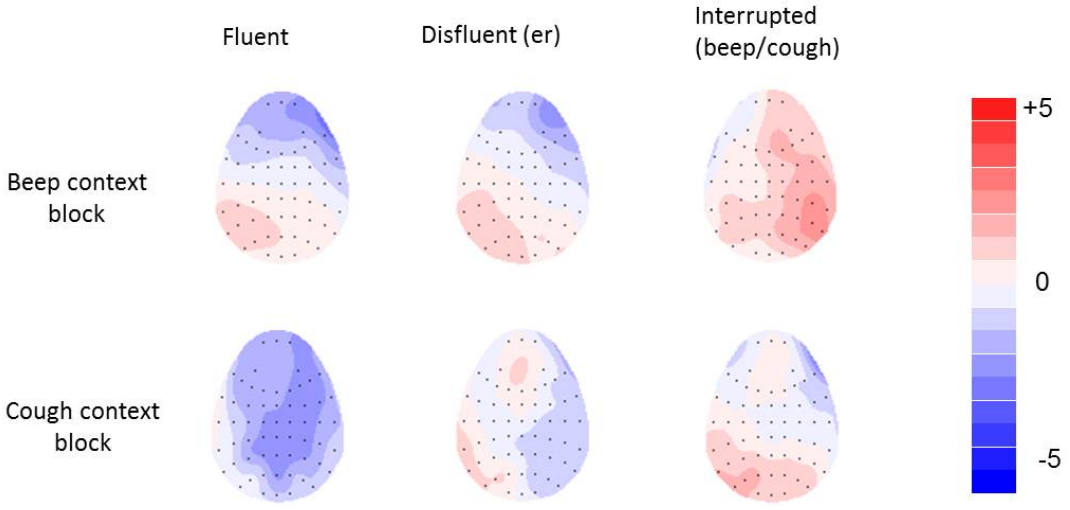

FIGURE 7.9: Scalp topographies (n=21) showing predictability effects in the 600-900ms time window for targets in utterances which were fluent (left), disfluent (middle) and interrupted (right), in a beep context (top) and a cough context (bottom).

7.4 Effects Over Time

Although in most of the predictability and fluency conditions, unpredictable words elicited a relative positivity in the later (600-900ms) time window, fluent words which

had been encountered in the cough condition elicited an ERP which was more negative to unpredictable than predictable words. To establish whether this should be considered to be a continuation of the N400, the rescaled difference waveforms from each fluency and predictability condition were topographically compared across time windows. The comparison made use of the same ANOVA structure as was used for topographic comparisons within time windows (see Sections 7.3.1 and 7.3.2), with an added factor of epoch *(first, second)*, and incorporated data from the electrodes used for global analysis. Every condition was found to reveal differences in scalp distribution between the two windows, including the cough-context fluent condition, for which the ANOVA revealed significant interactions of epoch with hemisphere [$F(1,20) = 5.31$, $\eta^2_G = 0.003$, $p < 0.05$], epoch with location [$F(1.2,23.0) = 6.57$, $\eta^2_G = 0.007$, $p < 0.05$] and epoch with location and site [$F(2.7,54.4) = 12.72$, $\eta^2_G = 0.0007$, $p < 0.00001$]. This reflects the fact that the negativity to unpredictable words is more frontal and right hemisphere biased in the later epoch than in the N400 epoch. As such, this effect cannot be quantitatively compared to the preceding N400, although inspection of the waveforms in Figure 7.2 and comparison of the scalp topographies in Figures 7.7 (bottom left) and 7.9 (bottom left) suggests that it may be reasonable to consider it to be an extension of the N400 in the case of fluent targets in the cough-context block.

300-500ms						
Fluency	fluent		disfluent (er)		interruption	
Context	beep	cough	beep	cough	beep	cough
Predictability	(*)	*	*	*	*	*
Predictability : Location		(*)				
Predictability : Hemisphere					(*)	
Predictability : Site		(*)		*	*	(*)
Predictability : Location : Hemisphere			*			
Predictability : Location : Site				(*)		
Predictability : Hemisphere : Site						
Predictability : Location : Hemisphere : Site			*		*	

TABLE 7.2: Summary of significant [*] ($p < 0.05$) and marginally significant [(*)] ($p < 0.1$) results from global ANOVA amplitude analyses of ERPs to targets in fluent, disfluent and interrupted utterances in the beep and cough contexts, 300-500ms. Subsequent quantitative comparison of these N400 effects revealed no magnitude differences across fluency and context conditions.

600-900ms						
Fluency	fluent		disfluent (er)		interruption	
Context	beep	cough	beep	cough	beep	cough
Predictability		*			*	
Predictability : Location	*					
Predictability : Hemisphere		(*)	*	*	*	
Predictability : Site		*				
Predictability : Location : Hemisphere						
Predictability : Location : Site	(*)	*	*	*	*	(*)
Predictability : Hemisphere : Site		(*)	(*)	(*)	(*)	
Predictability : Location : Hemisphere : Site						

TABLE 7.3: Summary of significant [*] ($p < 0.05$) and marginally significant [(*)] ($p < 0.1$) results from global ANOVA amplitude analyses of ERPs to targets in fluent, disfluent and interrupted utterances in the beep and cough contexts, 600-900ms.

7.5 Summary and Discussion

This experiment combined the manipulations used in the first two experiments (reported in Chapters 5 and 6) with the aim of investigating whether the contrasting results in Experiments 1 and 2 should be considered to be an artefact of different participant cohorts, or a consequence of the differing speech environments in which fluent, disfluent and interrupted targets were encountered.

As expected, unpredictable words elicited negative ERPs relative to predictable words, over the standard N400 time window, 300-500ms. The timing and topography of this negativity are consistent with its identification as an N400 effect. Importantly, there

was no evidence to suggest that the magnitude of the N400 effect was affected by the fluency or context of utterances, although the N400 did appear to be longer lasting for fluent words in the cough context.

Inspection of the data in the later 600-900ms time window, in which a possible LPC has been identified in previous experiments, revealed a slight left parietal positivity for unpredictable words in some conditions. In one condition this relative positivity has a more right hemisphere bias, and in one condition it fails to appear all together. Although the LPC is typically described as having a mid-frontal distribution (Van Petten et al., 1991; Federmeier et al., 2007), careful inspection of the results presented in MacGregor's thesis (MacGregor, 2008) show a mid-frontal distribution with a left parietal component, and so it remains possible that the effects seen in these data may be interpreted as such. However, given the weakness of the effects, and the variations in distribution between fluency and context conditions, no quantitative comparison between fluency and context conditions was attempted.

The reason for the apparent lack of disfluency effects on the N400 deserves some thought. Inspection of the graph showing mean N400 effect amplitudes for each condition at midline centro-posterior electrode sites (Figure 7.8) confirms the impression given by the amplitude analyses of data containing a lot of variance. This is visually confirmed by the large error bars on the graph, demonstrating that one standard error of the mean is comparable to the mean size of the N400 effect, particularly for fluent targets in the beep condition.

With particular regard to ERPs to fluent utterances in the beeps context, which appeared to fail to produce a strong N400, appropriate checks for errors in the timing and coding of the experiment were made to establish whether this weak predictability effect was an

artefact of experimental error. It was concluded that this was not the case. Although the number of trials per participant does not appear to be significantly lower for this condition than the remaining eleven conditions (see Table 7.1), inspection of the ERP waveform, shown in Figure 7.1 shows that the data are very noisy. This is particularly the case when considered in comparison to the waveforms from the same condition in the Beeps experiment (Chapter 5), shown in Figure 5.1. Thus it seems likely that the relative weakness of the N400 in this condition should be considered to be due to large variance between participants. Despite the lack of N400 effect in this condition, however, the later time window does appear to show an LPC, theorised to reflect memory control processes, and which has been observed in similar experiments following an N400 effect (MacGregor, 2008).

Given the lack of effect of fluency on N400 effects, post-hoc analysis of effect sizes and statistical power was carried out to determine whether this experiment would have had the power to detect such effects, were they present[1]. For the quantitative comparison global ANOVA, a table of effect sizes and statistical powers of main effects and interactions of interest is given below (Table 7.4). As is obvious from the table, the effect sizes of all of the effects and interactions of interest is small, and consequently the power to detect them within this experiment (given the number of participants incorporated in the analysis) is also very small; the power to detect a main effect of predictability is 0.39, and for each of the interactions of interest, including the interaction between predictability and fluency, the power no greater than is 0.06. These fall well short of the recommended statistical power of 0.8 (Cohen, 1988, 1992). Thus it is possible that the failure to detect differences in N400 amplitude between fluency conditions is a function of low statistical power, given the very small size of any effects.

[1]Statistical power analysis was carried out using the freely available G*Power software (Faul, Erdfelder, Lang, & Buchner, 2007).

Main Effect or Interaction	Generalised Eta Squared	Statistical Power
Predictability	0.13	0.39
Predictability : Fluency	0.00037	0.05
Predictability : Context	0.0040	0.06
Predictability : Fluency : Context	0.0016	0.05

TABLE 7.4: Table showing the effect sizes (generalised eta squared, as calculated by ezANOVA, R) and power for the main effect of predictability, and key interactions of interest to the experimental manipulation.

An alternative way to interpret the effect sizes and statistical power presented in Table 7.4 is to consider whether the small effect of fluency on predictability indicates that disfluency has only a minimal impact on comprehension, or whether EEG is too insensitive or noisy to detect these effects. These possibilities will be discussed further in Chapter 10.

As the patterns of data across the two epochs vary slightly between contexts, each context will now be considered separately. First the pattern of results from targets in the beep context will be considered, followed by the pattern of results from targets in the cough context, before moving on to discuss how these findings, if reliable, affect our understanding of disfluency.

Results from Targets in the Beep Context

For targets in the beep context, there was no effect of fluency on N400 effect size. Importantly, there was no reduction in N400 size for disfluent words compared to fluent words. The N400 effect is followed by a relative positivity for unpredictable words. This positivity has a left parietal maxima for fluent and disfluent words, but is greater over the right hemisphere for words interrupted by a beep. Previous experiments have found an LPC following disfluent but not fluent words (MacGregor, 2008). It has been theorised that the LPC reflects memory control processes and may be associated with

the resumption of sentence comprehension following a disfluency. If indeed the left parietal positivity seen in a later time window should be interpreted as a LPC, then its appearance for the fluent condition is rather surprising. That said, the N400 to these fluent items was remarkably unreliable, as can be seen in Section 7.3.1. The lack of robust N400 effect, combined with a subsequent relative positivity for unpredictable words would appear to make it valid to question whether these stimuli were really being interpreted as fluent by participants, or whether these stimuli were subject to some of the same effects that have previously been reported following disfluent fillers (in Chapter 5 of this thesis and by Corley et al. (2007). If these results are to be believed, then one would have to conclude that cohort may play a role in determining responses to disfluency, particularly given the difference between the results of the beeps context block in this study, and the study reported in Chapter 5.

Results from Targets in the Cough Context

For targets in the cough context, there was no significant effect of fluency on N400 size. Although visual inspection of the data appears to reveal a slight reduction in N400 size for disfluent words, this fails to reach statistical significance. In contrast to the beep context, fluent utterances elicited a robust N400 effect as would traditionally be expected for highly unpredictable words (Kutas & Hillyard, 1980). Following the N400 a relative positivity for unpredictable words emerged in the disfluent and interrupted conditions. If this positivity is to be interpreted as an LPC, then its presence only in the disfluent and interrupted condition is consistent with the interpretation that it reflects memory control processes involved in resuming fluent comprehension, following the disruption. This is, however, in contrast to the findings reported in Chapter 6, in which a relative positivity for unpredictable words appeared in all three fluency conditions. This would again lead

to the conclusion that, to some extent, cohort determines the results. However, even within the single cohort used in the current experiment, there is a difference in the pattern of the LPC affects across contexts. This brings us to conclude that cohort and context interact.

7.5.1 Discussion

If the outcomes of this experiment are to be believed, then the conclusion should be that predictability effects as indexed by the N400 do not vary with fluency. We might additionally suggest that cohort is a determining factor, and that the fluency effect observed in the Experiment 1 (Chapter 5) was either an unreliable effect, or was particular to the cohort of participants accepted for that study.

A review of ERP disfluency experiments reveals that a reduced N400 following disfluency is not necessarily prevalent. MacGregor (2008) compared the size of the N400 effect between fluent and disfluent words in a series of experiments using disfluent fillers *(er)*, silent pauses and repetitions. Of these, only disfluent fillers produced a smaller N400 than fluent utterances. MacGregor concluded that this indicated that it was the phonetic form of the filler which was key, but it is also possible that these results should be interpreted as showing that the N400 is not necessarily sensitive to disfluency. It is worth bearing in mind that the initial study which reported a disfluency effect on the N400 (Corley et al., 2007), also reported in MacGregor, (2008), and using fillers *(er)* included only twelve participants. Such a low participant count may increase the likelihood of individual idiosyncracies having a significant effect on the outcomes.

There are, however, a number of factors to consider before drawing definite conclusions,

particularly with regard to the experiment reported in this chapter. The current experiment suffered from relatively low trial numbers (see Table 7.1). Low trial numbers may be particularly problematic in auditory speech experiments, in which the speech signal unfolds over time. These slight offsets between the identification times for each individual stimulus may reduce the power of averaging to improve the signal to noise ratio. Certainly, visual inspection of the waveforms reveals the data to be noisy, which raises the question of whether this experiment would have had the power to detect a relatively small effect such as the reductions in the N400 reported by Corley et al. (2007) (approximately $2\mu V$ difference between fluency conditionss at CPz maxima) or in Chapter 5 of this thesis (approximately $1.5\mu V$ difference between fluent and disfluent conditions at CPz maxima). Additionally the failure of fluent utterances in the beep context to produce a strong reliable N400 and the large standard error in this condition demonstrate that the data may be considered unreliable due to the low number of trials included.

One point which cannot be overlooked, given the aim of the experiment of clarifying whether cohort or context may account for the differences between ERPs results reported across experiments, is that within this experiment contexts are not fully separated. Participants heard utterances in one context, and then the other. Thus the processing of utterances heard in the second context encountered may have been influenced by the first. However, when participants are separated by context order no noteworthy differences emerge, although it is important to acknowledge that when only half of the participants are considered the actual number of trials incorporated into grand means is very low, resulting in extremely noisy data. It is also possible that in the current experiment there are simply too few trials for each stimulus item. It was previously mentioned that across all participants, each target word occurred only twice in each condition before processing,

and the associated epoch- and subject rejection. Like participants, words have their own innate properties, which are usually averaged over in linguistic experiments. However, taking account of these properties may give a better understanding of the data (Baayen, Davidson, & Bates, 2008). This will be further explored in Chapter 9.

Having considered the limitations of the current experiments, discussion will now turn to the possibility that the N400 is not sensitive to disfluency.

Is the N400 Sensitive to Disfluency?

A disfluency related attenuation of the N400 effect was first reported by Corley et al. (2007). The rationale behind expecting an effect on the N400 is clear in light of eye-tracking evidence that participants are more likely to fixate on discourse new or abstract items following a disfluency (Arnold et al., 2007, 2004). Similarly, mouse tracking evidence showing listeners to be faster to click on abstract shapes following disfluency than in fluent utterances, and that they began moving the mouse cursor towards the abstract shape even during the disfluent filler (Barr, 2001). Items which are of low frequency, unfamiliar or discourse new are likely to produce bigger N400s than familiar or expected items (Van Petten & Kutas, 1990; Van Petten, 1993), and so it follows logically that where listeners appear to be expecting these items, as in the eye-tracking and mouse-tracking experiments, that one might expect to see a reduction in the N400.

However, whilst the initial rationale appears sound, there is nonetheless good reason to challenge this assumption. One point to bear in mind with regard to the studies mentioned above is that they all used closed sets of possible referents and had task demands other than simply listening and understanding. In each case, participants knew that they would be required to make a response to one of a small number of referents,

visible on the screen, and so it seems likely that they would be preparing their response, and actively predicting which of the available items would be mentioned. By contrast, in natural conversation, and to some extent in the utterances used in the present study, listeners ongoing predictions are not bounded by a set of four or five visual references. As such it may be that the predictive effects reported by Barr (2001) and Arnold et al. (2004, 2007) may not necessarily generalise beyond the narrow context of the testing environment. Alternatively, a closed referent set may simply strengthen any predictions about upcoming material, based on fluency. Even in natural discourse, where people are not bounded by closed referent sets, people maintain models of the discourse record, and so it is possible that following disfluency, listeners suppress expectations for discourse-old referents (Arnold, Fagnano, & Tanenhaus, 2003). If this is the case, then it may be that the very small effects of fluency on prediction can only be reliably detected when strengthened by the addition of a closed referent set.

However, this is not to dismiss the immediate effects of disfluency on comprehension. Increased attention to the speech stream has been demonstrated by Collard et al. (2008), who showed that acoustically deviant targets elicited MMN and P300 ERP components, associated with the capture and orientation of attention. These effects were attenuated following a disfluency, and Collard argues that the disfluency had already oriented attention to the speech stream, and so attention could not be further raised by the acoustically deviant stimulus.

How does attention affect the N400?

Selective attention paradigms have demonstrated that although N400 effects may be elicited by unattended targets in the context of unattended primes, the N400 and is larger when both primes and targets are attended (Otten et al., 1993). As such, one

might extrapolate that as the attention focused on the target is increased, the magnitude of an N400 elicited may also increase. If then, disfluency increases attention to the speech stream, as proposed by Collard et al. (2008), we might then assume that N400 effects would actually be larger following disfluency than in fluent utterances.

Here it is important to also question the hypothesised relationship between attention and disfluency in these experiments. If N400 magnitude is larger when items are attended, it is reasonable to ask whether the attenuated N400 reported by Corley et al. (2007), and the attenuated P300 reported by Collard et al. (2008), are actually a product of reduced attention, perhaps as a function of an attentional blink following the disfluent filler. Collard et al. (2008) reject this on three counts: Firstly, the N400 is not attenuated in an attentional blink. Secondly attention attenuation is greatest approximately 300 ms after the onset of the first stimulus, in this case the attention orienting filler. The speech experiments with which we are concerned have a much longer lag between filler onset, and word onset. Thirdly, and most importantly, a memory benefit has been demonstrated for words affected by fillers (Collard et al., 2008; Corley et al., 2007); this is the opposite of what we expected if fillers were causing an attentional blink. Even if there is doubt over the reduction of the N400 effect following disfluency, the timing of an attentional blink and the memory benefits conferred by disfluency are enough to make it plausible that Collard's findings do in fact point to raised attention, rather than suppressed attention to the speech, resulting from the disfluent filler. As such, a dip in attention to the speech stream following a disfluent filler cannot account for the attenuation in N400 reported in some experiments.

Another possibility is that there are, in fact, two processes occurring simultaneously, which both affect N400 magnitude. It may be that N400 size is reduced either by participants expecting the unexpected (Arnold et al., 2004, 2007; Barr, 2001), or through

the increased difficulty in integrating predictable words following disfluency, at the same time as increase in attention levels increases the size of the N400 effect. If there are two competing processes occurring, then it is possible that they interact causing variable results depending on overhead conditions and cohort. If the N400 is not directly sensitive to disfluency, but attention is, then a robust memory effect might be expected, even in the absence of disfluency effects on the N400. It is to the memory outcomes of the experiments reported in the last three chapters that discussion will now turn.

Chapter 8

Memory Performance and ERP Results

8.1 Introduction

The previous chapters have reported three ERP experiments comparing the immediate effect of fillers, coughs and beeps on online language processing. These experiments also incorporated a surprise memory test component, the outcomes of which are described and discussed in this chapter.

8.2 Disfluency and memory

A relationship between disfluency and memory was first demonstrated by Ozuru and Hirst (2006), whilst investigating how surface features of speech, such as pauses and intonation, affect listeners' credibility judgements. In their experiments, Ozuru and Hirst asked participants to listen to pre-recorded question and answer sessions between two

people. They found that although listeners had difficulty remembering pause lengths, and using pauses to inform subsequent credibility judgements, answers which had been preceded by a long pause (5 seconds) were better remembered than answers which had been preceded by a short pause (1 second). The authors posit that this indicates that during the pause, listeners were thinking about the content of the utterances. Given that the pauses occurred before the answers were uttered, it might be possible to construe "thinking about the content of the utterances" to mean focusing attention on the upcoming utterance.

Corley et al. (2007) also linked disfluency and subsequent memory. Using the paradigm upon which the experiments in this thesis are based, Corley et al. (2007) found that disfluency had a long term effect on the representation of words in memory, such that words which had been preceded by a disfluent filler were more likely to be recognised in a subsequent memory test, up to 55 minutes after participants had initially heard them. This effect was primarily driven by an increase in the memorability of predictable words. Similar results are reported in MacGregor et al. (2010), in which silent pauses were used in place of fillers.

In a slight variation of the same paradigm, MacGregor et al. (2009) considered the effect of repetition disfluencies on immediate language processing and subsequent memory. In this experiment, sentences were rendered fluent or disfluent not by the addition of a filler or a silence, but by the repetition of one or two words before the sentence-final target. An example of a disfluent sentence from this experiment would be:

> Everyone's got bad habits and mine is biting my my nails.

Disfluent utterances were created from fluent recordings by copying pre-target words and splicing them into the speech stream to create a repetition. As in the experiments

reported in this thesis, stimulus sentences were created in pairs, so that the predictable ending of one sentence formed the unpredictable ending of another. Interestingly, in this experiment, MacGregor et al. (2009) did not find an effect of fluency on memory. Nor did they find an effect of fluency on the size of the N400 during the listening task, although there were fluency dependent differences in the ERPs later in the epoch. This lack of N400 effect, co-occurring with a lack of memory effect, leads to the suggestion (also put forward by Corley et al., 2007) that any disfluency advantage in recognition memory for items which are presented disfluently may reflect differences in processing indexed by the N400. Corley et al. (2007) suggest that this tells us something about how disfluency affects immediate processing; the increase in memory performance they report is mainly driven by increased accuracy for predictable words, and Corley et al. (2007) suggest that the processing of these words is made more difficult (more effortful) by the preceding disfluency.

However, despite Corley et al.'s, (2007) suggestion that any memory advantage depends on processing effects indexed by the N400, it remains possible that immediate processing and subsequent memory effects are independent, and reflect slightly different mechanisms. Although Corley et al. (2007) and Collard et al. (2008) report no memory advantage for unpredictable words following disfluency, it does not necessarily follow that predictable words were made more difficult to process, and that it was this more effortful processing which led to more successful memory recognition. It seems likely that the raised attention following fillers (demonstrated by Collard et al., 2008) would lead to increased subsequent memory for the affected items, but this does not preclude the possibility of dissociation between N400 effects and memory performance.

One explanation for the failure of previous experiments to find a disfluency advantage on memory performance for unpredictable words is that they may have experienced a ceiling

effect. Participants were already accurate at identifying unpredictable words in fluent sentences, and so the added benefit of a filler may not have made any significant difference to performance. Alternatively, an explanation could be found with regards to attention: Collard et al. (2008) demonstrated that under fluent presentation, acoustically deviant words elicited robust MMN and P300 effects, but that these were very much reduced under disfluent presentation. The authors reasoned that this indicated that attention had already been captured and oriented by the disfluency, and so could not be further heightened by the acoustically deviant stimulus. Following similar logic, it is possible that attention is raised for unpredictable words even in fluent presentation, leading to more successful encoding, and so adding a disfluent filler before the unpredictable word does not raise attention any higher than it would already have been for that item, had it been presented fluently.

It is also possible that listeners use their experience with the distribution of disfluency to make predictions about upcoming material, as has been demonstrated in eye tracking studies such as Arnold, Fagnano and Tanenhaus (2003) and Arnold et al. (2004) , and that it is this change in prediction strategy which leads to the attenuated N400 effects reported for disfluent words (Corley et al., 2007). Collard et al. (2008) have provided a convincing demonstration of an effect of disfluency on attention, likely to lead to a memory advantage. Whilst it is possible that this increase in cognitive resources made available to the speech stream facilitates prediction changes, it does not necessarily follow that changes in prediction, indexed by the N400, and subsequent memory are co-dependent effects.

If Corley et al.'s, (2007) interpretation is correct in so far as memory performance and size of N400 effect at presentation reflect the same processes, then we would expect to see a memory advantage for words affected by disfluency in Experiment 1 (Chapter

5, Comparing Fillers to Beeps), but would not necessarily expect this advantage to extend to words which were preceded by a beep. We would not expect to see a memory advantage for words preceded by a beep as Experiment 1 showed a small difference in N400 effect size between fluent and disfluent words, but no significant difference in N400 effect size for words which had been preceded by a beep. By contrast, for Experiment 2, (Chapter 6, Comparing Fillers to Coughs), we would not necessarily expect to see any memory advantage, as this experiment, we found no effect of fluency on N400 effect size.

8.3 Comparing Fillers to Beeps

8.3.1 Comparing Fillers to Beeps — Memory Performance

This section reports the memory outcomes of the experiment reported in Chapter 5.Memory performance was assessed as the probability of participants correctly identifying old words. In the memory test, five words were inadvertantly repeated. Removal of these items from analysis resulted in 319 distinct targets. Overall, 66% of the old words and 61% of the new words were correctly recognised (39% false alarm rate). Participants were more successful at correctly identifying items which had appeared in unpredictable than predicable contexts, and more successful at identifying words that had been affected by a disfluency or a beep than words that had been presented in fluent contexts (see Table 8.1).

Participants likelihood of correctly identifying old words was assessed using a multilevel ANOVA, with factors of fluency and predictability, and using stimulus as a random factor. This ANOVA revealed significant main effects of predictability [$F(1,318) = 34.01$,

	predictable	unpredictable
fluent	60%	68%
disfluent (*er*)	65%	71%
beep	65%	70%
new	62%	

TABLE 8.1: Memory performance comparing fluent utterances with utterances interrupted by fillers and beeps. The table shows the mean probability of previously heard and new target words being correctly identified(n=26).

η^2_G = 0.028, p <0.00001] and fluency [F(2,636) = 7.09, p <0.001]. There was no significant interaction between predictability and fluency. Planned subsequent contrasts revealed significant differences between fluent and disfluent utterances [t(637) = 3.20, p <0.005], and between fluent utterances and those interrupted by a beep [t(637) = 3.20, p <0.005], but no significant difference between memory for targets which had been disfluent, and those which had been interrupted by a beep.

The data were also analysed for effects of fluency and predictability on participants reaction time to correctly identified targets, and their self-assessed confidence about their judgements. Reaction time and confidence effects were assessed using ANOVA with factors of predictability and fluency, and using stimulus as a random factor. Where any one stimulus did not have at least one incidence of a correct response in each of the six presentation conditions, all incidences of that stimulus were removed[1].

Words which had been unpredictable at presentation generally elicited slightly faster reactions than words which had been predictable (see table 8.2). Reaction time varied significantly with predictability [F(1,274) = 4.81, p <0.05], but did not vary with fluency, or show any significant interaction between predictability and fluency. Confidence judgements also showed a significant effect of predictability [F(1,274) = 16.35, p <0.0001],

[1]Failure to remove stimuli which did not have at least one incidence in each condition violated ANOVA as this would result in an unbalanced dataset, where all levels of fluency and predictability did not occur for all levels of stimulus

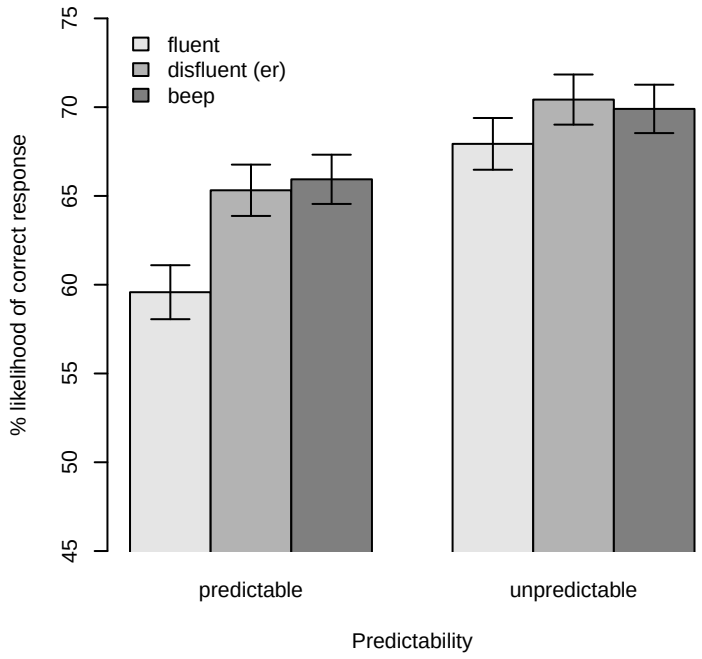

FIGURE 8.1: Probability of participants correctly identifying target words as old averaged across stimuli. Error bars represent one standard error of the mean. Unpredictable words were significantly more likely to be correctly identified than predictable words. Words from utterances which had been interrupted by a disfluent filler or a beep were significantly more likely to be remembered than words from utterances that had been fluent, but there was no difference between the types of interruption.

	predictable	unpredictable
fluent	905	916
disfluent (*er*)	917	902
beep	919	906
new	992	

TABLE 8.2: Mean reaction times (ms) to correctly identified old words comparing words from fluent utterances with those from utterances interrupted by fillers and beeps.

with participants slightly more confident in their responses to words which had been unpredictable at presentation. Neither the main effect of fluency nor an interaction of predictability and fluency reached significance.

	predictable	unpredictable
fluent	3.82	4.01
disfluent (*er*)	3.92	4.06
beep	3.95	4.08
new	3.26	

TABLE 8.3: Mean confidence levels for correctly identified old words comparing words from fluent utterances with those from utterances interrupted by fillers and beeps. Upon identifying target words as old or new, participants rated their own confidence in the answer they had just given using a five point scale, '5' representing "very confident", and '1' representing "very unsure".

8.3.2 Comparing Fillers to Beeps — ERP Results

Behavioural evidence (detailed in Section 8.3.1) revealed improved memory for items which had been interrupted, whether by a filler or a beep, at auditory presentation. The small size of previous studies has not allowed them to collect enough trials to investigate the electrophysiological correlates of this retrieval process, and to establish whether these effects vary with the condition of the stimulus at presentation.

EEG was collected from participants during the surprise memory test. ERPs were formed by processing the raw EEG in the same way as the EEG collected during the listening task. ERPs were quantified by comparing the ERP waveforms to targets which had been previously heard and were correctly identified as "old" to waveforms for targets which had not been heard, and were correctly identified as "new". This comparison focussed on two time windows of interest, 300-500ms after stimulus onset, where one might find an FN400 believed to reflect familiarity, and 500-800ms, where one might find the Left Parietal Old/New Effect (LPONE), believed to reflect recollection.

Data are based on 19 participants who generated the minimum usable 16 trials per condition. EEG was processed in accordance with the process laid out in section 4.8. As participants viewed 319 new words, and 319 old words, which were divided between

	fluent		er		beep		new
	predictable	unpredictable	predictable	unpredictable	predictable	unpredictable	
minimum	16	18	21	18	19	17	76
maximum	42	39	39	43	43	47	224
mode	16	39	33	30	25	27	N/A
mean	25.89	28.11	29.79	30.58	27.11	30.42	145.74

TABLE 8.4: Numbers of trials (per participant) included in ERP analysis for each condition (n=19), comparing fillers to beeps.

six conditions, there are far more trials incorporated in the grand average waveform to "new" words than to any of the "old" conditions, and this accounts for the difference in noise levels between the old and new conditions, apparent in the waveforms. The number of trials from each subject incorporated into the ERP for each fluency and predictability condition is detailed in Table 8.4.

Previously encountered words elicited ERPs which were generally more negative than ERPs to new words. This negativity onset at around 300ms at frontal locations, spreading backwards across the scalp and evident at parietal locations by 650ms. The negativity onsets slightly earlier in the right hemisphere than the left.

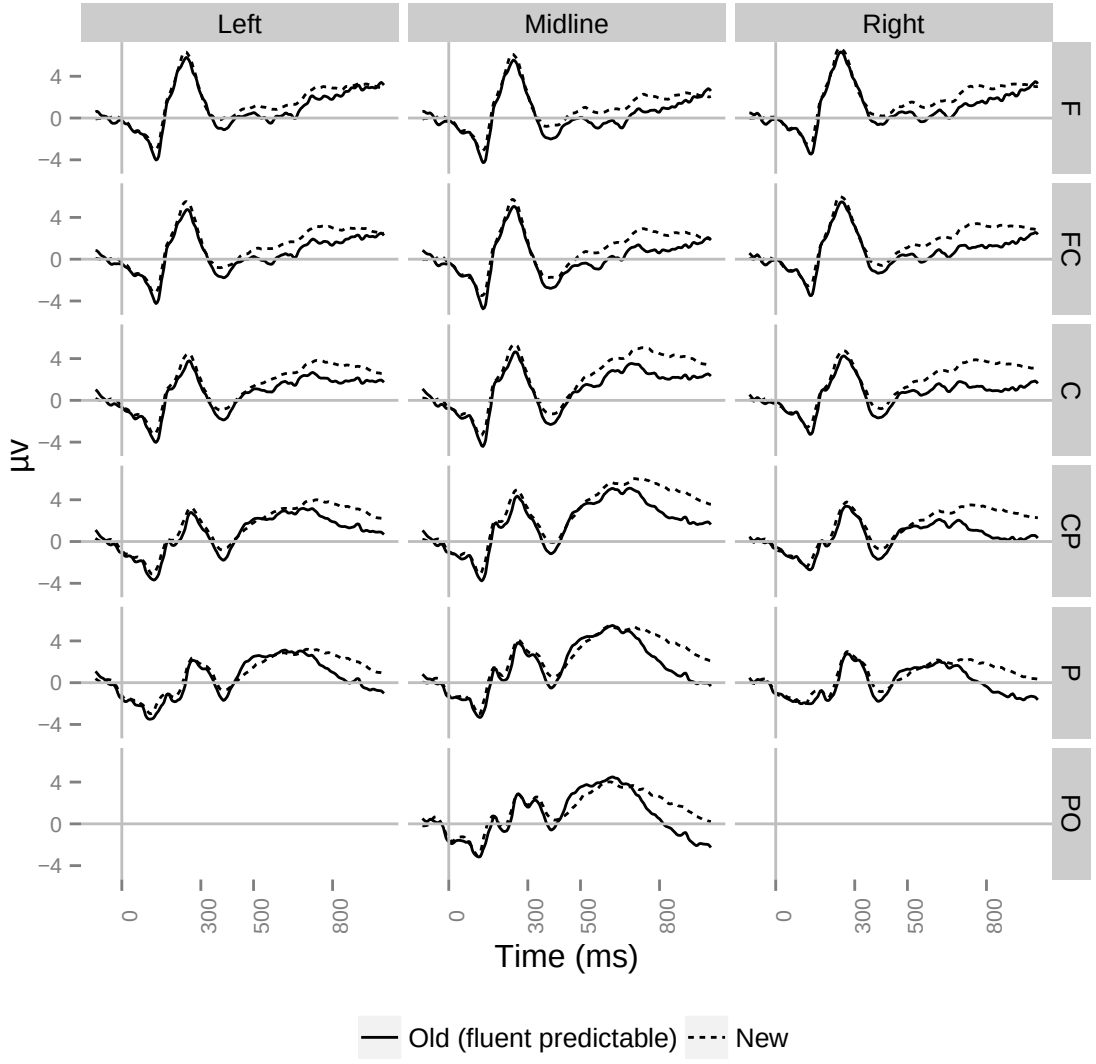

FIGURE 8.2: Grand average ERPs (n=19) for targets which had been predictable and fluent at original presentation, as well as new targets. Shown here are ERPs as measured at frontal (F), fronto-central (FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline.

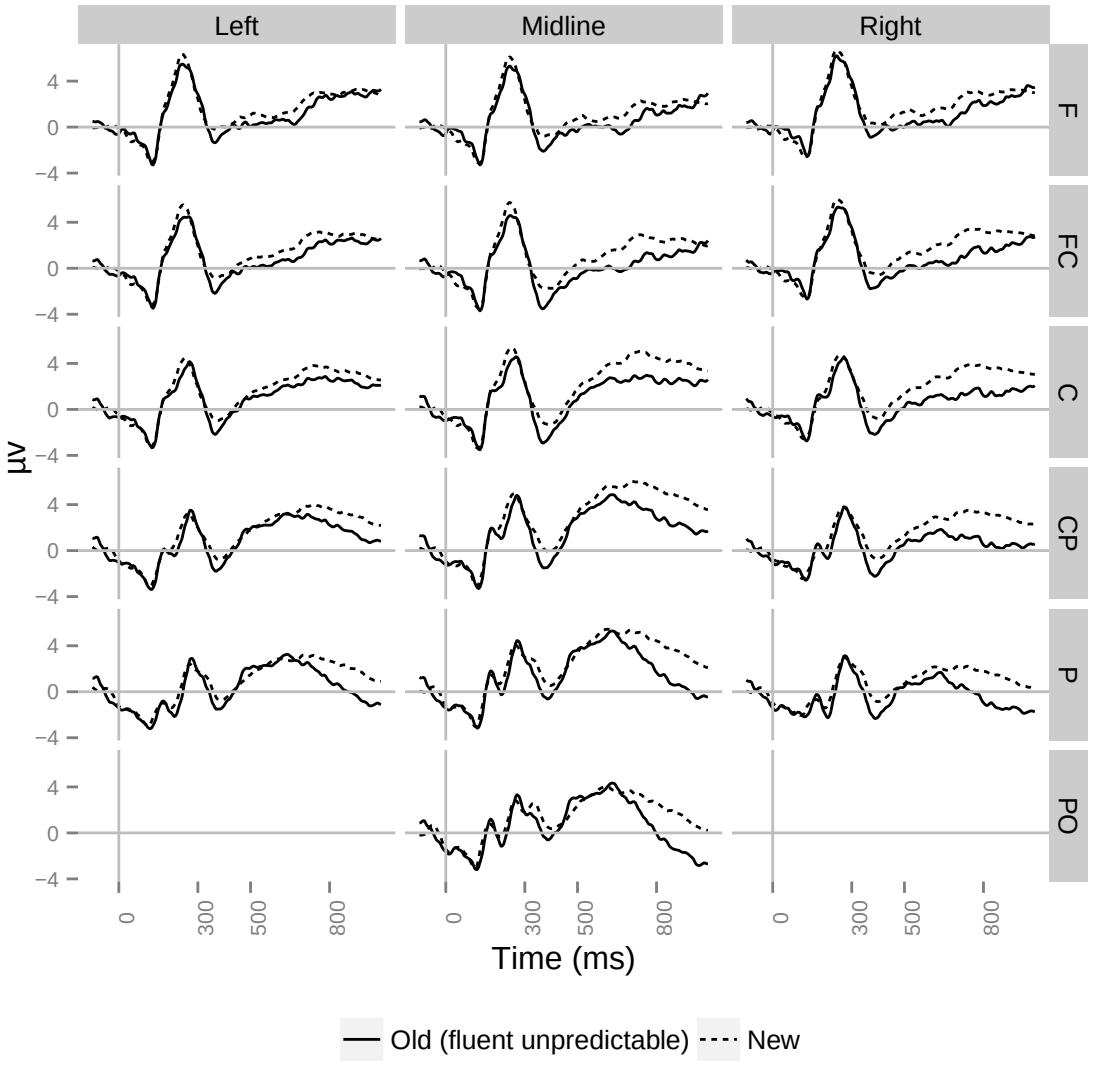

FIGURE 8.3: Grand average ERPs (n=19) for targets which had been unpredictable and fluent at original presentation, as well as new targets. Shown here are ERPs as measured at frontal (F), fronto-central (FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline.

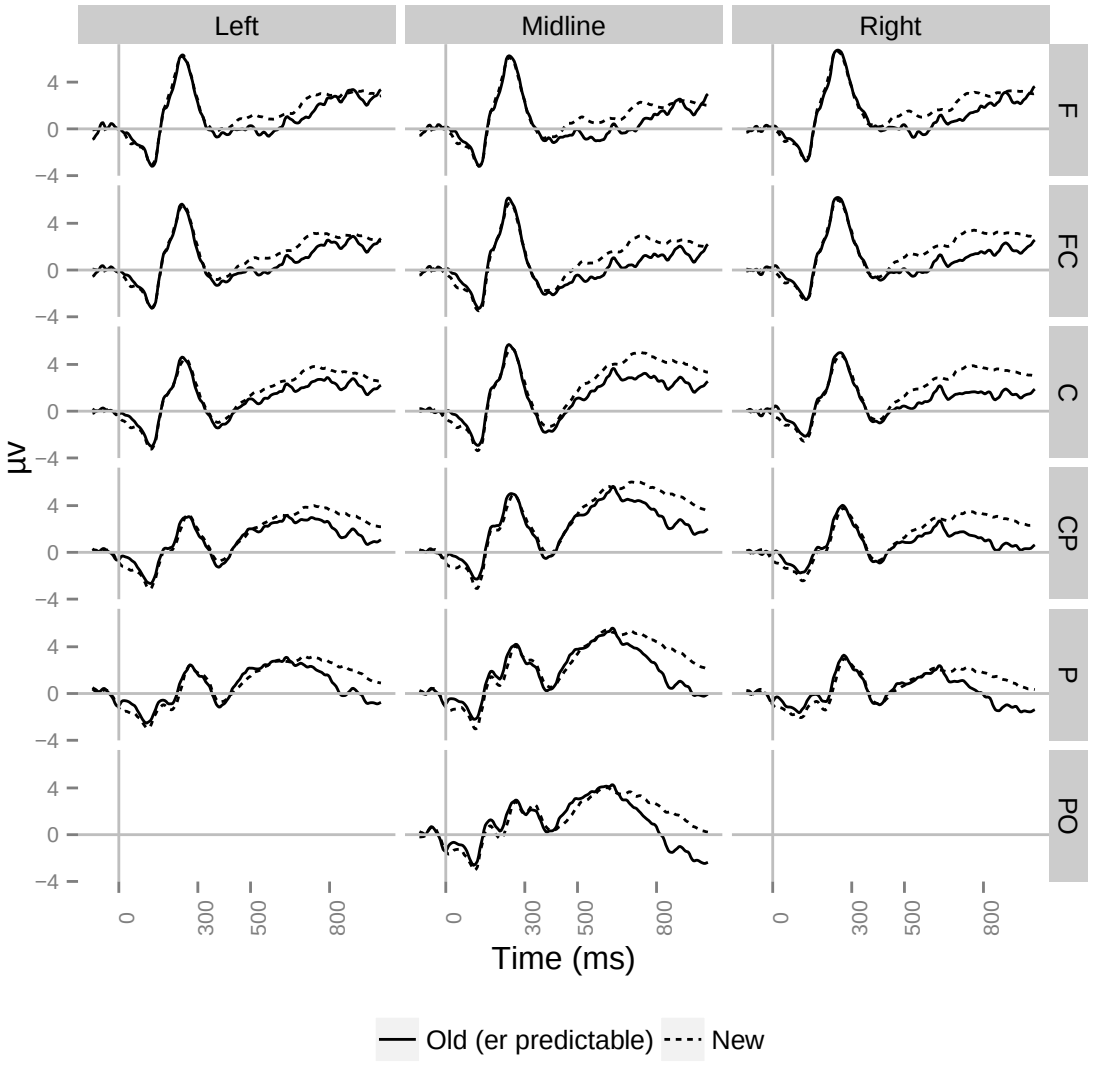

FIGURE 8.4: Grand average ERPs (n=19) for targets which had been predictable and disfluent at original presentation, as well as new targets. Shown here are ERPs as measured at frontal (F), fronto-central (FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline.

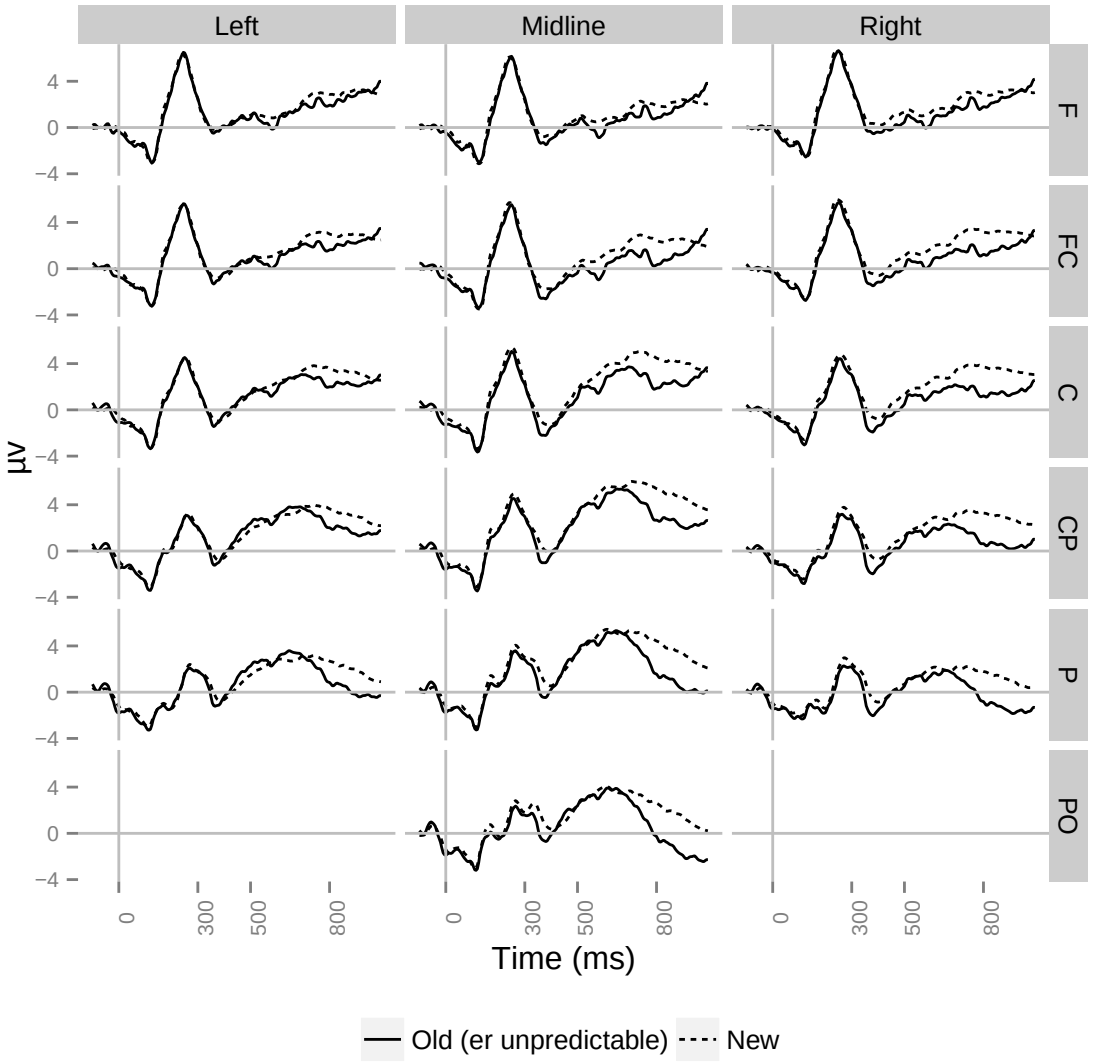

FIGURE 8.5: Grand average ERPs (n=19) for targets which had been unpredictable and disfluent at original presentation, as well as new targets. Shown here are ERPs as measured at frontal (F), fronto-central (FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline.

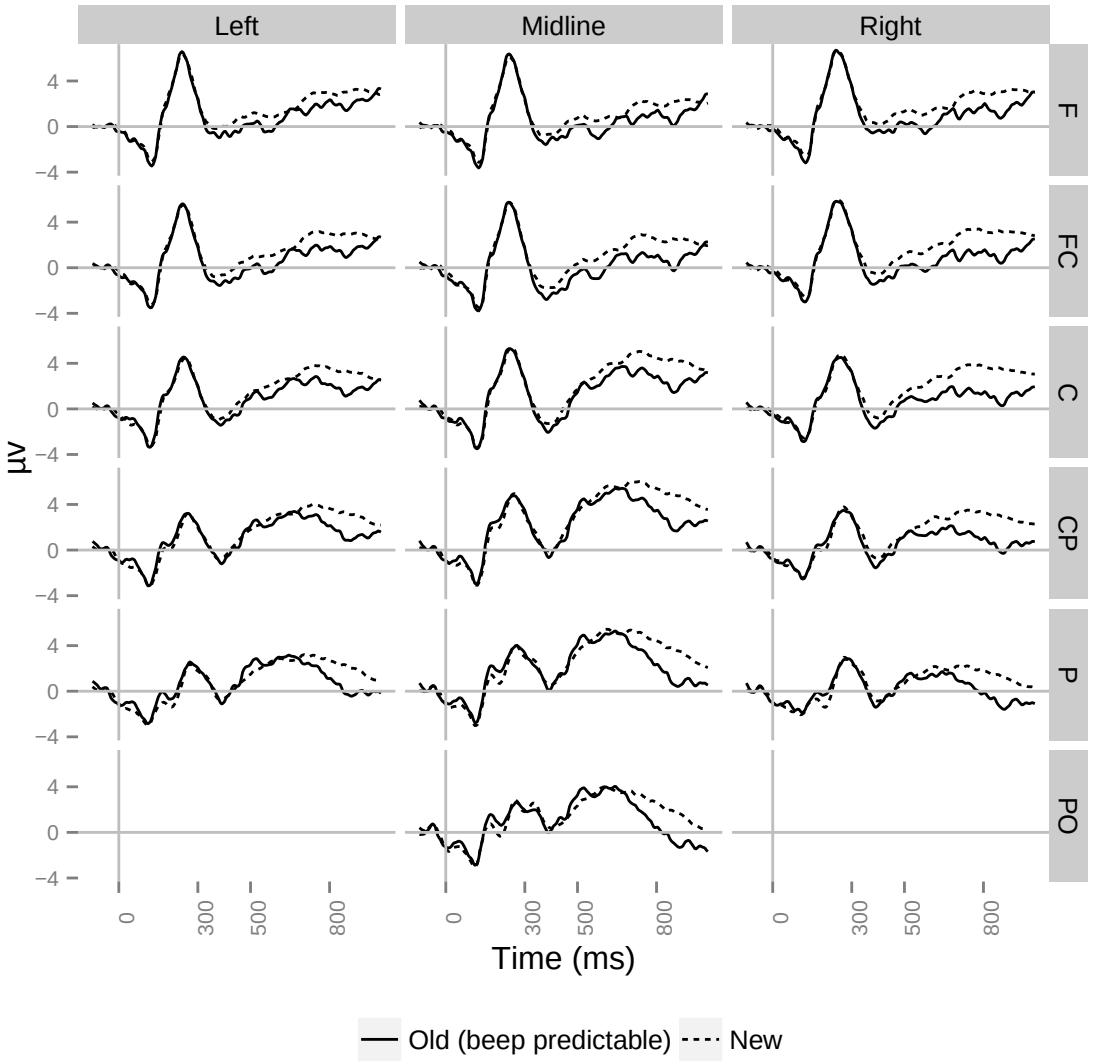

FIGURE 8.6: Grand average ERPs (n=19) for targets which had been predictable and interrupted by a beep at original presentation, as well as new targets. Shown here are ERPs as measured at frontal (F), fronto-central (FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline.

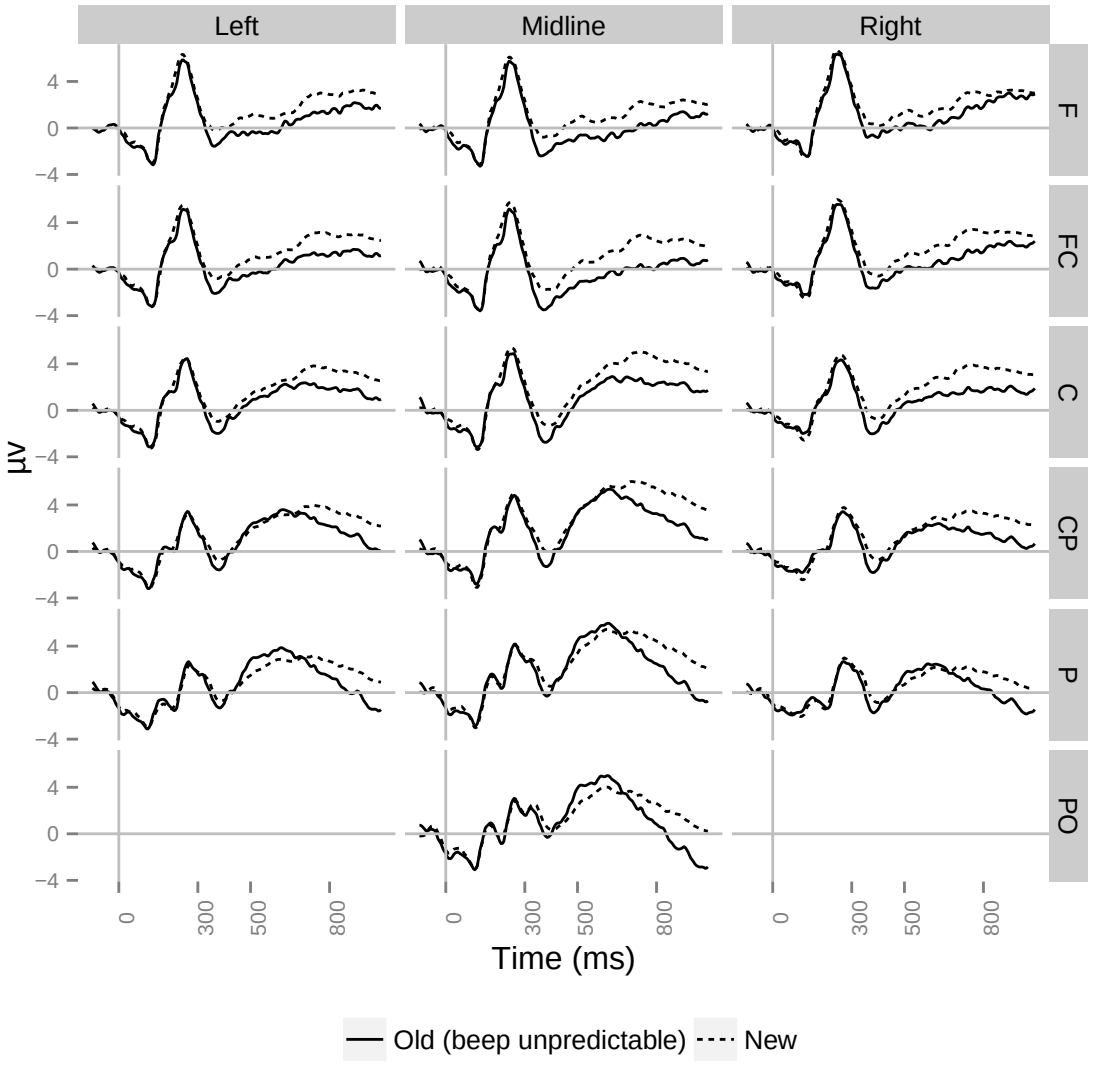

FIGURE 8.7: Grand average ERPs (n=19) for targets which had been unpredictable and interrupted by a beep at original presentation, as well as new targets. Shown here are ERPs as measured at frontal (F), fronto-central (FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline.

Comparing Fillers to Beeps — 300-500ms

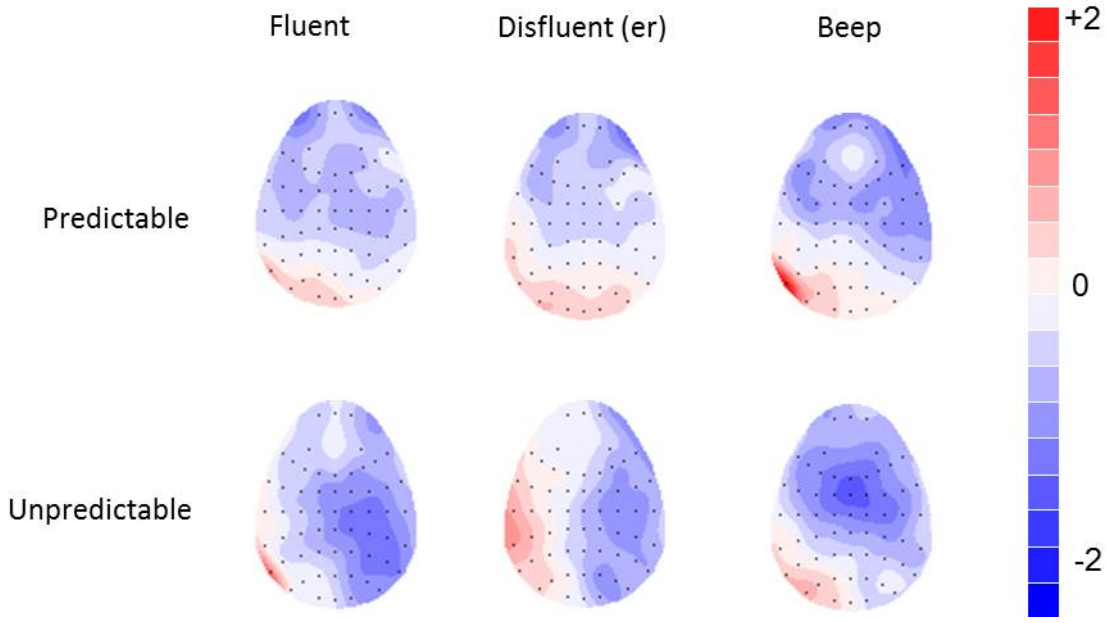

FIGURE 8.8: Retrieval effects (hits - correct rejections) in the 300-500ms time window (n=19). No mid-frontal old/new effects are seen, but for predictable words which had been affected by disfluency or a beep at initial presentation, there is a parietal positivity.

In the 300-500ms latency range, visual inspection of the data did not find any evidence of a mid-frontal old/new effect (FN400), but instead suggested a small left parietal positivity for previously studied items compared to new items. To confirm whether the impressions gained from visual inspection of the data are statistically robust, the data from each fluency and predictability condition were submitted to ANOVA with factors of condition *(*old, new), location *(F, FC, C, CP, P)* and hemisphere *(left, right)*, and incorporating data from the electrodes specified for global ANOVA in Figure 4.2. These ANOVA partially confirmed the impression given by Figure 8.8; for words which had featured as predictable endings in fluent, disfluent and interrupted utterances, no signficant or marginally significant differences were found between old and new items.

For unpredictable words which had featured in fluent sentences, there was a marginally significant interaction of condition with hemisphere [$F(1,18) = 3.64$, $\eta^2_G = 0.001$, $p < 0.1$] and a marginally significant interaction of condition with location and hemisphere

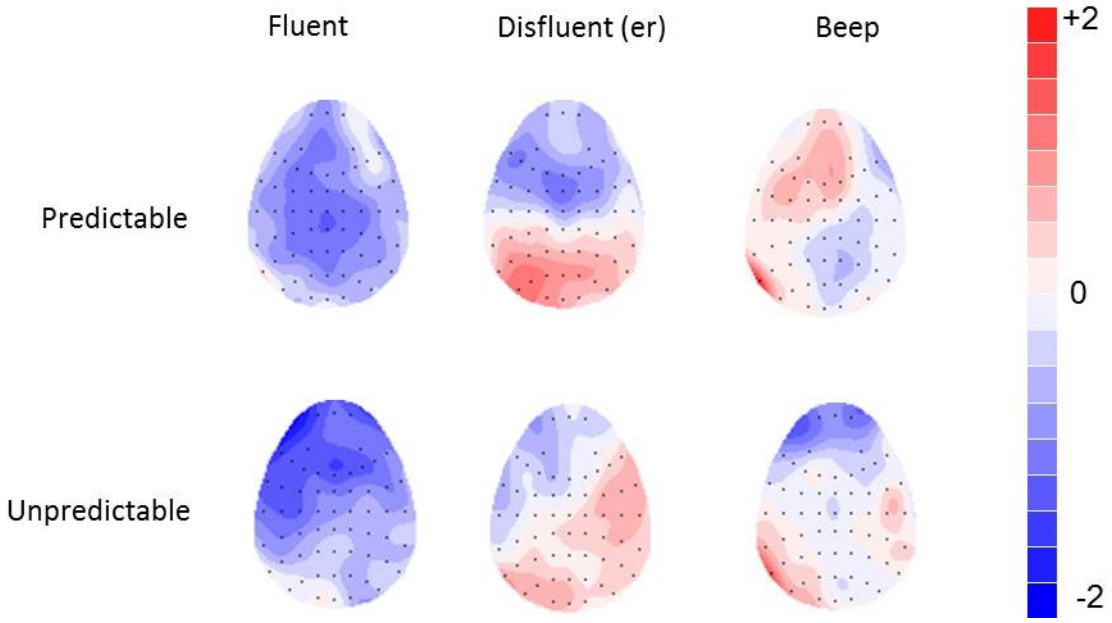

FIGURE 8.9: Repetition effects (misses - correct rejections) in the 300-500ms time window (n=19). Effects seen here reflect processing differences between forgotten 'old' words, and genuinely new items. The parietal positivity seen in response to forgotten words which had been disfluent or interrupted by a beep may represent a repetition or implicit memory effect.

[F(1.68,30.28) = 3.40, η^2_G = 0.0001, p <0.1], reflecting an ERP effect which was more positive in the left hemisphere at posterior locations, but not at frontal locations. Unpredictable disfluent words also featured a gradient across the hemispheres, with the left hemisphere again showing a more positive ERP than the right, reflected in the significant interaction of condition with hemisphere [F(1,18) = 10.78, η^2_G = 0.003, p <0.005]. For unpredictable words which had been preceded by a beep, however, there was no significant effect of hemisphere; but a marginally significant interaction of condition with location [F(1.16,20.89) = 3.04, η^2_G = 0.002, p <0.1], reflecting an effect which was more positive towards posterior locations.

Comparison of the ERP for each old condition with the ERP for new words appears to show relative positivity for old words, which is greater at posterior sites and in the left hemisphere (see Figure 8.8). A left parietal positivity for old items in the 300-500ms latency range has been reported as a repetition effect (Rugg et al., 1998; Bridson, Fraser,

Herron, & Wilding, 2006), which does not predict the accuracy of memory judgements. Repetition effects should thus also be found when old items which were incorrectly identified as new (misses) are compared to genuinely unstudied items. This difference between forgotten items and new items has led to this repetition effect being referred to as a marker of "implicit memory" (Rugg et al., 1998). In order to investigate whether these left parietal positivities should be considered repetition effects or an early onset-ting LPONE, misses in each condition were inspected. Topographic maps illustrating difference effects between misses in each condition and correctly identified new items can be seen in Figure 8.9. There appeared to be a parietal positivity for forgotten predictable words which had been preceded by disfluency. For words which had been unpredictable and disfluent, and words which had been preceded by a beep, there was some evidence of parietal positivity, but this was less clearly defined. For words which had been unpredictable and fluent, the ERP was less negative at posterior sites, possibly interpretable as posterior positivity masked by a widespread negativity. No evidence of parietal positivity was seen for forgotten words which had been predictable fluent and fluent at initial presentation.

It should of course be borne in mind that comparing hits and misses from the same subjects is not entirely straightforward - by virtue of having reached the minimum criterion of sixteen good trials per condition, most of these subjects did not produce sixteen misses which also passed all of the filtering and artifact rejection criteria. The imbalance in contribution to noise levels across subjects resulting from the uneveness in trial numbers renders ANOVA analysis of these data weak, and visual inspection spurious, and so no further analysis is reported on these potential 'repetition effects'. However, Figure 8.9 is included here to help in understanding what may be driving the posterior 'memory effects' seen in FigureBeeps ERP memory 300-500ms topo.

As there was no evidence for any mid-frontal old/new effects in the 300-500ms latency range, no further analysis on this time window is reported.

Comparing Fillers to Beeps — 500-800ms

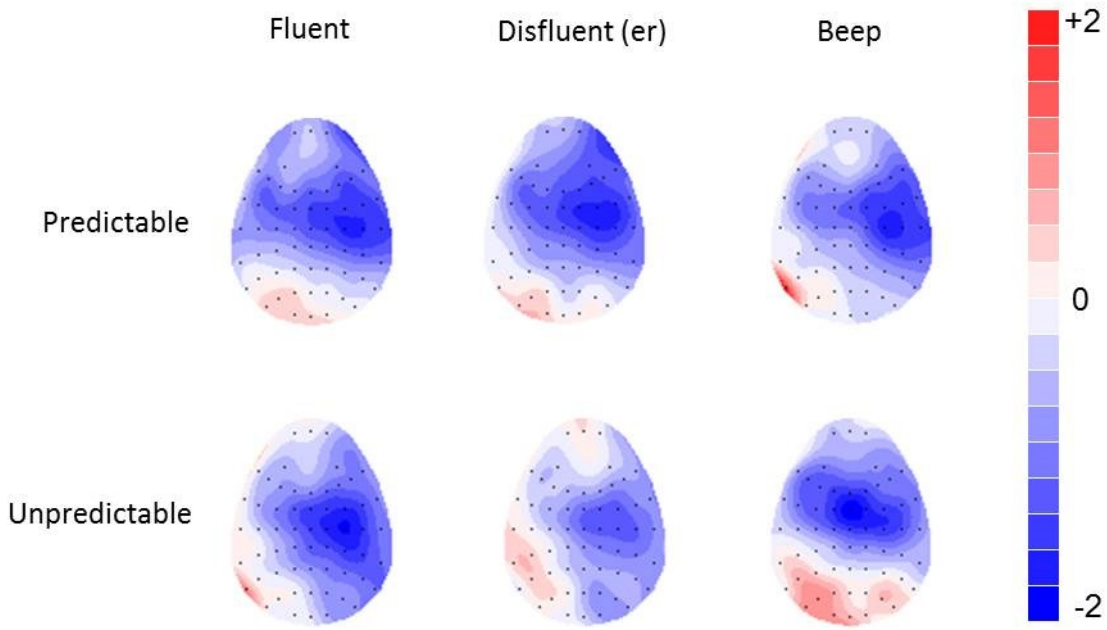

FIGURE 8.10: Retrieval effects (hits - correct rejections) in the 500-800ms time window (n=19). Possible left Parietal Old-New Effects (LPONE) are seen for all six conditions, although the ERP is dominated by a large mid-frontal relative negativity with a right hemisphere bias.

In the 500-800ms latency range, visual inspection of the data reveals a small left parietal positivity occurring for items in all conditions, although this effect was not broadly spread over the scalp as may be expected for a LPONE, but rather tightly focussed over parietal-occipital sites in the left hemisphere, while a relative negativity for old words dominated most of the scalp. Multilevel ANOVA with factors of condition *(old, new)*, location *(F, FC, C, CP, P)* and hemisphere *(left, right)* was performed for each fluency and predictability condition, in order to establish whether any observed effects were reliable.

For predictable words, ANOVA revealed significant or marginally significant effects of condition *(old, new)* in all three fluency conditions; fluent [F(1,18) = 6.81, η^2_G = 0.28, p <0.05], er [F(1,18) = 3.67, η^2_G = 0.028, p <0.1], beep [F(1,18) = 3.60, η^2_G = 0.023, p <0.1], reflecting an overall relative negativity for previously heard words. The ANOVA also revealed marginally significant interactions of condition with location for items which had been fluent [F(1.19,21.47) = 3.71, η^2_G = 0.004 p <0.1] or interrupted by a beep [F(1.17,21.15) = 3.09, η^2_G = 0.002, p <0.1], reflecting the fact that the negativity to old words was greater at fronto-central and central locations. No other interactions involving condition reached significance.

For items which had been unpredictable and fluent at auditory presentation, there was a significant effect of condition [F(1,18) = 5.62, η^2_G = 0.024, p <0.05], reflecting a general relative negativity to old words, and a marginally significant interaction of condition with hemisphere [F(1,18) = 3.60, η^2_G = 0.003, p <0.1], reflecting the right hemisphere bias of this negativity. Similarly, words which had been unpredictable and disfluent also showed an interaction of condition with hemisphere [F(1,18) = 5.17, η^2_G = 0.003, p <0.05]. Unpredictable words which had been preceded by a beep at auditory presentation, however, showed no effects of hemisphere but the ANOVA did reveal a significant interaction of condition with location [F(1.12,20.11) = 9.83, η^2_G = 0.010, p <0.005], reflecting an ERP difference which was negative at central and frontal locations, and more positive at posterior locations.

It is not clear to what extent these ERP effects can be interpreted as LPONE, as it appears likely that any LPONE would be masked by the large mid-frontal negativity for old words. Another way to view these data is, of course, to think in terms of a whole-head negativity, which is being pushed out of the left hemisphere and posterior locations by a LPONE.

Topographic Comparison

To establish whether any left parietal old/new effects vary across the six fluency and predictability conditions, the rescaled difference waveforms were assessed for variability in topographic distribution. A multilevel ANOVA with factors of predictability *(predictable, unpredictable)*, fluency *(fluent, er, beep)*, location *(F, FC, C, CP, P)* and hemisphere *(left, right)* revealed significant main effects of predictability [$F(1,18) = 13.21$, $\eta^2_G = 0.024$, $p < 0.005$] and fluency [$F(1.63,29.45) = 67.08$, $\eta^2_G = 0.316$, $p < 0.0001$], as well as a significant interaction of predictability with fluency and location [$F(2.72,48.92) = 4.89$, $\eta^2_G = 0.004$, $p < 0.01$] reflecting the fact that the focus of the negativity varies by both fluency and predictability. Given these topographic differences between conditions, quantitative comparison of all six fluency and predictability conditions is not justified. Visual inspection of the scalp topographies shown in Figure 8.8 suggests a strong similarity between the distributions of effects for predictable words. Neither global topographic ANOVA nor a follow-up midline ANOVA revealed any significant or marginally significant differences in topography across fluency conditions for predictable items. There is no reason to suppose that different neural generators underlie these effects and so they are quantitatively compared below. By contrast, topographic global ANOVA of old/new effects for unpredictable items revealed a marginally significant interaction of fluency with location [$F(1.85,33.37) = 2.52$, $\eta^2_G = 0.005$, $p = 0.1$].The effect of location becomes slightly more significant when data from the midline only are considered; [$F(2.43,43.78) = 2.72$, $\eta^2_G = 0.253$, $p < 0.1$]. Therefore, no quantitative comparison of old/new effects is reported for unpredictable words.

Quantitative Comparison

Difference waveforms from fluent, disfluent and interrupted (beep) items which had been predictable at auditory presentation were submitted to ANOVA with factors of fluency *(fluent, er, beep)*, location *(F, FC, C, CP, P)* and hemisphere *(left, right)*. Difference waveforms were used (rather than raw data and a factor of condition) because there were no fluency conditions for new data, making ANOVA structure impossible. This ANOVA revealed no main effects or interactions implicating fluency, indicating that there are no significant differences between old/new effects in the three fluency conditions, for predictable words at least.

8.3.3 Comparing Fillers to Beeps — Summary of Memory Results

Despite performance differences in the recognition test, demonstrating that participants are more successful at recognising words that had been unpredictable in their contexts, and words which had been either disfluent or interrupted, it is not clear that this experiment has produced ERPs consistent with these findings.

Improved memory performance for words which had been upredictable, as well as for those which had been disfluent, replicates the findings of Corley et al. (2007), upon whose paradigm this experiment is based. Corley and colleagues reported significantly higher recognition rates for words which had been unpredictable at auditory presentation, as well as for words which had been preceded by a disfluent filler *(er)*, particularly when these disfluent words had been predictable at original presentation. Although the performance data presented here does not present a significant interaction between predictability and fluency, visual inspection of the results (*c.f.* Figure 8.1) does suggest a slightly larger effect of fluency for predictable words. Importantly, this experiment has

uncovered no difference between the recognition rates of words which had been disfluent, and those which had been interrupted by a beep.

The findings of this experiment may be considered to replicate the findings of MacGregor et al. (2010), who, using the same paradigm as Corley et al. (2007), reported improved memory for items which had been preceded by a silent pause at initial auditory presentation. In contrast, Fraundorf and Watson (2011) reported that in a story-retelling task, fillers facilitated participants' long term memory, while coughs, presented as a non-linguistic interruption to speech, impaired recall. It is possible that the task differences between the present experiment and that reported by Fraundorf and Watson (2011) may account for the difference in findings between that experiment and the data presented here, or it may be that a beep and a cough are treated as fundamentally different by the listener, perhaps on the basis of whether or not the interruption is speaker generated or controlled.

In light of the clear performance differences between fluency and predictability conditions, the failure of this experiment to reveal ERP retrieval effects is somewhat suprising. No evidence was found of mid-frontal old/new effects in the 300-500ms latency range, and in the 500-800ms range, only weak evidence was revelealed for LPONE. It was not possible to definitively establish to what extent an early parietal positivity seen at 300-500ms may have reflected a repetition effect, or whether this should be considered the beginnings of an LPONE. Visual inspection of the ERP waveforms and scalp topographies suggests that an LPONE may be present, but is heavily masked by a whole-scalp relative negativity to old words, which onsets between 300ms and 400ms at frontal locations in the right hemisphere, spreading to encompass the whole scalp by ca. 600ms.

In the 500-800ms latency range, a small parietal positivity was observed for words which

had been predictable at auditory presentation, as well as for unpredictable words which had been preceded by a beep. For predictable words, no significant differences in scalp topography or amplitude were found between fluency conditions. It might be possible to interpret this positivity as a LPONE, but here there should be a note of caution; inspection of the unfiltered ERPs (incorporating both hits and misses) also appears to reveal very similar effects, making it difficult to be confident that these are retrieval effects, and not simply a continuation of any repetition effects seen earlier in the epoch.

8.4 Comparing Fillers to Coughs

8.4.1 Comparing Fillers to Coughs — Memory Performance

This section reports the memory outcomes of the experiment reported in Chapter 6. As in the previous experiment, five words were inadvertantly repeated. Removal of these items from analysis resulted in 319 distinct targets. Overall, 70% of old words and 59% of the new words were correctly identified (41% false alarm rate). Words which had been unpredictable were more successfully recognised than predictable words (see Table 8.5). The probability of words being correctly identified as old was assessed using a multilevel ANOVA, with factors of predictability and fluency, and using stimulus identity as a random factor. This revealed main effects of predictability [$F(1,318) = 62.51$, $p < 0.00001$] and fluency [$F(1,636) = 5.17$, $p < 0.01$], but no significant interaction of predictability with fluency. Planned subsequent contrasts revealed significant differences between the memorability of words in fluent and disfluent utterances [$t(637) = 2.87$, $p < 0.005$], and words in fluent utterances and those interrupted by a cough [$t(637) = 2.97$, $p < 0.005$]. There was no significant difference between words in utterances interrupted by a disfluent filler and those interrupted by a cough.

	predictable	unpredictable
fluent	62%	73%
disfluent (*er*)	67%	75%
cough	67%	75%
new	59%	

TABLE 8.5: Memory performance comparing fluent utterances with utterances interrupted by fillers and coughs. The table shows the mean probability of previously heard and new target words being correctly identified (n=24).

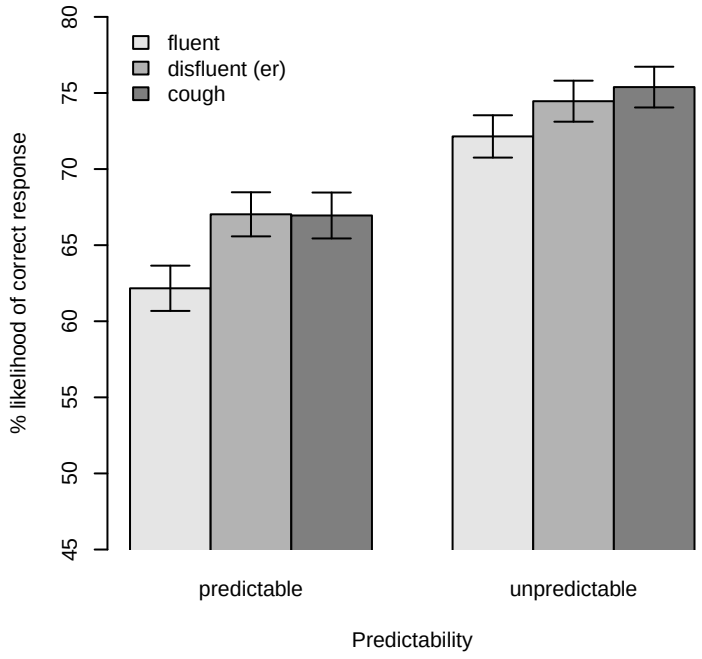

FIGURE 8.11: Probability of participants correctly identifying target words as old averaged across stimuli. Error bars represent one standard error of the mean. Unpredictable words were significantly more likely to be correctly identified than predictable words. Words from utterances which had been interrupted by a disfluent filler or a cough were significantly more likely to be remembered than words from utterances that had been fluent, but there was no difference between the types of interruption.

	predictable	unpredictable
fluent	885	902
disfluent (*er*)	892	881
cough	881	896
new	965	

TABLE 8.6: Mean reaction times (ms) to correctly identified old words comparing words from fluent utterances with those from utterances interrupted by fillers and coughs. Reaction time does not vary significantly with predictability or fluency.

The impact of predictability and fluency on reaction time and confidence judgements was analysed using an ANOVA with factors of predictability and fluency, and taking stimulus as a random factor. Where any stimulus items did not contribute at least one correct response in each condition, all incidences of that stimulus were removed from analysis, leaving 279 distinct targets. Reaction time to correctly identified old stimuli was not significantly impacted by the predictability or fluency of words (see Table 8.6), but participants' confidence judgements about their correct answer were affected by the words' predictability at original presentation [$F(1,278) = 12.59$, $p < 0.0005$], as well as their fluency [$F(2,556) = 4.28$, $p < 0.05$]. This reflects slightly higher confidence judgements for words which had been preceded by a cough than words that had been fluent [$t(1103) = 2.61$, $p < 0.01$], but no significant difference between fluent words and those preceded by a disfluent filler, or between disfluent words and those preceded by a cough. There was no significant interaction between predictability and fluency. For a summary of mean confidence judgements, see Table 8.7.

8.4.2 Comparing Fillers to Coughs — ERP Results

Data are based on 21 participants who contributed a minimum of 16 usable trials per condition. Numbers of trials incorporated into the ERP waveforms and subsequent analysis are are detailed in Table 8.8.

	predictable	unpredictable
fluent	3.72	3.84
disfluent (*er*)	3.82	3.93
cough	3.81	3.97
new	3.18	

TABLE 8.7: Mean confidence levels for correctly identified old words comparing words from fluent utterances with those from utterances interrupted by fillers and coughs. Upon identifying target words as old or new, participants rated their own confidence in the answer they had just given using a five point scale, '5' representing "very confident", and '1' representing "very unsure". On average, participants confidence was higher in response to words that had been unpredictable, and slightly higher for words preceded by a cough than fluent words.

	fluent		er		cough		new
	predictable	unpredictable	predictable	unpredictable	predictable	unpredictable	
minimum	17	17	17	21	19	21	41
maximum	33	40	40	39	35	41	201
mode	23	29	32	32	22	31	150
mean	25.19	29.90	28.71	31.71	26.52	31.48	138.48

TABLE 8.8: Numbers of trials included in ERP analysis for each condition per participant (n=21), comparing fillers to coughs.

As in the previously described experiment, ERPs to previously heard words were typically more negative than ERPs to new words. This negativity onset around 300ms at frontal locations for words which had been fluent at original presentation, spreading over the scalp and becoming evident at parietal locations around 650ms. For words which had been disfluent, or interrupted by a cough, the relative negativity for old words onset somewhat later, between 350ms and 500ms post-stimulus, again spreading backwards over the scalp. Despite the general positive shift for new words, ERPs to previously encountered items do appear to show a relative positivity at posterior locations between 400ms and 650ms.

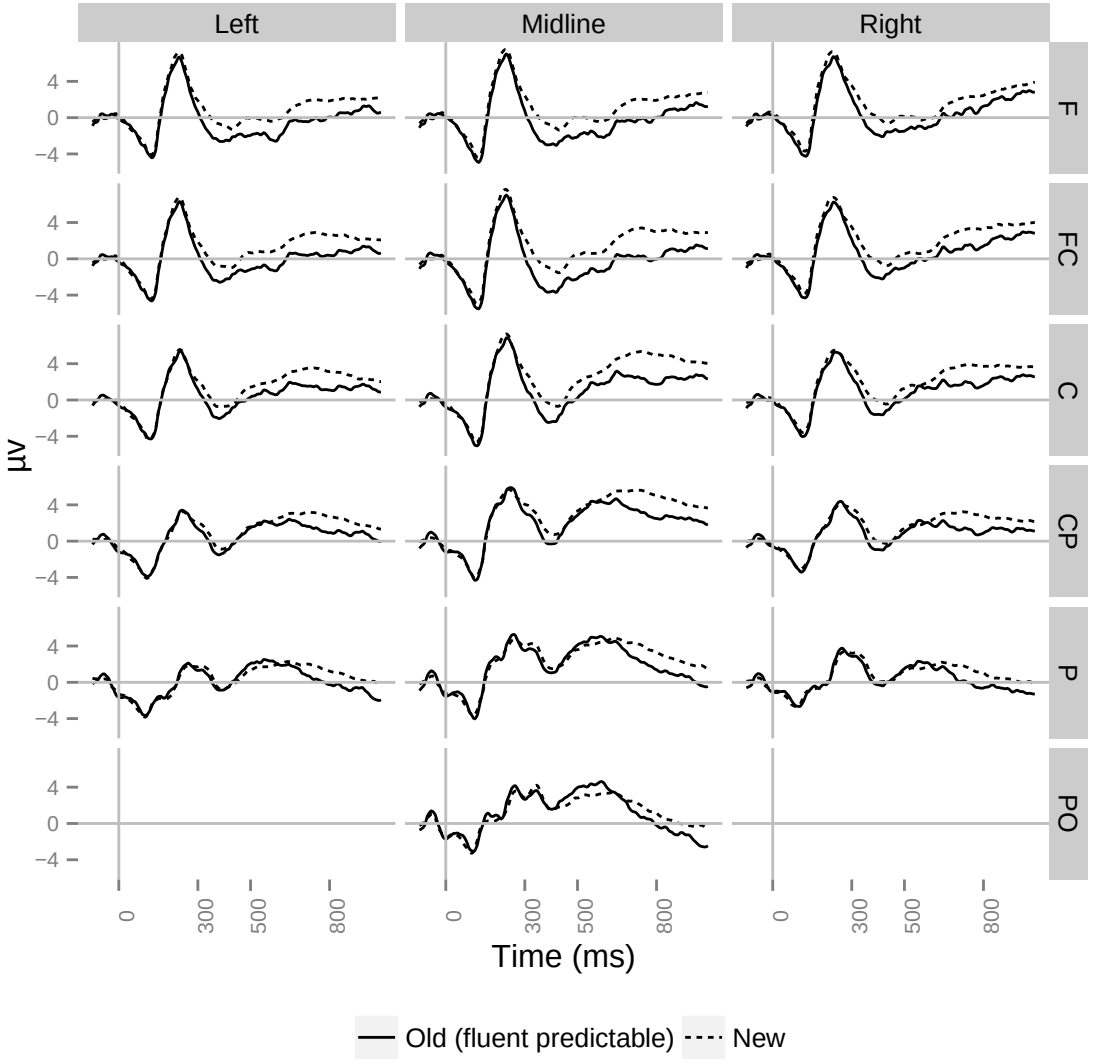

FIGURE 8.12: Grand average ERPs (n=21) for targets which had been predictable and fluent at original presentation, as well as new targets. Shown here are ERPs as measured at frontal (F), fronto-central (FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline.

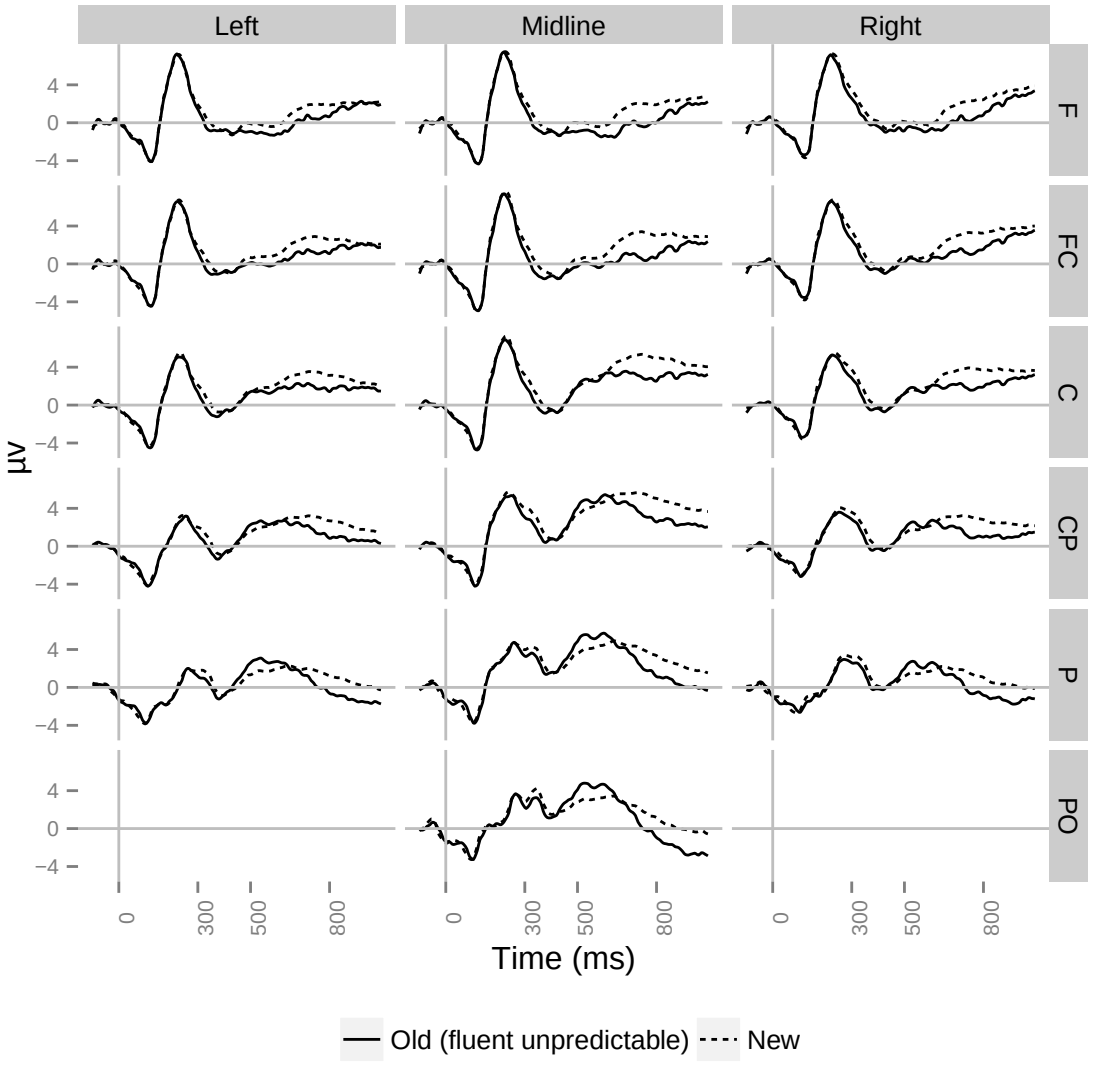

FIGURE 8.13: Grand average ERPs (n=21) for targets which had been unpredictable and fluent at original presentation, as well as new targets. Shown here are ERPs as measured at frontal (F), fronto-central (FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline.

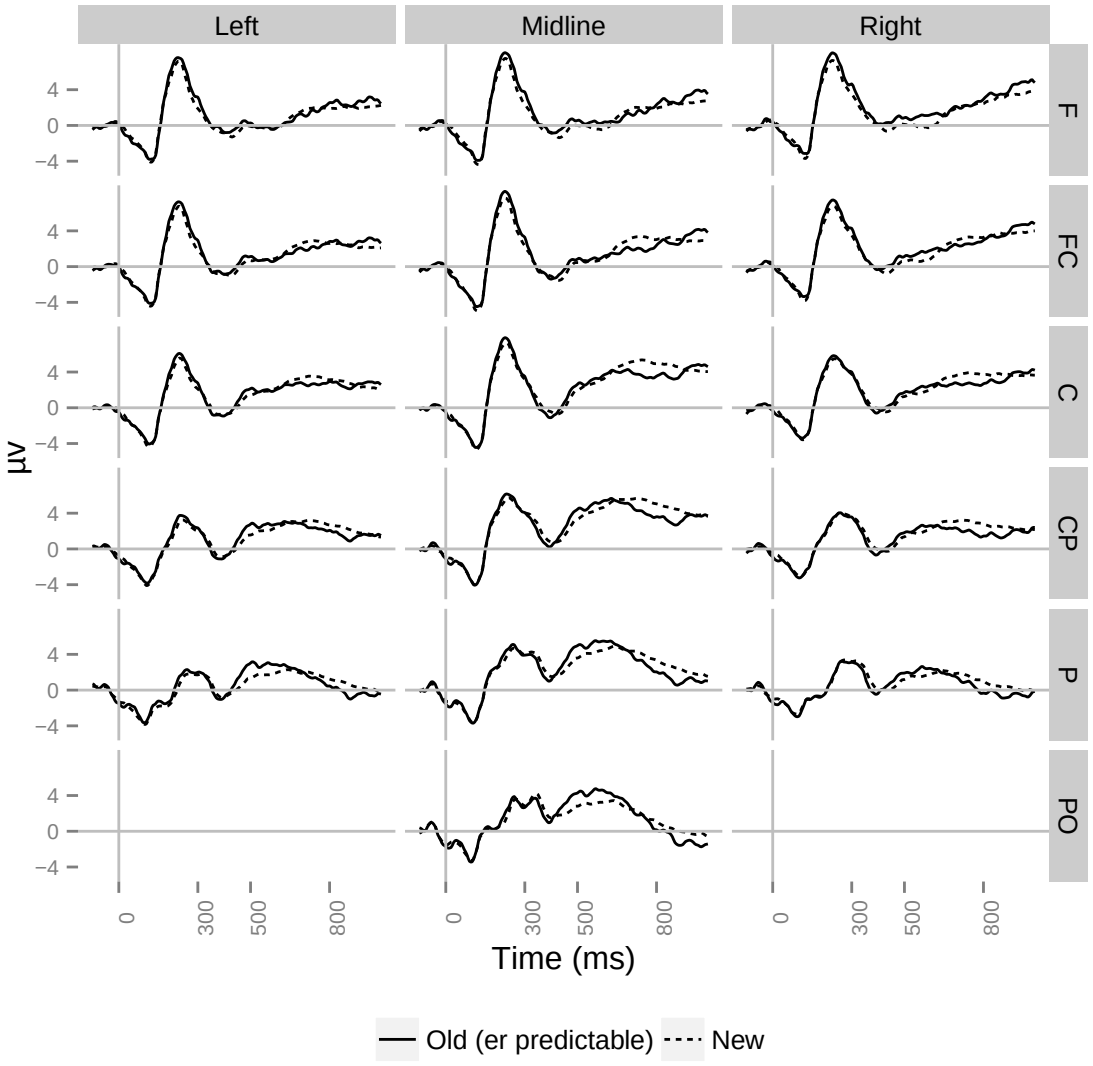

FIGURE 8.14: Grand average ERPs (n=21) for targets which had been predictable and disfluent at original presentation, as well as new targets. Shown here are ERPs as measured at frontal (F), fronto-central (FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline.

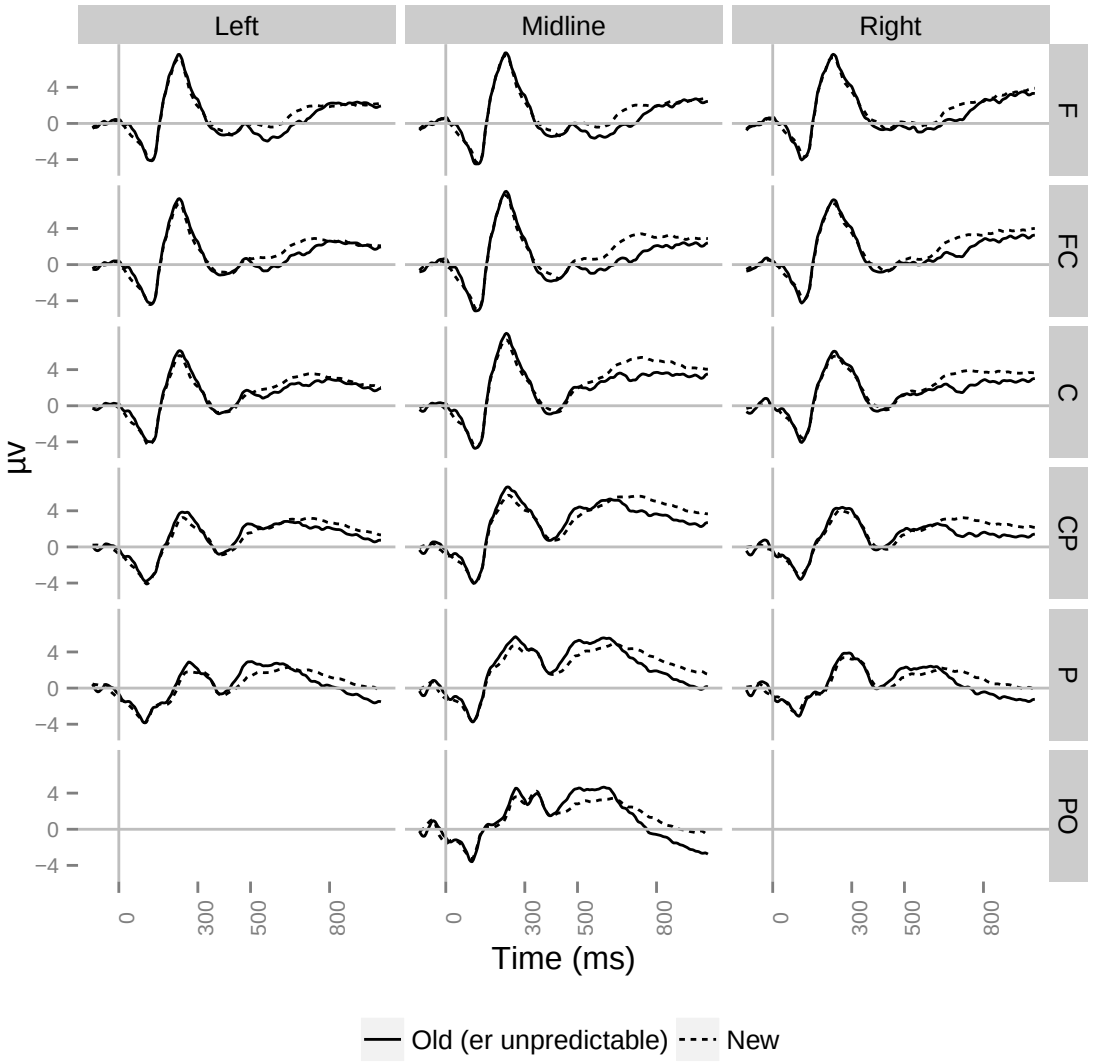

FIGURE 8.15: Grand average ERPs (n=21) for targets which had been unpredictable and disfluent at original presentation, as well as new targets. Shown here are ERPs as measured at frontal (F), fronto-central (FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline.

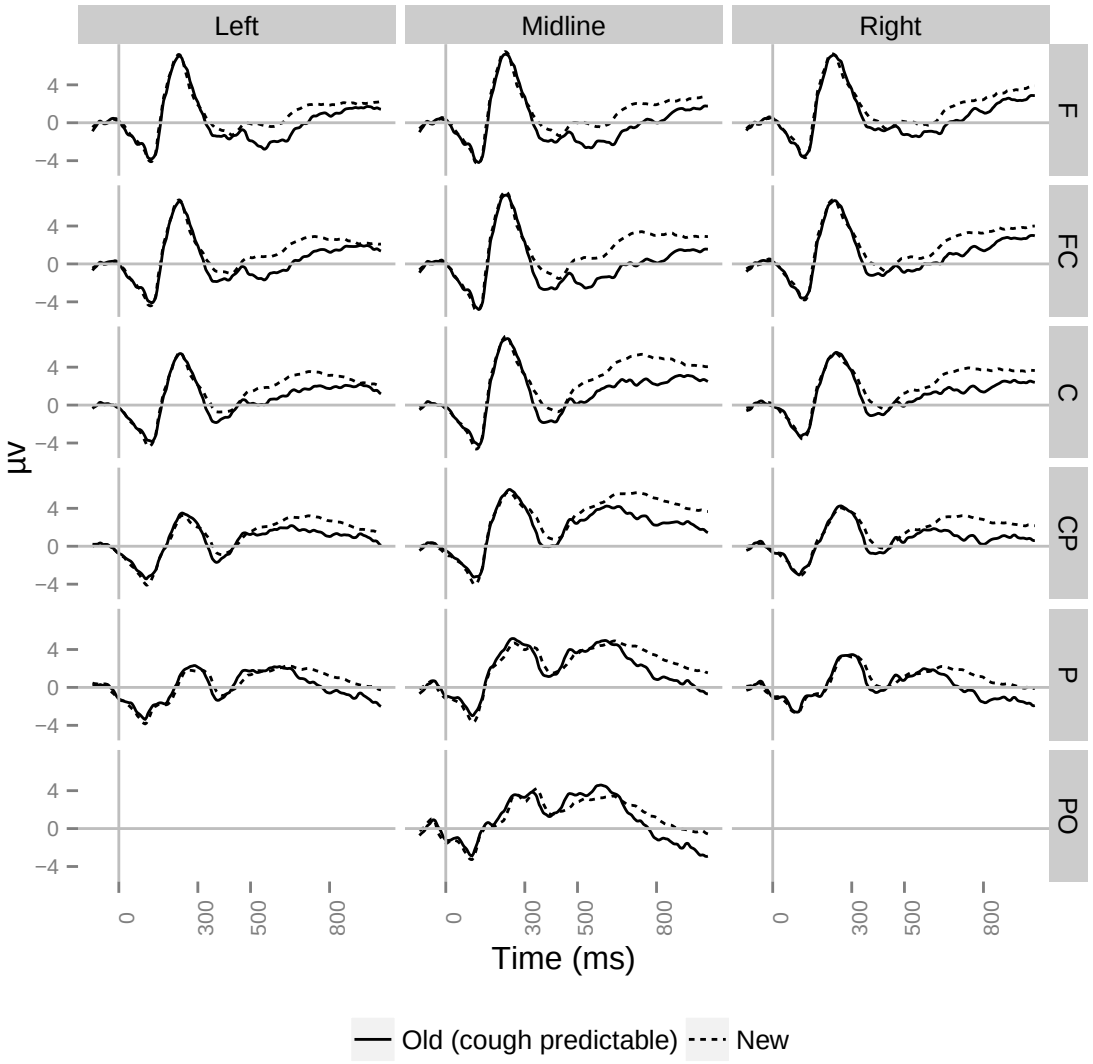

FIGURE 8.16: Grand average ERPs (n=21) for targets which had been predictable and interrupted by a cough at original presentation, as well as new targets. Shown here are ERPs as measured at frontal (F), fronto-central (FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline.

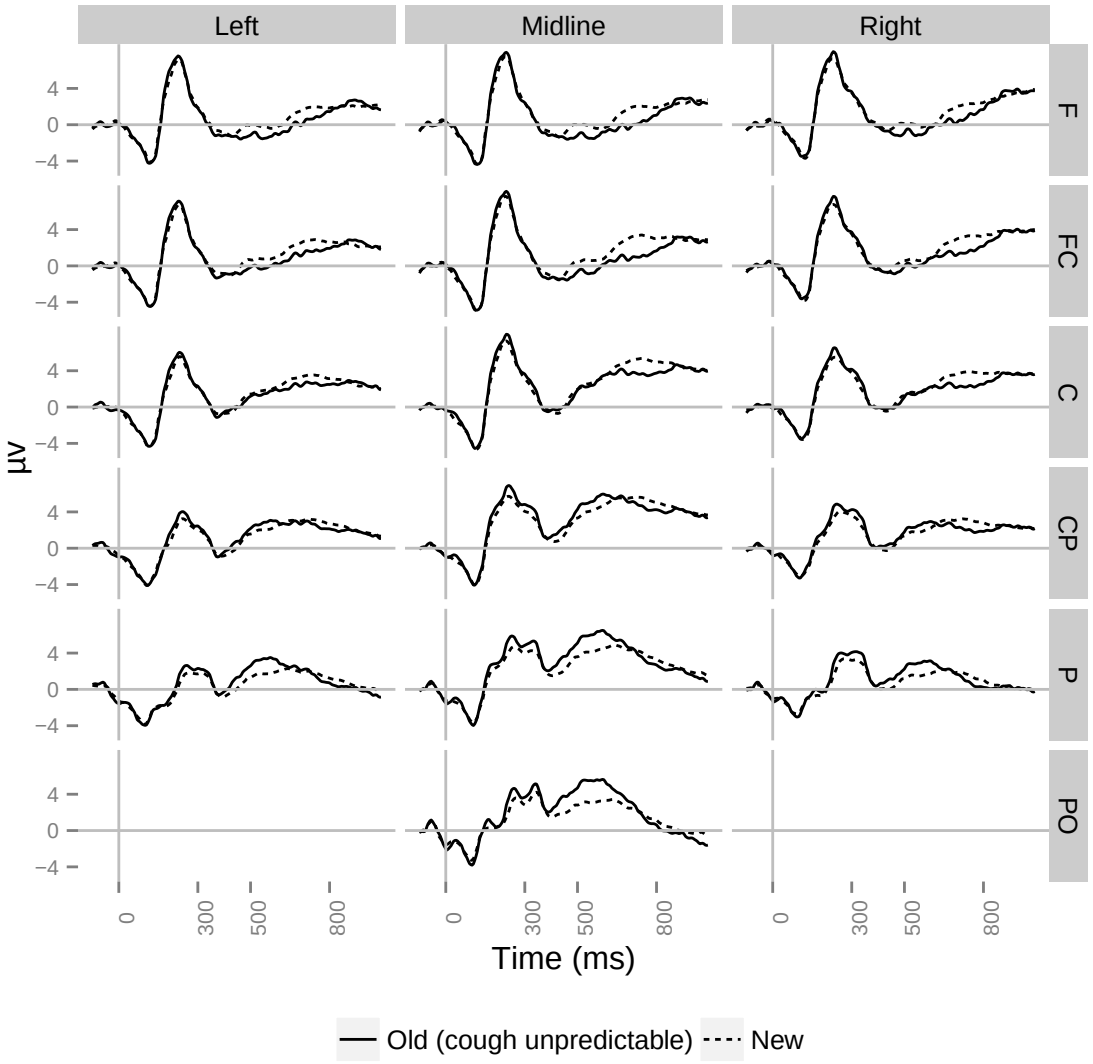

FIGURE 8.17: Grand average ERPs (n=21) for targets which had been unpredictable and interrupted by a cough at original presentation, as well as new targets. Shown here are ERPs as measured at frontal (F), fronto-central (FC), central (C), centro-parietal (CP), parietal (P) and occipito-parietal (PO) locations, for electrodes grouped over left (electrodes 1,3,5) and right (electrodes 2,4,6) hemispheres, and the midline.

Comparing Fillers to Coughs — 300-500ms

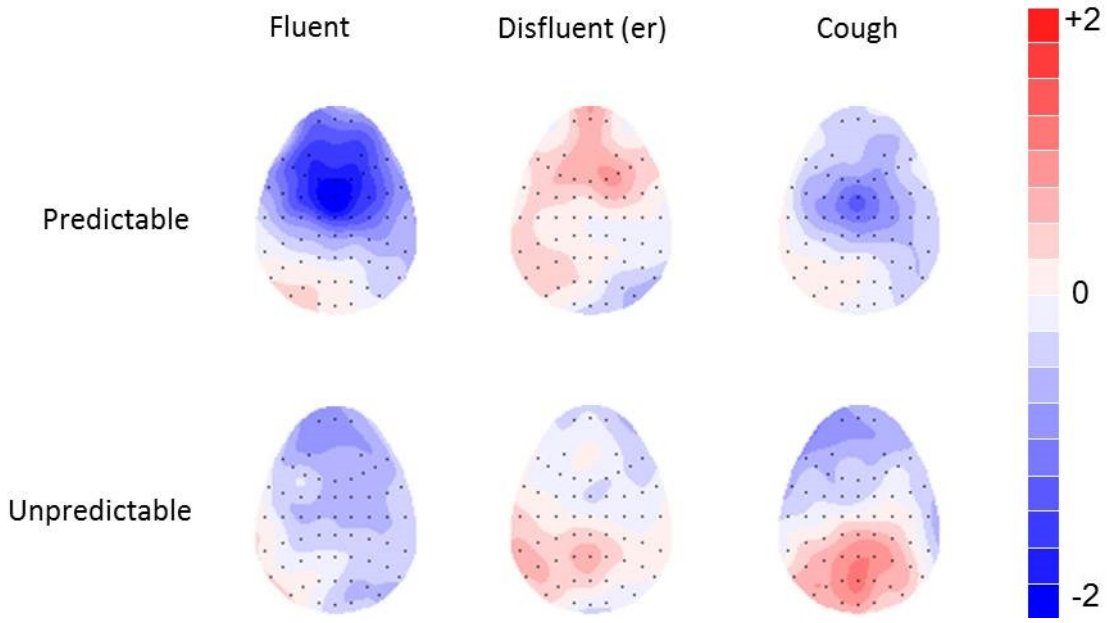

FIGURE 8.18: Retrieval effects (hits - correct rejections) in the 300-500ms time window. Whilst there appears to be something akin to a mid-frontal old/new effect for predictable disfluent words, this is not seen for any of the other conditions, which show a general relative negativity for old words. Unpredictable words which had been preceded by a cough show a parietal positvity, which may be interpretable as a repetition effect.

In the 300-500ms time window, where one might typically expect to see mid-frontal positivity for old words, only one condition appeared to elicit an effect which could be interpreted in this way. Items which had been predictable and disfluent at initial presentation appeared to elicit a relative positivity maximal around F2, and spreading backwards over the left of the scalp (see Figure 8.18). For these predictable, disfluent items, ANOVA with factors of condition *(old, new)*, location *(F, FC, C, CP, P)* and hemisphere *(left, right)* revealed a significant interaction of condition with location and hemisphere [$F(1.64,32.75) = 4.18$, $\eta^2_G = 0.0002$, $p < 0.05$], reflecting the fact that old words elicited ERPs more positive at the front of the scalp in the right hemisphere, whereas in the left, the difference between frontal and posterior locations was not so strong.

Each of the other fluency and predictability conditions was submitted to the same ANOVA. For items which had been fluent and predictable, this analysis revealed a significant main effect of condition [$F(1,20) = 6.49$, $\eta^2_G = 0.013$, $p < 0.05$], reflecting the overall relative negativity elicited by old words, an interaction of condition with location [$F(1.15,22.92) = 8.69$, $\eta^2_G = 0.004$, $p < 0.01$], reflecting a weak parietal positivity for old words. No significant main effects or interactions were found for words which had been predictable and preceded by a cough.

For items which had been unpredictable and preceded by a cough, the multilevel ANOVA revealed a significant interaction of condition with location [$F(1.19,23.86) = 8.97$, $\eta^2_G = 0.003$, $p < 0.005$]. This interaction reflects the clear parietal positivity elicited by old words in this condition. It may be possible to interpret this parietal positivity as a repetition effect (Rugg et al., 1998; Bridson et al., 2006), although why this should be found for only one condition is unclear. ERPs to unpredictable words which had been fluent or disfluent did not show any significant main effects of interactions implicating condition.

As there is no clear evidence of mid-frontal old/new effects, or repetition effects across the fluency and predictability conditions, no further analysis is reported for the 300-500ms latency range.

Comparing Fillers to Coughs — 500-800ms

In the 500-800ms latency range, the widespread relative negativity for old items persisted, and visual inspection of the data shows the negativity to have increased in amplitude, compared to the earler 300-500ms range previously analysed. This is the latency range in which one might expect to find LPONE, and all six experimental conditions

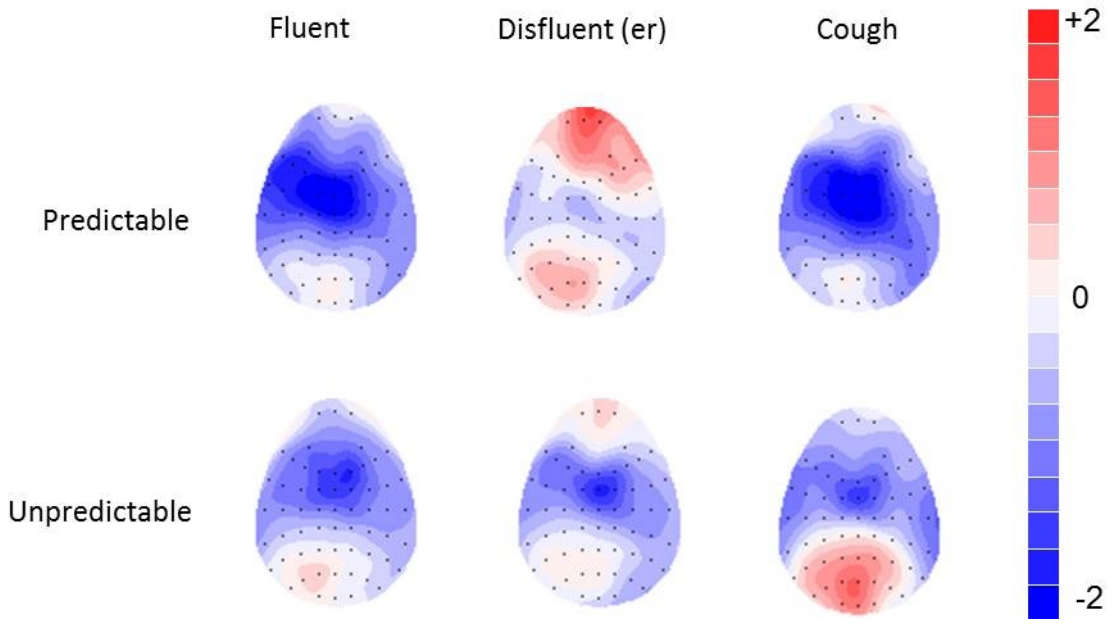

FIGURE 8.19: Retrieval effects (hits - correct rejections) in the 500-800ms time window. Across all six conditions, left parietal sites appear to show a greater relative positivity. This is most marked for unpredictable words which had been preceded by a cough. Predictable disfluent words also show a frontal positivity to old words.

show a less negative ERP towards posterior electrodes, although any positivity seems to be masked by the broadly spread relative negativity for old words.

In order to establish whether any of the fluency and predictablity conditions elicited ERPs consistent with an LPONE, the data from each condition were submitted to multilevel ANOVA with factors of condition *(old, new)*, location *(F, FC, C, CP, P)* and hemisphere *(left, right)*. For items that had been predictable at auditory presentation, fluent items elicited a main effect of condition [(F1,20) = 7.34, $\eta^2_G = 0.004$, p <0.05] and an interaction of condition with location [F(1.25,15.07) = 8.60, $\eta^2_G = 0.006$, p <0.005], reflecting a general relative negativity to old words which is stronger at mid frontal than posterior locations. Predictable disfluent words elicited a significant interaction of condition with location and hemisphere [F(1.42,28.30) = 4.87, $\eta^2_G = 0.0008$, p <0.05], reflecting the fact that frontal and parietal locations showed more positive ERPs for old words, and that at frontal locations this positivity had a right hemisphere bias, whereas

at posterior locations, the positivity was greater in the left hemisphere. Predictable words which had been preceded by a cough resulted in a significant effect of condition [F(1,20) = 11.47, η^2_G = 0.030, p <0.005], a significant interaction of condition with location [F(1.39,27.70) = 4.66, η^2_G = 0.002, p <0.05], and a marginally significant interaction of condition with location and hemisphere [F(1.52,30.42) = 2.91, η^2_G = 0.0005, p <0.1]. This again reflects a broadly spread relative negativity for old words, which is greater at fronto-central locations than towards the rear of the scalp. The marginally significant three way interaction reflects the fact that when each hemisphere is considered separately, the interaction of condition with location is significant in the left hemisphere, but not in the right.

For words which had been unpredictable in their contexts at auditory presentation, all three fluency conditions elicited significant interactions of condition with location; fluent [F(1.51,30.29) = 11.18, η^2_G = 0.003, p <0.001]; er [F(1.30,26.00) = 5.28, η^2_G = 0.002, p <0.05]; cough [F(1.22,24.41) = 11.65, η^2_G = 0.006, p <0.005]. This reflects the fact that there was a gradient across the scalp for old words, with frontal locations showing a relative negativity, while posterior locations showed a small positivity. Items from fluent utterances also elicited a significant main effect of condition [F(1,20) = 4.53, η^2_G = 0.013, p <0.05], reflecting the general negativity elicited by old words.

Topographic Comparison

Given that a relative parietal positivity to old items is seen to a greater or lesser extent across all six fluency and predictability conditions, the rescaled difference waveforms are compared topographically, in order to establish whether they should be considered to represent different sets of neural generators. A multilevel ANOVA incorporating factors of predictability *(predictable, unpredictable)*, fluency *(fluent, er, cough)*, location

(F, FC, C, CP, P) and hemisphere *(left, right)* reveals a significant interaction of predictability with location and hemisphere [F(1.43,28.72) = 3.75, η^2_G = 0.0003, p <0.05]. Given this difference in scalp topography across fluency and predictablity conditions, it is not possible to quantitatively compare effects in the 500-800ms latency range accross all six fluency and predictability conditions. However, when only ERPs to words which had been unpredictable at their initial presentation are considered, no significant or marginally significant effects implicating topographic differences between fluency conditions are found. This was the case for both global and midline ANOVA, and so the old/new effects for unpredictable words were quantitatively compared across fluency conditions.

Quantitative Comparison

This comparison drew on data from the difference waveforms (old - new) for unpredictable words, and employed a multilevel ANOVA with factors of fluency *(fluent, er, beep)*, location *(F, FC, C, CP, P)* and hemisphere *(left, right)*. No significant differences were found between fluency conditions, leading to the conclusion that the recognition effects reported for unpredictable words do not vary with fluency.

8.4.3 Comparing Fillers to Coughs — Summary of Memory Results

The memory performance results for data from the cough context replicate those reported for data from the beep context. In accordance with expectations, participants were more successful at recognising words which had been unpredictable in their original contexts. Both fillers and a cough improved memory performance compared to fluent utterances, but there was no interaction between predictability and fluency. As in the

previous experiment, however, visual inspection of the results (as seen in Figure 8.11) does seem to indicate a slightly larger difference in performance between fluent and disfluent items for words which had been unpredictable than predictable.

As with the previously reported experiment, the memory performance outcome reported here replicates previous studies which reported increased recognition rates for items which had been preceded by a filler (Corley et al., 2007) or a pause (MacGregor et al., 2010) at initial auditory presentation. Further, the data presented here do not provide any evidence for a difference in memorability between items preceded by a filler and those preceded by a cough. In the context of this experiment, a cough adds delay to the utterance without connotations of disfluency — it does not necessarily indicate that the speaker is experiencing linguistic difficulty. Unlike the beep used in the previously described experiment, a cough is ecologically valid, and should constitute a plausible explanation for the temporary cessation of speech. The results of this study contrast with those of Fraundorf and Watson (2011), who found coughs to be of no benefit to participants attempting to retell a story. Indeed, they reported that the presence of a cough actually reduced the likelihood of a plot point being recalled at a subsequent retelling, and concluded that whilst fillers were beneficial to participants' memory, coughs acted as a distraction, reducing memory performance.

It is not possible to directly contrast the results reported by Fraundorf and Watson (2011) with those of this study, given the significant difference in tasks. Whereas in the experiment reported here, the memory test came as a surprise, in order to prevent participants using memory strategies during the auditory section of the experiment, the participants in Fraundorf and Watson's experiment knew they would be retelling the story they heard, and so were presumably making an effort to remember each plot point

as they heard them. In addition, it is unclear how the effects of coughs on single item recognition should map onto the effects of coughs on holistic, discourse level memory.

The ERP data gathered in this study are again somewhat inconclusive. There are neither clear mid-frontal nor clear left-parietal old/new effects, despite the evidence that participants were attending to the task and were successful in recognising previously heard items. The relatively high false alarm rate may offer some explanation of the failure to find recognisable retrieval ERP effects; if participants were incorrectly identifying 43% of new words as old, this suggests that rather a lot of their answers may have been guesses. If a large proportion of the correctly identified old trials, or "hits" are really made up of guesses, without recognition, then that should dilute any recognition ERP effects, possibly to the point of being unidentifiable. It also seems possible that left-parietal old/new effects may be present in the data reported here, but are masked by the large centro-frontal relative negativity to old words, the cause of which is unclear. Perhaps the most convincing LPONE type ERP is that seen in response to words which had been unpredictable and interruped by a cough at their initial presentation. Given, however, that a significant parietal positivity is also seen in the earlier 300-500ms latency range for these words, it is unclear whether the positivity seen between 500ms and 800ms should be interpreted as a LPONE, or a continuation of a repetition effect.

8.5 Comparing Fillers, Beeps and Coughs

8.5.1 Comparing Fillers, Beeps and Coughs — Memory Performance

This section reports the memory outcomes of the experiment reported in Chapter 6.4.4, in which participants were exposed to two listening blocks, one contrasting fillers with

block	fluency	predictable	unpredictable
beeps	fluent	64%	71%
beeps	disfluent (*er*)	66%	72%
beeps	beep	66%	70%
coughs	fluent	63%	72%
coughs	disfluent (*er*)	70%	73%
coughs	cough	65%	73%
	new	58%	

TABLE 8.9: Memory performance comparing fluent utterances with utterances interrupted by fillers, beeps and coughs. The table shows the mean probability of previously heard target words being correctly identified (n=24).

beeps and one contrasting fillers with coughs, before completing the surprise memory test. This added a dimension to the experimental paradigm used in the two previous experiments; one of global context. Fluent and interrupted items had been heard in the context of a speaker with a cold (coughs), or artificially interrupted speech (beeps). For a fuller explanation of the experimental paradigm, see Chapter 7.

Participants were more accurate in identifying words that had been presented in unpredictable than predictable contexts. Overall, participants correctly identified 69% of old items and 55% of new items (45% false alarm rate). To assess whether these numeric differences were reliable, predictability data was assessed using a multilevel ANOVA with factors of predictability, fluency and context, and using stimulus as a random factor. As there were 12 conditions and 24 participants, each individual stimulus was seen by only two participants in each condition. In the event that any stimulus did not provoke a timely response from either participant in any one condition, all incidences of that stimulus were removed from the data. In total, 16 stimuli were removed, leaving 308 individual stimuli. The ANOVA revealed a significant main effect of predictability [$F(1,307) = 28.6$, $p < 0.00001$]. No other main effects or interactions reached significance. Mean accuracy rates for each condition can be seen in Table 8.9.

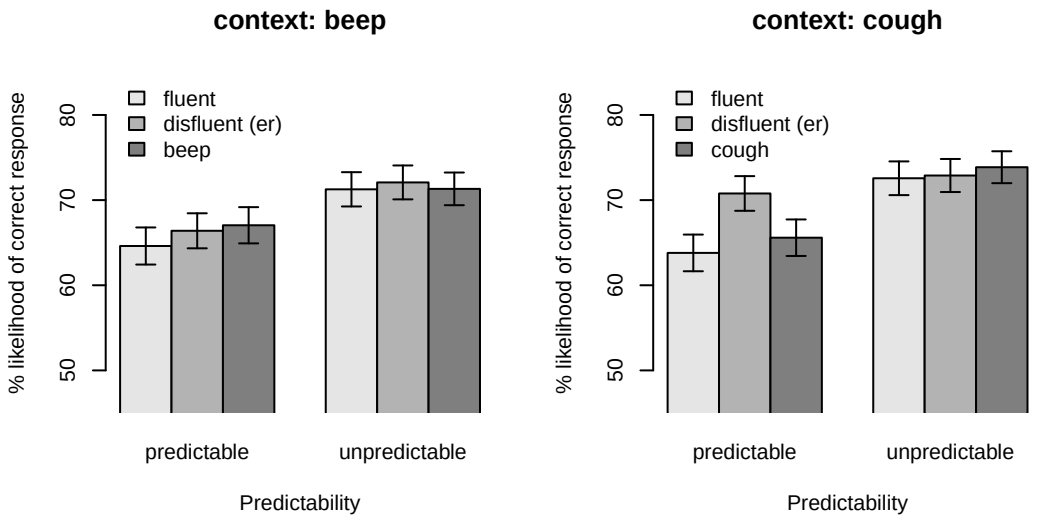

FIGURE 8.20: Probability of participants correctly identifying target words as old averaged across stimuli. Error bars represent one standard error of the mean. Unpredictable words were significantly more likely to be correctly identified than predictable words in both the beep and the cough contexts. There were no significant main effects or interactions involving fluency or context.

The effects of fluency, predictability and context on reaction time and confidence were assessed using multilevel ANOVAs with factors of predictability, fluency and context. Where any stimulus failed to produce at least one correct response in each fluency, predictability and context condition, all incidences of this stimulus were removed, resulting in 72 distinct targets. The ANOVAs revealed no significant main effects or interactions affecting either participants' reaction times to correctly identified old targets, or their confidence in memory judgements. Mean reaction times and confidence judgements for correct responses in each condition can be seen in Tables 8.10 and 8.11.

8.5.2 Comparing Fillers, Beeps and Coughs — Summary of Memory Results

Participants in this experiment listened to fluent, disfluent and interrupted sentences with predictable and unpredictable final words. Interruptions took the form of either

block	fluency	predictable	unpredictable
beeps	fluent	999	1007
beeps	disfluent (*er*)	1007	1004
beeps	beep	998	997
coughs	fluent	1018	966
coughs	disfluent (*er*)	1018	1007
coughs	cough	1029	982
	new	1168	

TABLE 8.10: Mean reaction times (ms) to correctly identified old words comparing words from fluent utterances with those from utterances interrupted by fillers, coughs and beeps. In the coughs block, unpredictable words elicit faster reaction times than predictable words. In the beeps section, reaction time is not significantly influenced by predictability or fluency.

block	fluency	predictable	unpredictable
beeps	fluent	3.05	3.00
beeps	disfluent (*er*)	2.96	2.88
beeps	beep	3.06	3.06
coughs	fluent	2.96	3.01
coughs	disfluent (*er*)	2.99	3.06
coughs	cough	2.98	2.91
	new	3.03	

TABLE 8.11: Mean confidence levels for correctly identified old words comparing words from fluent utterances with those from utterances interrupted by fillers and coughs. Upon identifying target words as old or new, participants rated their own confidence in the answer they had just given using a five point scale, '5' representing "very confident", and '1' representing "very unsure". Mean confidence judgements did not vary with predictability or fluency in the coughs block of the experiment. In the beeps section, there was a significant effect of fluency.

an artificial beep, or a speaker generated cough. Both types of interruption were time-matched to the disfluent filler in the equivalent disfluent sentence. The materials were presented in two blocks, one using each type of interruption. Following presentation of all the auditory materials, participants performed a surprise memory test, in which words which had been the final words in the auditory sentences were visually presented, intermixed with new words. In the memory test, all the old target words were intermixed. Items which had originally been preceded by a beep or a cough were not separated.

The results of this experiment show a significant effect of predictability, with participants more successful at identifying words which had formed unpredictable endings to sentences. No effects of fluency or context on accuracy were found. Reaction times and participants' self assessed confidence in their responses did not vary significantly with predictability, fluency or context. Inspection of the graph in Figure 8.20 (left) suggests a small numeric increase in accuracy for disfluent and interrupted predictable words which were first presented in the context of a beep type interruption. This pattern is very similar to that seen in the first two experiments (see Figures 8.1 and 8.11), although analysis of data from the "beep block" alone also fails to reveal a significant effect of fluency on accuracy rates. Examination of memory performance in the "coughs block" (Figure 8.20, right) reveals a slightly different numeric pattern. Whilst unpredictable words show very little variation across fluency conditions, disfluent predictable words appear to have been more accurately recalled than fluent predictable words, but words preceded by a cough did not appear to enjoy the same memory benefit. Analysis of the data from the "cough block" only did not, however, reveal any main effects or interactions implicating fluency.

It is possible that this experiment has failed to reveal fluency effects on memory as a result of the experiment's relatively low trial numbers and noise in the data. An alternative interpretation might posit that participants are distracted by the change in stimuli between blocks, rendering fluency in some way less significant at the point of recognition. Overall accuracy does not seem to be diminished by the extra level in this experiment. The overall hit rate (69%) and correct rejection rate (58%) are very similar to those the first two experiments reported in this chapter (Beeps: hit rate 67%, correct rejections 61%, Coughs: hit rate 70%, correct rejections 58%).

No ERP results have been presented for this memory test. A relatively low number of

trials presented per condition (27) meant that once trials with incorrect responses and those affected by artifacts and drift in the EEG signal were excluded, very few remained for analysis. Only three participants passed the criterion of providing a minimum of sixteen useable trials per condition, making it impractical to analyse these data using traditional ANOVA analysis. Low trial numbers are not an uncommon problem in ERP experiments, and an alternative approach is explored in the next chapter.

8.6 Chapter Summary

This chapter has reported the memory outcomes from three experiments, in which participants were required to make old/new judgements to visually presented words. Half of the words had been previously encountered in a listening task, in which participants heard sentences ending predictably and unpredictably and in which were either fluent, or contained a disfluent hesitation or interruption before the final, target word. The final word of the auditory sentences constituted the target word for the purpose of these experiments, and so it was these final words which were presented as part of the memory test. In the first experiment, utterances were presented in three fluency conditions: fluent, disfluent (er) and interrupted by a beep. The second experiment built on this by replacing the beep interruption with a cough, to investigate whether effects observed in the first experiment remained robust when the interruption was speaker generated. In the third experiment reported, an extra factor is added, and the paradigms from experiments 1 and 2 are effectively combined, in order to investigate whether global context affects participants interpretation and use of disfluency.

Experiment		predictability	fluency	context	predictability:fluency	predictability:context	fluency:context	predictability:fluency:context
Exp1: Beeps	Accuracy	*	*	N/A		N/A	N/A	N/A
	RT	*		N/A		N/A	N/A	
	Confidence	*		N/A		N/A	N/A	
Exp 2: Coughs	Accuracy	*	*	N/A		N/A	N/A	N/A
	RT			N/A		N/A	N/A	N/A
	Confidence	*	*	N/A		N/A	N/A	N/A
Exp 3: Context	Accuracy	*						
	RT							
	Confidence							

TABLE 8.12: Table summarising significant ($p < 0.05$) outcomes of behavioural analyses. Data from the first two experiments (comparing fillers to fluent items and beeps, and comparing fillers to fluent items and coughs) were analysed using multilevel ANOVAs with factors of predictability *(predictable, unpredictable)* and fluency *(fluent, er, beep/cough)*. For data from the third experiment, a factor of context *(beep context, cough context)* was added. Analysis was carried out for three dependent variables; Accuracy — the likelihood of previously heard items being correctly identified as 'old', reaction time — the time between the appearance of a previously heard target word on the screen and participants' button press indicating that they correctly identified the word as 'old', and confidence — participants' self rated confidence in their answer (5 point scale) where they correctly identified a word as 'old'.

Memory Performance Summary

In both of the first two experiments reported, participants were significantly more accurate in recognising words which had been unpredictable at presentation, and more accurate in recognising words which had been interrupted by either a disfluent filler or a non-linguistic interruption (a cough or a beep). In both of these experiments, there was no difference between the memorability of words which had been preceded by a filler and those which had been interrupted with a cough or beep. In the first experiment, greater accuracy for unpredictable words was accompanied by faster reactions and greater confidence in response to these targets. In the second experiment, reaction time did not vary, but participants' confidence was higher in response to unpredictable words. Additionally, participants were slightly more confident in their recognition of words which had been interrupted by a cough than those which had been fluent, although there was not a significant difference between confidence levels between interrupted word and disfluent words, or between fluent words and disfluent words. The third experiment showed a significant effect of predictability on accuracy — participants were more likely to recognise unpredictable words, but unlike the first two experiments, fluency had no significant effect on accuracy, confidence or reaction time.

The first two studies appear to replicate previous findings using a similar paradigm, which reported increased recognition accuracy for unpredictable words, and greater accuracy for words which had been preceded by a disfluent filler (Corley et al., 2007) or a silent pause (MacGregor et al., 2010). The data from the experiments presented in this chapter do not show the interaction between predictability and fluency reported by Collard et al. (2008), who found no predictability effect after words which had been acoustically deviant at initial presentation. The finding that both coughs and beeps

improved memory just as much as disfluent fillers suggests that the delay introduced by these was sufficient to raise attention enough to cause a lasting effect — improving memory for items up to two hours after initial presentation.

Corley et al. (2007) reported that memory performance benefits from disfluency, and that the N400 effect decreases following a disfluent filler. They suggest that the presence of a disfluency makes predictable word more difficult to integrate because listeners are predicting unexpected word. According to this logic, disfluency leads to more effortful processing, particularly of predictable words, and this increased processing effort results in a long term memory benefit. In addition, MacGregor et al. (2009) found no reduced N400 and no memory effects for words which were affected by repetition disfluencies. They suggest this demonstrates that the N400 and memory benefits pattern together.

In contrast to MacGregor et al.'s (2009) suggestion, the first two experiments reported in this chapter revealed memory benefits for unpredictable and interrupted targets, even in the absence of clear N400 effects. In experiment 1, the N400 effect was slightly reduced by the presence of a disfluent filler *(er)*, but no difference in N400 amplitude was detected between words which had been fluent and those preceded by a beep (see Figure5.6). Despite this, the behavioural evidence presented in Section 8.3.1 shows a clear memory benefit for words which had been preceded by either a filler or a beep, and no difference between the two. In experiment 2, the N400 effect did not vary significantly with fluency (see Figure 6.6), but once again, a clear recognition memory advantage emerged for words affected by either a disfluent filler or a cough, and no difference was found between the two (see Section 8.4.1). This may reflect a dissociation between the processes underlying the reduction of the N400 following disfluency and the processes underlying improved memory for affected items. Alternatively, it may reflect the difficulty of reliably capturing differences in N400 effect size in auditory presentation

experiments, in which the data is made noisier by the unfolding over time of the target words, and more complex processing nature of sentence based stimuli. If this is the case, and the fluency effect on N400 effect sizes is not robust, or is hard to detect, then the failure of MacGregor et al. (2009) to find fluency effects on the N400 must also be viewed in this light, which means that the possibility has not been ruled out of a dissociation between the immediate processing effects (as indexed by the N400) and lasting memory effects of disfluency.

ERP Results Summary

ERP results are reported from the first two experiments presented in this thesis; one experiment comparing the impact of fillers and beeps and one comparing the impact of fillers and speaker-generated coughs. A third experiment, in which a third dimension was added to the experimental paradigm, did not produce enough full datasets (in which each participant provides 16 usable trials in each condition) to allow meaningful analysis.

ERP data were analysed for the presence of two old/new effects. The first of these was the mid-frontal old/new effect, sometimes referred to as the FN400, maximal at mid-frontal locations in the 300-500ms latency range, and widely believed to index familiarity based retrieval (Nessler et al., 2001; Curran, 2000; Curran, Tepe, & Piatt, 2006). The second old/new effect considered was the left-parietal old/new effect (LPONE). This relative positivity to previously studied words is typically maximal at left parietal locations between 500ms and 800ms, and is thought to indicate recollection.

Neither of the two experiments reported in this chapter provided evidence of mid-frontal effects. Inspection of the data in the 300-500ms latency range showed a parietal positivity to old words in some conditions, which may be interpreted as a repetition effect (Rugg

et al., 1998; Bridson et al., 2006), but analysis of this effect is hampered by trial number limitations. If this observed positivity is a repetition effect, then it should not predict memory performance, and should be present when studied words incorrectly identified as new (misses) are compared with unstudied words. Trial number limitations mean that very few of the subjects who provided sixteen usable hits in every condition also provided sixteen usable misses. This means that whilst it is possible to observe a possible repetition effect, it is not possible to comment further on whether it is indeed a repetition effect and whether it is in any way modulated by the condition of the stimulus at presentation.

Visual inspection of the data from both of the experiments reported here shows a slight positivity at left-parietal electrode sites for previously studied words, although the ERPs continued to be dominated by a relative negativity for studied words, which onset between 300 and 400ms at anterior right hemisphere electrode sites, spreading to reach posterior left hemisphere sites by 500-700ms after stimulus onset. The reason for this unexpected negativity for old items is unclear, but it may go some way to explaining the lack of clear LPONE, as they are masked by the broad spread of the negative component. Visual inspection of data from the first experiment suggested a weak left parietal positivity for old words in all conditions, but ANOVA revealed a that the apparent difference between anterior and posterior locations was only significant for words which had been preceded by a beep. Differences in topographic distribution were also found for the old/new effects elicited by the six fluency and predictability conditions, although when predictable data only were considered, no significant topographic difference emerged. Quantitative comparison of recognition effects for predictable words revealed no differences between fluency conditions. Analysis of ERPs from the second experiment (comparing fillers with coughs) revealed significant interactions of fluency

with location for all six fluency and predictability conditions. Differences in scalp topography meant that a magnitude comparison of effects across all six conditions was not possible, but across fluency conditions for unpredictable words (where topography did not vary significantly), no evidence was found for any variation in the magnitude of the LPONE with fluency.

The negative shift to old words does not seem to be a side effect of changing modality; from aural presentation at study, to visual presentation at test. Although Curran, Schachter, Johnson and Spinks (2001) found no FN400 effects in a visual memory test for words which had been aurally presented, Nessler et al. (2001) did report FN400 effects in a similar paradigm. In an experiment combining visual and auditory stimuli, followed by visual presentation at test, Joyce, Paller, Schwartz and Kutas (1999) do report ERP differences at retrieval between items which had been presented visually and auditory stimuli, but this does not extend to a reversal of the typical old/new effect, in which old words usually produce more positive ERPs than correct rejections. Comparison of ERPs from other cross modal auditory study / visual test experiments shows a uniform pattern of more positive ERPs to hits than correct rejections (Curran & Dien, 2003; Wilding & Rugg, 1997; Wilding et al., 1995). Thus the cross-modal aspect of these studies cannot explain the whole-scalp negativity seen for old words.

It seems plausible that the long-lasting positive shift observed for words attracting a correct "new" response could be an artifact of response preparation or decision making. Responses to new targets are slower (see Tables 8.2, 8.6 and 8.10). If there is a readiness potential or similar, time-locked to a response, then the difference in the time-courses of responses between new and old words would lead to a difference in amplitude between the waveforms. Indeed, inspection of the waveforms presented earlier in this chapter does seem to suggest a lag for ERPs to new words. It is possible that this difference

could partially be the result of Lateralised Readiness Potentials (LRPs). The offset in negative shifts associated with response preparation would result in an apparent positive shift for the condition with slower responses (new items), which may account for the apparent negativity to old items. This effect would be largest in the contra-lateral hemisphere to the response hand; in this experiment, participants' response hands were counterbalanced so that when all the data are taken into account, no overall hemisphere bias should be apparent, whereas the negativity reported in these experiments is slightly greater in the right, casting some doubt over whether LRPs can fully account for the negativity observed. The Late Posterior Negativity (LPN) often reported for old compared to new words also seems unlikely to be responsible for the outcome of this experiment. This is is typically observed later in the epoch, and has a posterior distribution.

Taken together, the ERP results from these two experiments appear to suggest that left-parietal old/new effects might be expected, were it not for the fronto-central negativity seen in the ERP. This might be avoided were the experiments to be repeated with a forced delay before participants were able to make their response to the old/new judgement. Relatively low trial numbers (see Tables 8.4 and 8.8) almost certainly contribute to a higher than desirable noise to signal ratio in the data, but increasing trial numbers is somewhat problematic. The experiments described here had a testing duration of approximately two hours in addition to set-up, instructing and debrief time. To increase trial numbers meaningfully would not only have ethical and practical implications in terms of the amount of time demanded of voluntary participants, but may also lead to a decrease in accuracy, resulting in many more lost trials. In debriefing, participants reported finding the surprise memory test difficult, and this is substantiated by the relatively low discrimination rates (17.5% in the beeps experiment; 28% in the coughs experiment). It is possible that increasing the length of the experiment would lead to

more guessing by participants, diluting recognition ERPs still further.

A different analysis approach may be able to shed some light on these data. Like many ERP experiments which suffer from trial number problems, these experiments are rather inefficient in terms of data. In order to avoid problems associated with unbalanced data, any participant who does not provide sixteen usable epochs to all seven of the testing conditions (six old conditions, one new) is rejected from analysis. This results in analysis incorporating a rather smaller than ideal pool of participants, whilst many hours worth of EEG data are discarded. One approach which may go some way to deal with this problem is mixed-effects modelling. This alternative to a standard ANOVA approach to ERPs may allow the analysis of unbalanced datasets, opening the door to complex multilevel experiments such as these described here without the need for excessive trial numbers, and is explored in the next chapter.

Chapter 9

Mixed Effects Models — An Alternative Approach to ERP Analysis

9.1 Introduction

In the previous chapter I alluded to the possibility of increasing the interpretability of ERP data with a change in statistical approach. In this chapter, I introduce the linear mixed-effects regression modelling technique for statistical analysis as a potential tool for ERP research. Much ERP research is designed around ANOVA analysis, and the newer and more flexible analysis structure allowed by linear mixed effects modelling has not been widely adopted in the field. This type of modelling is, however, gaining traction within psychology generally, and discussion surrounding the way psycholinguistic data should be analysed has been ongoing for some time.

9.1.1 Accounting for the Effects of Stimuli

One of the problems psycholinguistic researchers face is deciding how to deal with effects of the stimuli they choose. Stimuli in psycholinguistic experiments are typically words or groups of words, often presented within certain manipulations. However, every word has its own inherent properties, including variation in concreteness, imagability, length, frequency, memorability etc., and the individual words selected for psycho-linguistic studies have some bearing on the pattern of results. Responses in experiments are likely to show some similarity across participants to any one stimulus word compared to other stimulus words, much the same as all of the responses of one participant, presented with a set of stimuli, are likely to be more similar to one another than the responses of another participant, exposed to the same stimuli. Despite this, many psycholinguistic experiments tend to gloss over differences between stimuli, and assume that all stimuli within a category behave in the same way.

Coleman (1964) pointed out that linguistic materials should really be treated as random effects, as they are drawn from a much larger set of possible linguistic materials. The sometimes conceptually difficult distinction between fixed and random effects can be clarified by imagining repeating an experiment. If, for example, one of the experiments in this thesis were to be repeated exactly, using the same stimulus recordings, then the experimental conditions (fluency, predictability) would be the same; these are fixed effects. The individual participants, however, who are selected randomly from a large pool of possible participants, would differ. Participant identity, while not the focus of the experiments, nonetheless has an effect on data outcomes.

In psychological experiments, including ERP experiments, it is very common to use subject as a random factor in ANOVA analysis. This allows the intercept to vary

for each participant. For example, in an experiment investigating whether students' likelihood of submitting an essay on time was affected by whether or not they had been invited to a party the night before, including random intercepts for participants would allow the researcher to take into account that some students are generally more likely to submit on time than others. Essentially, we are asking: "Does y change as a result of x, given subject variation?" The necessity of having random intercepts for subjects is particularly clear when considering ERP experiments, where physiological differences between participants can have a significant impact on the measured data. In the same way as treating participants as random variables allows us to establish the likelihood of any reported effects being replicated were the experiment carried out again on a separate group of participants, treating items as random effects would allow us to to predict the likelihood of effects being repeated were the experiment repeated with a different sample of language.

A technical solution to the problem of incorporating random effects for both subjects and items into F-statistic based analysis (such as ANOVA) was proposed by Clark (1973), using a quasi-F statistic, derived from the minimum-F statistics from separate by-participants and by-items analyses. Although Clark's solution has been part of the psycholinguistic literature for over forty years, it has not been widely taken up for use in ERP analyses, partly because of the constraints it would place on experimental design size.

An alternative solution to the language-as-fixed-effect problem, not requiring separate by-participants and by-items analyses has emerged in the form of linear mixed-effects regression modelling. Linear mixed-effects models allow the incorporation random effects for both participants and simultaneously. As off-the-shelf implementations for these models have become available, they have begun to gain traction in psycho-linguistics. For

example, in 2008, a special issue of the *Journal of Memory and Language* was dedicated to emerging data techniques, and within this issue were included three separate articles on mixed-effects models.

9.1.2 What are Linear Mixed-Effects Models?

Linear mixed-effects models are extensions of linear regression models. Responses are simulated as a function of fixed effects, random effects, and a noise term. Random effects can include both random intercepts and random slopes and interactions. In terms of the type of language experiment discussed thus far, this means not only that we can incorporate random intercepts for the identity of the participant and the stimulus, but also model the sensitivity of the participant or stimulus to the other predictors in the model.

Independent (predictor) variables can be grouped, which means that assumptions of independence and homogeneity are not necessary. Clustering of response variables, for example the tendency for all responses from one participant, or all responses to a stimulus item to display some similarity, can be incorporated into the model. When all participants encounter all the stimulus words, the random effects are fully crossed. Alternative, nested models are also possible, in which groups of participants are exposed to different groups of words. Both of these designs can be modelled within the random effects structure of mixed-effects models, relieving the need for multiple F-tests, and increasing the range of experimental designs possible.

Mixed-effects models allow the inclusion of factorial and linear fixed and random effects, and as the name suggests, these can all be incorporated within one model to best describe the data. Whilst Clark's (1973) solution allows an estimated F which

guards against over-confidence in the significance of effects, it does not allow inclusion of multiple random effects, nor modelling of interactions between them. In contrast, mixed-effects modelling presents an alternative which not only prevents over-confidence by allowing incorporation of both participants and items as random effects, but also allows researchers to include interactions between multiple random effects. Moreover, whereas Clark's solution only allows the calculation of significance, mixed-effects modelling also allows the calculation of coefficient estimates.

In the past few years, the debate about the validity of Null-Hypothesis Significance Testing (NHST) has gained importance, with the *American Psychological Association* and *Psychological Science* both recommending that wherever possible, authors use estimation and base their interpretation of results on point and interval estimates. A general movement of the field away from NHST and towards estimate based analysis also highlights the need to investigate how 'new statistics', including mixed-effects models, can be used to allow a better understanding of the data available to us than that afforded by traditional and familiar ANOVA and NHST. That being said, it is however important to note that mixed-effects models can be used in a variety of ways, and it is entirely possible to use mixed-effects models for NHST, as well as for estimation. A detailed discussion of how mixed-models can be used to improve NHST can be found in Barr et al. (2013), which will be referred to later in this chapter.

9.1.3 Missing Data

In addition to providing a solution to the language-as-fixed effect fallacy (Clark, 1973), which is particularly pertinent to psycho-linguistic experiments such as those described in this thesis, mixed-effects models are also relatively robust to missing data. This makes

them particularly interesting as an option for ERP investigations, in which missing data are inevitable.

In ERP experiments, each trial is accepted or rejected not only on the basis of a response (as in a memory test), but also on the basis of the quality of the EEG signal measured during that trial. Trials with excessive drift, or electrical noise (for example resulting from muscle movement) are rejected. This can lead to to a number of problems. The first is that the experiment may become unbalanced as trials are not rejected from all conditions equally. The traditional response to this problem has been to average across trials in each condition, but averaging necessarily comes at the expense of detail in the data. A second, and perhaps more obviously problematic outcome is the inefficiency of data collection if the rate of loss is high. In order to achieve a reasonable signal/noise ratio when averaging together trials, many experiments require a minimum of 16 trials per condition per subject (see Section 3.3.3). This means that if a subject provides less than 16 good, usable trials in any one condition, all of that subject's data is rejected from analysis. This was the criterion used in the experiments in this thesis, and the implications of failing to collect the minimum number of good trials for each participant and condition become particularly clear when one considers the duration of the experiments in this thesis. For example, in Experiment 3, reported in Chapter 7, (Comparing Fillers, Beeps and Coughs), there were twelve experimental conditions. If a participant performed well in eleven conditions, but provided only fifteen trials in one condition, then all of the data from that participant, which amounts to approximately two and a half hours of recording, and nearly four hours of laboratory time, in addition to processing time, is wasted. This is not only costly in terms of EEG consumables, but also in terms of hours of researcher and participant time. This is particularly evident when we consider the memory test data from the same experiment; although it was possible

to analyse the behavioural outcomes, the ERPs were un-interpretable as only three (of twenty-four) participants provided 16 correct, usable trials in all twelve conditions.

One approach to the problem of inefficiency and data-loss is to increase the number of trials in experiments, so that participants cannot possibly fail to produce enough trials for analysis. While this approach is fairly common in many types of ERP experiment, for example single word visual presentation language or memory tasks, it becomes impractical for experiments including a large number of conditions, particularly in tasks where each trial takes longer than one or two seconds. With increased experiment length, participant fatigue increases, increasing in turn the number of errors made due to lapsing attention, and the number of trials which are rejected due to muscle movement and alpha activity (a low frequency, 8-10Hz wave which appears particularly over posterior scalp electrodes, and is associated with lapsing concentration, a relaxed state, or fatigue).

For repeated measures ANOVA based analysis, data are averaged over all trials that a participant contributes to each condition, leading to the requirement for a minimum number of trials per condition, to prevent noise having an undue influence on the outcome. Thus each participant contributes a single averaged waveform for each condition. By contrast, mixed-effects modelling allows the incorporation of each individual trial that has been accepted for analysis after processing, meaning that far more individual data points are incorporated into the analysis. This necessarily means that the number of trials contributed by each participant will vary, but the robustness of mixed-effects models to missing and unbalanced data means that it is more likely to be possible to proceed with analysis.

Having now given a broad introduction to the topic of mixed effects modelling in the context of some of the problems encountered in language ERP experiments, the next

section will move on to discuss the application of this approach to ERP data.

9.2 Applying Mixed Effects Modelling to ERP data

Despite the potential advantages of a mixed effects approach, only a small number of attempts have thus far been made to apply it to electrophysiological data. These have variously focussed on modelling of mean amplitudes from a single electrode under a small number of conditions (Bagiella, Sloan, & Heitjan, 2000); wavelet analysis (Davidson, 2009); and point by point analysis of an EEG signal (Janssen, van der Meij, & Barber, 2013). However, the approach taken in the second and third studies mentioned above requires powerful computing resources and plenty of memory storage. Here, we attempt a compromise between traditional ANOVA analysis, and the highly intensive approach taken by Davidson (2009) and Janssen et al. (2013).

In the approach trialled here, each admissible trial is included for analysis, but rather than using every datapoint separately, we significantly reduced the size of the dataset by extracting the mean voltage over a pre-determined time window for each electrode in each trial. As data were digitised at 250Hz, a recording epoch of 2 seconds produced 500 time points for each electrode. Taking an average over a time window of interest reduced this down to one time point for each electrode for each epoch. The decision to attempt a 'lightweight' approach to mixed effects modelling with ERP data was mainly pragmatic, based on the computing resources available.

Whereas ordinary regression is fit to the data using a least-squares method, linear mixed-effects models are fit by iteration. Very big datsets (such as those obtained in ERP experiments) become very complicated to fit, leading quickly to convergence problems. For this reason, the data presented in this chapter were simplified still further by analysing

only a subset of electrodes. For N400 effect data, six midline electrodes were selected, whereas when analysing data expected to show a left-parietal old/new effect (LPONE) data were averaged over electrodes in each of the four quadrants of the scalp, creating four virtual electrodes.

Preparing data for mixed effects analysis required processing which deviated from that described in Chapter 4 (General Methods), and so the processing stream developed for this purpose, along with some of the reasoning and the limitations encountered, follows below.

9.3 Data Processing for Mixed Effects Modelling

9.3.1 EEG to ERP

One of the advantages of mixed models is that they use data more efficiently, as they can incorporate every available trial. However, given that single word data was not analysed in the original analysis of the data, a new processing stream was required to extract single trials from Neuroscan 4.5.

Data had to be completely reprocessed, beginning with the raw EEG (*.cnt*) files. First, the identity of each target word had to be attached to the appropriate epoch. To achieve this, each target word was first assigned a three digit numeric identity. Using R, and the output from each participant's *.edat* file (a file produced by the stimulus presentation software, which reports stimuli, presentation timings, responses etc.), the identity of the target presented to each participant in each individual trial was established. These data were incoroporated into the *.dat* file, in a column usually reserved for Response Latency. The *.dat* file is a file used to merge task data with EEG in Neuroscan. It was not

possible to incorporate stimulus identity into the trigger codes for each trial at the time of experiment presentation. This was because it was necessary to know the condition (fluency and predictablity) of each word for standard analysis. Combining 324 distinct stimulus identities with 6 (or 12 in Experiment 3) presentation conditions resulted in a greater number of individual trigger codes than the 528 permitted by Neuroscan.

For data from memory experiments, it was necessary to merge an original task data file (*.dat*) with the raw EEG at this point, to incorporate information about correct and incorrect responses. At this point, for all datasets (whether focussed on the N400 or recognition memory) a Neuroscan *.tcl*[1] file was run to exchange the trigger codes marking the beginning of each epoch with the stimulus identity codes found in the response latency column of the newly created dat files.

Following this, data were subjected to the same *Mark and Reject* and *Ocular Artifact Reduction* procedures as for standard processing. It was not possible to incorporate the stimulus identity data into any EEG file which had already been treated to *Mark and Reject* or *Ocular Artifact Reduction*, as each 'point' in the data file was counted as an "event"; for example, the beginning and of a rejected block would each be regarded by the software as an "event" requiring a trigger code. Any mismatch between the number of trigger codes and the number of events in the EEG file caused the process of exchanging old trigger codes for stimulus identities to fail. Had this not been the case, stimulus identites would have been wrongly assigned. Where data had already been processed for standard analysis, and it was desirable to be able to compare the outcomes of standard and mixed effects analysis, great care was taken to ensure that as far as possible, the same data were submitted to analsysis. To achieve this, the mark and reject stage was carried out to match the blocks selected for standard analysis by

[1]The *.tcl* file was written by Ric Sharp (http://www.sharp-apps.com), in response to a request for assistance for this project.

extracting their start and end times (accurate to within .001 of a second), and the ocular artifact reduction was run using a regression file (*.ldr*) created during the first processing of the data, so that the sample blinks selected to create an average blink profile did not vary across processing streams.

Finally, the EEG files were treated to the same epoching, baseline correction, drift detection, artifact rejection and averaging procedures as data processed for standard analysis, as is described in Chapter 4[2]. The obvious difference is, of course, that each "average" file consisted of only one trial.

9.3.2 Extracting Data and Preparing for Analysis

Once ERP files (and finally numeric representations of the data, *.dat* files) had been created, voltage means were calculated from each electrode over time windows of interest. This data was imported into R, where the remainder of the analysis took place.

Because Neuroscan had created average files for each individual stimulus for each participant, averages were created even when the data from a trial had been rejected on some grounds (artifact contamination, incorrect response *etc.*). This resulted in a significant number of trials which containing an amplitude of zero across all electrodes. To eliminate these, each data point was squared (to make all data-points positive), and the resulting squares were summed for each trial for each participant. Any trials where the resulting sum of amplitudes at all electrodes was zero were dropped from analysis.

[2]In the processing described in Chapter 4, in which condition averages were created, the final processing stages, from epoching to creating average files was completed using an automated script which took approximately 30-40 minutes to run for an experimental dataset (24 participants). To extract individual trials from the same dataset, as is described here, this same automated script required approximately 145 hours, and was relatively unstable thoughout. The subsequent process, in which ERP waveforms were converted to numeric averages over time windows of interest required a further 170 hours, as compared to the 15-20 minutes required to run this process for the grand-average based methodology described in Chapter 4. Whilst this does not affect the interpretation of data reported in this chapter, it is worth mentioning for the benefit of any reader considering similar analysis.

Using R, and again with reference to the *.edat* file for each participant, each stimulus identity was matched with its condition (fluency, predictability) at presentation, and this information was imported into the dataframe. At this point, the data were fully processed and ready for mixed-effects modelling analysis.

9.4 Hardware

Analysis was carried out initially on a Samsung NP3530EC laptop computer, with a 2.30GHz i5 CPU and 6GB RAM, running Windows 7 (64bit). Once it became apparent that processing speed was a serious problem, and that it would be beneficial to be able to run models for several datasets at once, further processing was conducted using Amazon's EC2 cloud computing service[3]. Virtual machines were created using Amazon's C3.Large Instances. These were based on Intel Xeon E5-2680 v2 processors, with each machine assigned two virtual processors and 3.75GB of RAM, and running the Ubuntu 14.04 LTS ("Trusty") operating system. R3.1.1 was installed on each virtual machine, along with RStudio Server v0.98.1062. Analysis was carried out using the R package lme4.0[4] (Bates, Maechler, & Bolker, 2013).

9.5 Datasets for Linear Mixed-Effects Modelling Analysis

For the purposes of this thesis, four datasets were selected for mixed-effects modelling; two datasets incoporating data from the auditory 'on-line processing' phase of the experiments, focussing on the N400 effect, and two datasets from the memory test phase,

[3] aws.amazon.com/ec2/

[4] lme4.0 was used for all analyses and not upgraded during the project, as discussion within the R community suggested fit problems with some of the newer versions of lme4, leading to convergence failures (Bicknell, 2014b). Although it appears that newer versions of lme4 (released June 2014) do not suffer the same fit problems previous versions (Bicknell, 2014a), we continued to use lme4.0 throughout the project for the sake of consistency.

focussing on the Left Parietal Old/New Effect (LPONE). For each dataset, the chief predictors of interest are predictability, fluency and location. Primarily, we are interested in whether predictability and fluency, or an interaction between predictability and fluency, influence the pattern of voltage across the scalp. As we are also interested in incorporating both subject and stimulus as random effects, a random intercept for each of these is included as the basis of each model. The way in which model structures were selected is described below.

9.6 Selecting a Random Effects Structure

A fully maximal random effects structure, as recommended by Barr, Levy, Scheepers and Tily, (2013), was not possible for the data reported in this thesis. Fully maximal random effects structures resulted in models which failed to converge after the standard maximum of 300 iterations, and reinforced the point made by Barr and colleagues (2013) that one should "*not be in any hurry to publish*" when running large models (p.298); such maximal models on the data reported in this thesis took up to ten days to run 300 iterations (before resulting in a failure to converge). A marginal improvement in processing speed was achieved by running models on virtual machines on a cloud computing service, but the reduction in processing time was not sufficient to warrant increasing the number of iterations; rather, we focussed on a strategy of simplifying the models.

Where the maximal model failed to converge, an attempt to reach convergence was first made by removing subjects with very few items in any one cell (*c.f.* Barr et al., 2013, p.276). Consideration was also given to whether items should be removed which had produced very few data-points in any one cell (*i.e.* very few participants had contributed a usable data-point for any fluency and predictability condition). Because each item

appeared only once to each subject, across the experiment each item was presented only a handful of times in each condition. Once lost trials due to processing are taken into account, this meant that many items were altogether absent from one or more conditions at the point of analysis. Removing all items with very few points in any one condition proved to be impractical due to the low trial numbers available.

Where the removal of participants contributing very low trial numbers did not resolve convergence problems, a random effects structure was built up gradually instead. This always began with a model incorporating only a random intercept for subject, which was assumed to be the greatest source of variance in the data. Subsequently, a random intercept for stimulus identity was added. Random slopes for subject and stimulus identity were added one step at a time, and each new model compared against the previous one to check whether the fit of the model to the data had improved. Anything improving the fit of the model to the data at a fairly liberal level ($p < 0.2$, *c.f.* Barr et al., 2013, p.276) was retained. Only once a full random effects structure had been built up were fixed effects incorporated into the model. Although this procedure was time-consuming, with the more complex random-effects only models taking several days to run, and final models including fixed effects, upwards of a week, it did allow the fitting of mixed effects models to the data.

To help determine significant effects, p-values were estimated based on the t distribution, using the following formula, as described by Baayen (2008: 248);

$$p = 2 * (1\text{-pt}(\text{abs}(X), Y - Z))$$

where X represents the t value, Y the number of observations, and Z the number of fixed effect parameters taken into account in the analysis. *abs* represents a function

calculating the absolute value of the value specified by X ($\sqrt{X^2}$), and pt is a function allowing the calculation of the cumulative probability of the t distribution for the value specified.

9.7 Applying Linear Mixed-Effects Models to Experiments 1-3

9.7.1 Establishing the Relibility of the Method - Comparing Mixed-Effects Models to ANOVA for a Balanced Dataset

If this type of 'lightweight' mixed effects modelling for ERPs is to be useful, it is desirable to know whether the results obtained are comparable to the results obtained from a traditional repeated measures ANOVA analysis. Knowing this determines whether or not it is valid to use mixed models to analyse data which are unsuitable for ANOVA analysis (for example unbalanced datasets), and then discuss the outcomes of ANOVA treated data alongside mixed-effects treated data.

To facilitate comparison with ANOVA analsis, the N400 data from Experiment 1 (Chapter 5; Fillers and Beeps) was reprocessed for mixed effects analysis. To recap, ANOVA analysis of this dataset had revealed significant main effects of predictability, fluency, and an interaction between predictability and fluency.

The data was analysed with an almost full random effects structure — random correlations were removed, and no stimulus random slope for location was included. Predictability, fluency and location were incorporated as fixed effects. Importantly, for comparison

with ANOVA, this analysis incorporated the same data as the ANOVA analysis. No extra participants were included, and by virtue of the careful *Mark and Reject* and *Ocular Artifact Reduction* procedures, the same epochs were selected for both analyses.

Predictability was centre coded, which reduced the likelihood of Type II errors due to inflated standard errors, by reducing collinearity within the model. This also had the effect of aiding comparison with ANOVA analysis — any effect of predictability could be interpreted in terms of a main effect in ANOVA. Fluency was not centred. Dummy contrasts were set up to compare data from the disfluent *(er)* condition with the fluent and beep conditions.

The random effects structure was built up stepwise. The final model had the following structure in R[5]:

```
FinalModel <- lmer(V ~ 1 + s(Predictability)*Fluency*Location+
(1|Subject)+(0+s(Predictability)|Subject)+(0+Fluency|Subject)+
(0+Location|Subject)+(0+s(Predictability):Fluency|Subject)+
(0+s(Predictability):Location|Subject)+(0+Fluency:Location|Subject)+
(0+s(Predictability):Fluency:Location|Subject)+
(1|StimID)+(0+s(Predictability)|StimID)+(0+Fluency|StimID)+
(0+s(Predictability):Fluency|StimID),
REML=F)
```

Three fixed effects were incorporated: predictability, fluency and location. Removal of any one of these three, or interactions between them, caused the model to fail to converge.

[5]In the code below, and throughout this chapter, '**s()**' denotes a centring function in R. The symbol '**:**' denotes specific interactions between predictors, whereas '*****' denotes main effects and all possible interactions.

Electrode	Estimate	Estimated N400 effect (μV)	SE	t	p (estimated)
Predictability: Location (Fz)	0.698	-0.734	0.353	1.978	0.048
Predictability: Location (FCz)	0.299	-1.501	0.348	0.832	0.405
Predictability: Location (Cz)	-2.464	-2.631	0.509	-4.845	<0.001
Predictability: Location (CPz)	-0.042	-3.245	0.338	-0.124	0.901
Predictability: Location (Pz)	0.045	-3.428	0.348	0.129	0.897
Predictability: Location (POz)	0.321	-3.378	0.361	0.890	0.373

TABLE 9.1: Estimates for the effect of predictability across levels of location. As linear mixed-effects models are additive, the estimated voltage for the interaction between predictability and location at each electrode is found by summing relevant main effects. *E.g.*, the estimated magnitude of the N400 effect at electrode Fz is found by summing the estimate for the intercept, the estimate for the effect of predictability, the estimate for the effect of location at Fz, and the estimate for the interaction between predictability and location at Fz. The full model output table, including estimates for all main effects and interactions can be found in Appendix B.1. The interaction of predictability with location reaches significance at the Fz electrode.

Thus it was concluded that the fixed effects and their interactions were necessary for the fit of the model.

Having established the form of the model, its output was inspected. Only significant effects and interactions of interest are reported here. A full model table can be seen in Appendix B.1. The model revealed that unpredictable targets elicited ERPs which were significantly more negative than predictable words [estimate =-2.46μV, SE = 0.51, t = -4.48, p<0.00001]. This effect of predictability was revealed to be greater towards the back of the head and maximal over Pz, consistent with an N400 effect (although the interaction of predictability with location was significant only at the frontal Fz and central Cz electrodes). The estimates for the interaction of predictability with location are given in Table 9.1.

There was no evidence for an interaction of predictability with fluency. The model showed that the magnitude of the N400 effect did not vary with fluency; there was no significant interaction of predictability with fluency when the disfluent *(er)* condition was compared to the fluent condition [estimate = -0.537μV, SE = 0.57, t = -0.935 p>0.3], or when the disfluent *(er)* condition was compared to the interrupted (beep)

condition [estimate = 0.231μV, SE = 0.60, t = 0.387, p>0.65]. No other interactions involving predictability reached significance.

Comparison with ANOVA analysis

The output of this model differs somewhat from the ANOVA analysis reported in Section 5.3.2, in that the ANOVA revealed a significant interaction of predictability with fluency [F(1.98,47.59) = 3.63, p<0.05], whereas the output of the mixed effect model presented here has not revealed evidence for for such an interaction. It is worth noting that the ANOVA reported in Section 5.3.2 was performed on electrodes in the global array (due to evidence of hemispheric effects for some conditions), whereas the mixed effects model reported here made use of data from midline electrodes. This was to reduce the complexity of the model, and reduce convergence problems To check whether this discrepancy between the findings of the ANOVA analsis and the mixed effects model analysis could be due to the difference in electrodes selected, the ANOVA was repeated using the same midline electrodes as were incorporated into the mixed effects model. This ANOVA revealed a marginally significant interaction of predictability with fluency [F(0.47,11.27) = 2.92, p<0.1]. The interaction of predictability with fluency was significant between targets which were fluent and disfluent [F(1,24) = 5.84, p<0.05], but not between targets which had been disfluent and interrupted [F(1,24) = 1.09, p>0.3]. This again differs from the findings of the mixed effects model, which revealed no interaction of predictability with fluency.

Finally, as t-tests carried out on the N400 effect maxima (CPz) had revealed a marginally significant difference between the N400 effects to fluent and disfluent utterances [t(24)= -2.03, p = 0.054], but not between disfluent and interrupted conditions [t(24) = 1.01, p >0.3], the mixed effects model was applied to data from the CPz electrode only (the

fixed effect of location was omitted). This analysis also revealed no significant interaction between predictability and fluency, although the interaction of predictability and fluency approached marginal significance between utterances which had been fluent and disfluent [estimate = -0.696μV, SE = 0.43, p = 0.1].

Although the pattern of data described by the ANOVA analysis and mixed effects models analysis is similar, the results of the mixed effects models would not lead to a confident statement that the N400 predictability effect varies with fluency.

Including random effects for items should reduce Type I errors, by allowing us to test whether the same effects would be found if the experiment were repeated with a different set of items. Therefore, failing to find an interaction between predictability and fluency is not altogether surprising — we should expect some weak effects to disappear once we include random effects for items. The implications of this are discussed in Section 9.8.

9.7.2 Applying Mixed-Effects Models to an Unbalanced Dataset

Having demonstrated that this type of lightweight mixed-effects modelling provides slightly different insights into the the data than the ANOVA analysis for the N400 data presented in Experiment 1 (Chapter 5), this approach is now applied to the memory data from the Experiment 2 (presented in Chapter 6). In processing these data for ANOVA analysis, a significant proportion of subjects were lost, as they did not contribute enough trials in every condition, particularly as a result of errors in participant identifying trials as old or new. The ability of mixed models to handle unbalanced datasets means that these participants can now be incorporated into analysis, allowing a much bigger dataset to be analysed, and possibly providing more power for the analysis.

Inspection of the two memory datasets already analysed using ANOVA revealed that the data obtained during Experiment 2 (Coughs) provided the most usable trials overall. Additionally, this dataset contained less potentially unstable datasets (participants with less than 8 trials in any condition) than the dataset obtained during the memory test of Experiment 1. Consequently this dataset was selected.

ANOVA analysis of the LPONE (500-800ms) time window of this this dataset had revealed no significant differences between the recognition effects between words which had been fluent, disfluent or interrupted and unpredictable in their contexts. No quantitative comparison was made across all six fluency and predictability conditions, due to differences in scalp topography. As the mixed-effects model took into account particiapants previously excluded from analyses due to low trial numbers, the data incorporated are not the same, and so topographic differences uncovered in ANOVA may or may not exist in the current dataset. Ideally, the data analysed here would also have been submitted to topographic comparison, but computational limitations meant that this was not possible.

Difference ERP waveforms were created by subtracting the averaged waveform for 'new' items from the waveform for each 'old' trial. Data were analysed averaged over the classic Left-Parietal Old/New Effect (LPONE) time-window (500-800ms). To simplify the data for analysis, four regions of interest were selected. This is in contrast to the previous section, in which data from the mid-line was analysed. In the previous section, we were primarily interested in the N400 effect, which is strongest towards the mid-line, whereas here, we are interested in the LPONE, which takes the form of a relative positivity that is strongest over electrodes in the left-parietal region of the scalp. The four regions of interest selected constituted four strings of electrodes, one in each quadrant of the scalp (see Figure 9.1). These strings were Front Right (F2, F4, F6), Front Left (F1, F3, F5),

Parietal Right (P2, P4, P6), and Parietal Left (P1, P3, P5). Voltages were averaged across each of these strings.

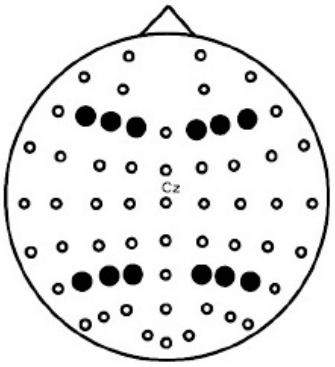

FIGURE 9.1: Map of electrodes on the scalp, showing the electrodes incorporated into the mixed-effects analysis of memory effects. Electrodes from four strings, front-left (F1, F3, F5), front-right (F2, F4, F6), parietal-left (P1, P3, P5), and parietal-right (P2, P4, P6), were used. Data were averaged over each string, so that one virtual electrode was created for each string.

A random effects structure was built up, beginning with random intercepts and adding random slopes and interactions, testing each new model to determine whether it provided a significantly better fit to the data than the previous model. As in the previous analysis, where a model failed to converge, or the fit was not better than the previous model (at a relatively liberal significance level, $p < 0.2$), the newly introduced random slope or interaction was discarded.

The model was planned with three fixed effects; string, predictability and fluency. However, incorporating all three, with interactions, into the model led to a failure to converge. Consequently, the fixed effects and their interactions were added to the model one at a time, and, as with the random effects, their contribution to the fit of the model was tested to establish which model provided the best fit to the data. Where the addition of a fixed effect or interaction led to a convergence failure, this effect or interaction was discarded from the model.

The fit of the model was significantly improved by the addition of a fixed effect of string [$\chi^2(3) = 8.943$, p <0.05] and further marginally improved by a fixed effect of predictability [$\chi^2(1) = 3.468$, p <0.1]. Adding a fixed effect of fluency resulted in convergence failure. The final model incorporated fixed effects of string and predictability; predictability was centred to reduce the likelihood of Type II errors. Subject and stimulus were used as random intercepts. The model also incorporated a subject random slope for string; a subject random slope for predictability and fluency (including interactions); a stimulus random slope for predictability; and a stimulus random slope for fluency; a stimulus random slope for string. The structure of the final model as specified in R is given below.

```
FinalModel <- lmer(V~1+String+s(Predictability)+
(1|Subject)+(0+String|Subject)+(0+s(Predictability)*Fluency|Subject)+
(1|StimID)+(0+s(Predictability)|StimID)+(0+Fluency|StimID)+(0+String|StimID),
REML = F)
```

Having established the structure of the final model, its output was inspected. The model output table can be seen in Appendix B.3.

The relationship between the estimated values at each of the four strings appeared to describe a LPONE, with the voltage more positive for the left parietal string than for the right parietal and frontal strings. This difference achieved significance between the left parietal string was compared to the right parietal strings [estimate = -0.976μV, SE = 0.308, t - 3.166, p <0.05]. Additionally, unpredictable words elicited a generally more negative ERP than predictable words, and this effect was marginally significant [estimate = -0.752μV, SE = 0.386, t = -1.951, p <0.1].

As the fit of the model was not improved by the inclusion of an interaction of predictability with string, there is no evidence to suggest that the distribution of the LPONE is affected by predictability, and as the inclusion of fluency also did not improve the fit of the model, there is no evidence that the memory effect was affected by the fluency of the utterance.

Comparison with ANOVA analysis

In ANOVA analysis of the same dataset (see Section 8.4.2), quantitative comparisons were not carried out across all six predictability and fluency conditions, as topographic comparison revealed significant differences in the distribution of old/new effects across conditions. Quantitative comparison was carried out between fluency conditions for predictable words, and revealed no significant differences. Although it is not possible to make a direct comparison, it seems as though the output of the ANOVA and mixed-effects analyses are broadly in agreement.

9.7.3 Applying Mixed-Effects Models to a Dataset Unsuitable for Standard ANOVA Analysis

Having demonstrated that unbalanced ERP datasets can be analysed using mixed-effects models, and extended for unbalanced datasets, it follows naturally to attempt to apply this to an ERP dataset not accessible to ANOVA analysis, such as that obtained in the memory test of Experiment 3 (Comparing Fillers, Beeps and Coughs). This dataset produced only three participants with 16 usable trials in all conditions; the minimum number of trials for ANOVA analysis of averaged ERP waveforms.

As with the previously detailed analysis (Experiment 2 [Coughs] memory), ERP data from across the scalp were condensed down into four strings, representing the four quartiles of the scalp.

As fixed effects, the model used string, predictability and fluency. Once again, predictability was centred around its mean value to reduce the likelihood of Type II errors. The random effects structure for the model was built up beginning with random intercepts, and random slopes and interactions were added individually. Each new model was tested to determine whether it provided a better fit than the previous model. Where adding a random slope or interaction did not improve the fit of the model, or caused a failure to converge, that random slope or interaction was not incorporated into subsequent models. As random intercepts, the final model used subject and stimulus. The model had a subject random slope for string, and subject random slopes and interactions for predictability and fluency. There was a stimulus random slope for predictability, and a stimulus random slope for fluency. The full model as specified in R is given below.

```
FinalModel <- lmer(V~1+String*s(Predictability)*Fluency
+(1|Subject)+(0+String|Subject)+(0+s(Predictability)*Fluency|Subject)
+(1|StimID)+(0+s(Predictability)|StimID)+(0+Fluency|StimID),
REML = F)
```

To test whether each of the three fixed effects significantly improved the model, the fit was compared between the full model, and the model with each fixed effect (and its interactions) removed. The fit of the model was significantly improved by the inclusion of the fixed effects and interactions of string [χ^2 (18) = 37422, p <0.0001]; predictability [χ^2 (12) = 37412, p <0.0001]; and fluency [χ^2 (16) = 37419, p <0.0001].

Once the structure of the final model had been established, the model itself was inspected. This can be seen in Appendix B.4 This did not reveal evidence of an LPONE scalp distribution, in which a positivity is expected at left parietal electrode sites. Rather, the model estimated more positive voltages at frontal strings than at the left parietal string (intercept); Front left [estimate = $5.375\mu V$, SE = 1.651, p <0.005]; Front right [estimate = $6.789\mu V$, SE = 1.921, p <0.0005]. Predictability did not have a significant overall effect [estimate = $0.461\mu V$, SE = 0.745, t = 0.619, p >0.5]. The interaction between predictability, location and fluency at frontal electrode sites produces a positive estimate when fluent targets are compared to disfluent targets, both in the left hemisphere [estimate = $2.373\mu V$, SE = 0.632, t = 3.758, p <0.0001] and in the right hemisphere [estimate = $1.274\mu V$, SE = 0.631, t = 2.018, p <0.05]. Importantly, this doesn't produce a pattern of voltage over the scalp identifiable as an LPONE - frontal electrodes are significantly more positive than posterior electrodes. No other effects achieved significance. Thus, this analysis has revealed no evidence for LPONE, or significant differences across conditions of fluency and predictability.

9.7.4 Incorporating a Linear Predictor

Earlier in the chapter, I mentioned that linear mixed-effects models permit the inclusion of data which are linear, as well as factorial. In the models reported thus far in this chapter, the models have all incorporated factorial fixed effects only; their structure has been somewhat analogous to ANOVA. In order to demonstrate the use of mixed-effects models in answering questions involving linear predictors, we now return to the the N400 data from Experiment 3 (Chapter 7), and incorporate a factor of trial number into the analysis. No significant interaction was revealed between fluency and predictability in the analysis reported in Chapter 7. Like most repeated measures experiment, this

experiment is subject to the potential criticism that participants may learn to expect the experimental manipulation over the course of the experiment. Specifically, in the experiments reported in this thesis, participants may become accustomed to the unexpected sentence endings, and so the unpredictable and predictable targets produce increasingly weak N400 effects as the experiment goes on. It is also possible that any interactions between predictability and fluency are observed only at the beginning of the experiment, and that over time, participants response to disfluency is reduced. It is possible to investigate this by incorporating a factor of trial number into the model for the data.

Analysis was carried out on data from mid-line electrodes. As was the case for the models reported earlier in this chapter, a fully maximal random effects structure led to a failure to converge. Consequently, the model structure was built up, beginning with random effects, and tested at each stage to ensure that each additional level of complexity was justified by significantly improving the fit of the model. The final model incorporated random intercepts for subject and stimulus, subject random slopes for location and context, and subject random slopes and interactions for predictability and fluency, as well as stimulus random slopes for predictability, fluency and trial number. To reduce the likelihood of Type II errors, predictability was centred. The final model, as specified in R, is given below.

```
FinalModel <- lmer(V~1+s(Predictability)*Fluency*Location*Context*Trial
+(1|Subject)+(0+s(Predictability)*Fluency+Location+Context|Subject)
+(1|StimID)+(0+s(Predictability)+Fluency+Trial|StimID),
REML = F)
```

Electrode	Estimate	Estimated N400 effect (μV)	SE	t	p (estimated)
Predictability: Location (Fz)	0.905	0.094	0.998	0.907	0.364
Predictability: Location (FCz)	0.371	-0.990	0.998	0.372	0.710
Predictability: Location (Cz)	-1.617	-2.463	0.843	-1.919	0.055
Predictability: Location (CPz)	-0.139	-3.444	0.998	-0.139	0.889
Predictability: Location (Pz)	0.064	-3.779	0.998	0.064	0.949
Predictability: Location (POz)	0.293	-3.719	0.998	0.294	0.769

TABLE 9.2: Estimates for the effect of predictability across levels of location. The magnitude of the N400 effect at each electrode is calculated by summing the estimates for relevant main effects, and interactions. As can be seen, the magntide of the predictability effect is greatest towards the rear of the scalp, consistent with an n400 effect. The interaction of predictability with location reaches marginal significance at the Cz electrode.

To test the contribution of each fixed effect, the fit of the model was compared between the full model and the model with the fixed effect and its interactions removed. The fit of the model was significantly improved by the inclusion of each of the fixed effect predictors and their interactions; predictability [χ^2 (72) = 98.93, p <0.05]; fluency [χ^2 (96) = 627.71, p <0.0001]; location [χ^2 (120) = 368.42, p <0.0001]; context [χ^2 (72) = 347.57, p <0.0001]; and trial number [χ^2 (72) = 118.36, p <0.0005].

Having established the structure of the final model, its output was inspected. The model output table can be found in Appendix B.5. Unpredictable items elicited ERPs which were more negative than predictable items [estimate = -1.617μV, SE = 0.843, t = -1.919, p <0.1]. This effect was greater at posterior electrodes than frontal electrodes, consistent with an N400 effect, although the interaction of predictability with location did not reach significance at any electrode. See Table 9.2 for estimates of the N400 amplitude at each location.

No interactions of involving predictability reached or approached significance, thus this analysis provides no evidence to suggest that the magnitude of the N400 effect was influenced by fluency, context or trial number. Importantly, the interaction between predictability and trial number was far from significant [estimate = -0.0005μV, SE =

0.004, t = -0.126, p = 0.90], demonstrating that the effect of the unpredictable targets did not appear to diminish throughout the experiment.

9.8 Summary and Conclusions

The previous sections have outlined mixed-effects models, and described how they might be of service to ERP researchers conducting psycho-linguistic experiments, in providing a way to model both stimulus and participant variance within one model, and being robust to missing data and unbalanced datasets. Further, I have described a processing stream to allow the extraction of single trial data from Neuroscan 4.5, and the application of mixed-effects models to the data generated from the experiments described in this thesis.

9.8.1 New Insights into the Data Analysed using Linear Mixed-Effects Models

No effect of fluency on Experiment 1 N400 Effect

Applying linear mixed-effects models with crossed random effects for subjects and items allowed some slightly different insights into the data than the factorial ANOVA analysis reported thus far in this thesis. This is particularly interesting in the case of the on-line processing (N400) data from Experiment 1, in which factorial ANOVA revealed a significant interaction of predictability with fluency, whereas mixed-effects models found no such interaction. Whilst the mixed-effects model incorporated crossed random effects for subjects and items, the ANOVA included random effects for subject only, and did not model the effect of stimulus identity as a source of variation. The fact that

the predictability-fluency interaction disappears when stimulus is included as a random effect might suggest that this interaction does not generalise beyond the selection of stimulus words used in the experiment.

If it is the case that N400 effect does not vary with fluency, then this is in line with the findings of Experiments 2 and 3, in which ANOVA analysis also revealed no evidence for a significant effect of fluency on the N400 effect. This would add to the evidence suggesting that a disfluency effect on the N400 effect is not robust.

No Effect of Fluency or Predictability on the LPONE for Experiment 2

Experiment 2 compared disfluent fillers to coughs. In the analysis reported in this chapter, the ERPs from the memory test were analysed, focussing on data from the classic LPONE time window (500-800ms). Analysis made use of data averaged over a string of three electrodes from each of the four quadrants of the scalp. A mixed effects model revealed the difference waveform (old-new) to be more positive in the left parietal quadrant than elsewhere on the scalp, as would be expected for an LPONE. Adding a fixed effect of fluency did not improve the fit of the model, and there was no evidence for a significant interaction between predictability and string.

ANOVA analysis also revealed a more positive ERP at posterior locations, but quantiative comparisons were not carried out across all six fluency and predictability conditions due to differences in topographic distribution, present when rescaled data were analysed. Computing limitations meant that no re-scaled topographic analysis was carried out on the data submitted to linear mixed-effects analysis.

The failure to detect an effect of fluency does not contradict the ANOVA analysis, in which LPONE effects were compared across unpredictable words, revealing no effect

of fluency. The finding of no effect of predictability is not directly comparable to the ANOVA analysis, but does not appear to contradict the impression given by visual inspection of the waveforms (Figures 8.12 to 8.17) and the scalp topographies (Figure 8.18) representing data from the same experiment. It should, however, be borne in mind that these waveforms and topographies do not represent exactly the same data; different trials were selected and averaged together for the ANOVA analysis in Chapter 8 than for the mixed-effects analysis discussed here.

The robustness of linear mixed-effects modelling techniques to missing and unbalanced data has allowed the incorporation of all of the available data into the analysis, representing a significant increase compared to the data incorporated into ANOVA. However, in this case we do not draw significantly different conclusions from the output of mixed-effects modelling to those we drew from ANOVA. Specifically, we have not revealed evidence of an LPONE whose magnitude varies with the likelihood of participants correctly identifying old words. Behaviourally, participants were more successful at identifying words which had been unpredictable at presentation, and those which had been disfluent or interrupted by a cough. This was not reflected in the ERP data, whether analysed using ANOVA or mixed-effects models.

No Clear LPONE for Experiment 3

The ERP data from the memory test in Experiment 3 (Comparing Fillers, Beeps and Coughs) was not acessible for ANOVA analysis. Applying the minimum criterion of 16 trials per condition, per participant, before averaging data for ANOVA analysis leaves only three participants qualifying for analysis. We do not need to apply this criterion for mixed-effects analysis, due to the models' robustness to unbalanced datasets.

Mixed effects analysis of the ERP data from the 500-800ms time window from the Experiment 3 memory epochs failed to reveal an LPONE, but has, nonetheless allowed analysis of the dataset.

Investigating Whether N400 Effects Vary Throughout an Experiment

In common with the ANOVA analysis of the N400 data from Experiment 3 (see Section 7.3.1) mixed-effects modelling of the data revealed no evidence for any effects of fluency or context on the N400 effect. Additionally, we have found no evidence to suggest that the N400 effect changed throughout the experiment. It is reasonable to question whether unpredictable sentence endings remain unpredictable throughout an hour-long experiment, and thus whether the N400 effect might be reduced in later trials; the model reported here has not revealed evidence of this.

One point to consider, however, is that any effect of learning within the experiment may not be linear. It is, in fact, likely that any adaptation effects may occur very early in the experiment. However, given the difficulty of fitting fairly basic models to these datasets, it seems apparent that to explore anything more complicated than linear models would be impractical. Thus this is provided as a demonstration-in-principle of the type of question one might investigate using LMEMs, and any discussion concerning changes in effects throughout the experiment should bear this in mind.

9.8.2 Running 'Lightweight' Linear Mixed-Effects Models on ERP Data

While previous explorations into using mixed-effects modelling to analyse ERP data have employed wavelet analysis (Davidson, 2009) and point by point analysis (Janssen et al., 2013), we chose to attempt a lightweight approach, reducing each epoch down to a

mean voltage over a time window of interest. This decision was mainly pragmatic, given the computing resources available. Nevertheless, the models took a significant amount of time to run, and part-way through the project it became apparent that more powerful computing was required even for the 'lightweight' approach attempted here, prompting a move to cloud-based computing.

From this attempt to run mixed-effects models on voltages averaged over time-windows of interest, a number of observations can be made. Perhaps the most pertinent with regard to deciding whether to run such models in future is the length of time taken to run the models. Models were run on a subset of electrodes from the scalp; four virtual electrodes for memory data, and six midline electrodes for N400 data. Despite this reduced dataset, running the lme4.0 default of 300 iterations for a maximal model took between six and ten days. This is in contrast to the ANOVA analysis reported in Chapters 5-8, in which analyses incorporating up to thirty electrodes were completed in under one second.

For all of the datasets examined here, fully maximal models resulted in failures to converge. Consequently, models were built up stepwise, beginning with random intercepts only, and adding random slopes and interactions, and finally fixed effects, testing at each stage to determine whether the additional complexity added was justified by an improvement in the fit of the model. This building process added significantly to the time required for analysis, as many of the models required in the region of twenty separate random effects models before fixed effects could be added. As random effects structures gained complexity, these too required several days to run, such that arriving at a final model required R to be running models continuously for eight to ten weeks.

Because the process of arriving at a final model structure is iterative, and the decision

on the structure of each model run depends on the outcome of the previous model, running multiple models simultaneously is an inefficient solution; taking into account main effects and interactions, the number of possible model structures is very large. Some limited benefit can be gained from more powerful computing resources to increase the speed of running each model. As mixed-effects models are iterative, they cannot easily be parallelized. At the time of writing, there is no readily available package allowing the parallelization of processes in R over multiple threads, suitable for use with linear mixed-effects modelling iteration procedures.

In addition to being unable to run models with fully specified random effects structures, models incorporating electrodes from the global array were not attempted, nor were models assessing topographic distribution of effects and making use of re-scaled data. It was also not practical to incorporate a predictor accounting for the co-variance between electrodes based on their proximity to one another. This would have fulfilled a function similar to the Greenhouse-Geisser correction used in ANOVA analysis, and although not strictly necessary for a mixed-effects approach, including such a predictor may have improved the fit of the models. All of these limitations were due to the time taken to run the simple models presented here, and so this chapter should be considered to be a demonstration-in-principle of the method, rather than a fully satisfactory analysis strategy.

9.8.3 Is it Worthwhile to Run Linear Mixed-Effects Models on ERP Data?

Whereas previous work has employed wavelet-analysis or point by point analysis of ERP data for mixed-effects modelling, this chapter has described a lightweight approach. For

data which had been previously been analysed using ANOVA, the output of mixed-effects models did not differ greatly from ANOVA output. However, mixed-effects modelling has permitted the analysis of a dataset inaccessible to ANOVA (Experiment 3 memory). Additionally, it has been possible to include linear factors into a model, allowing the investigation of the effect of trial number within the experiment on N400-effect amplitude. This is a question not easily answerable with ANOVA, and so for questions or data-sets unsuited to ANOVA, mixed-effects models are a useful option to have. For data which is accessible to ANOVA, however, it is not clear that the significant cost in computing time is necessarily well rewarded.

Chapter 10

General Discussion

10.1 Introduction

The experimental work in this thesis has investigated the effects of disfluency on listeners, specifically investigating the role of delay in driving any disfluency effects. The effect of disfluency has been measured using both behavioural and neural indices to explore how disfluency affects listeners on-line processing, and subsequent memory for disfluent speech. Mixed-effects models processing has been implemented as an alternative method for analysing ERP data, and for analysing ERP data inaccessible to traditional ANOVA analysis. In this chapter, the findings of these experiments are summarised, and an initial interpretation of results is provided.

Three experiments were carried out. Each experiment consisted of two tasks; a listening task, and a surprise memory test. In each experiment, participants heard fluent sentences, sentences containing disfluent fillers, and sentences containing an interruption to be compared with the disfluent fillers. The first experiment compared the effects of disfluent fillers to artificially edited beeps of the same duration. In the second experiment,

disfluent fillers were compared to mid-sentence coughs, also of the same duration as the fillers. The third experiment combined the paradigms of the first two experiments, and the listening task consisted of two blocks, one in which disfluent fillers were contrasted with beeps, and one in which disfluent fillers were contrasted with coughs.

10.2 Summary of Results

10.2.1 On-line processing ERP results

In Experiment 1 (Chapter 5), an effect of fluency was seen on the amplitude of the N400 effect. Target words preceded by a disfluent filler (*er*) elicited a marginally smaller N400 effect than target words in fluent utterances, and words preceded by a beep elicited an N400 which did not differ in magnitude from fluent or disfluent words. The reduction in N400 amplitude for disfluent words suggests that where a disfluent filler preceded the target word, predictable and unpredictable words were processed more similarly than when the sentence was fluent. Words preceded by a disfluent filler also elicited a late positivity, consistent with a Late Positivity Complex (LPC), theorised to reflect memory retrieval and control processes. However, when the data from the N400 time-window (200-500ms) were re-analysed using linear mixed-effects modelling, incorporating crossed random effects for subjects and items, no effect of fluency was revealed.

In Experiment 2 (Chapter 6), no difference emerged between N400 amplitudes for words which had been fluent, disfluent, and interrupted by a cough. An LPC was apparent in all three fluency conditions, with unpredictable words eliciting a relative positivity compared to predictable words. There was no evidence for any difference in the topographic distribution or the magnitude of the LPC across fluency conditions, suggesting that the

degree to which memory control processes were engaged in resuming the meaning of the sentence on encountering an unpredictable word was not affected by fluency.

In the third experiment reported in this thesis (Chapter 6.4.4), there was no evidence for a difference in N400 magnitude across fluency and context conditions. For all conditions, there was some evidence of a late positivity for unpredictable words, but as this varied in topographic distribution, no quantitative comparison was made. Re-analysis of the data from the N400 time-window also revealed no effect of trial number, demonstrating that participants sensitivity to unpredictable words did not degrade through the experiment.

10.2.2 Recognition Memory — Behavioural Results

Experiments 1, 2 and 3 all revealed significantly better memory performance for words which had been unpredictable than predictable. Additionally, in Experiments 1 and 2, a difference emerged between words which had been fluent, and those which had been affected by some form of delay. Participants were significantly more successful at identifying words which had been presented in disfluent or interrupted sentences than those which had been presented in fluent sentences, and no differences emerged between the disfluent and interrupted conditions. No interactions emerged between predictability and fluency. In Experiment 3, there was no effect of fluency on memory performance, although unpredictable words were more likely to be recognised than predictable words.

10.2.3 Recognition Memory — ERP Results

EEG was collected during the memory test in all three experiments. ERPs suffered from low trial numbers, as memory performance was relatively low. Two time-windows were

examined within the ERPs, corresponding to the mid-frontal old/new effect, and the Left Parietal Old/New Effect (LPONE).

In Experiment 1, there was no evidence of a mid-frontal effect, although there did appear to be a weak LPONE. Differences in topographic distribution meant that a quantitative comparison was not carried out across all six fluency and predictability conditions, but within predictable words, where no significant topographic differences emerged, there were no significant differences in magnitude.

Similarly, Experiment 2 did not reveal mid-frontal recognition effects, but there was some evidence of an LPONE. Topographic distribution differences prevented comparison of the LPONE across all six predictability and fluency conditions, but within the unpredictable condition, where there were no significant differences in topographic distribution, fluency did not significantly affect LPONE magnitude. Re-analysis of the LPONE data using mixed-effects modelling also suggested an LPONE distribution across the scalp, with the left parietal electrodes exhibiting more positive voltages than elsewhere on the scalp. There was no evidence to suggest that this LPONE varied with fluency, but words which had been unpredictable did elicit an ERP effect which was generally more negative than predictable words. It should be remembered here that only a subset of electrodes were incorporated into mixed-effects modelling analysis, and that topographic comparisons were not conducted.

The ERP data from the memory test in Experiment 3 were not analysed using ANOVA, as insufficient trial numbers were returned by all but three participants. Linear mixed-effects analysis on a subset of the data from the LPONE time-window (500-800ms) revealed no evidence of an LPONE, but rather a relative negativity at left-parietal electrodes for correctly recognised old words.

10.3 Interpretation of Results

Although this project set out to investigate whether delay could be responsible for the attenuated N400 effects reported in Corley (2007), MacGregor et al. (2010), consistent N400 attenuation for disfluent words was not found in the experiments reported in this thesis. However, this does not indicate that participants were impervious to the fluency manipulation; disfluent words were significantly more likely to be remembered than words that had been presented in fluent context. This was the case even when no N400 attenuation had been observed. In the conclusion to Chapter 7 we suggested that if the N400 was not directly sensitive to disfluency, but attention was, then we might observe a memory performance effect for disfluency, even in the absence of N400 attenuation. This is indeed the pattern observed in the experiments in this thesis.

There is theoretical reason to believe that the temporal and prosodic disruption introduced by disfluency could be responsible for raised attention and improved subsequent memory. British English is a relatively strongly stress-timed language (Deterding, 2001), meaning that equal temporal distance between stressed syllables is more important than equal duration of syllables. As such, any disruption to the flow of speech will disrupt the regularity of the rhythm of stressed syllables, and is likely to be salient to the listener. Additionally, attention is easily grabbed by changes to the physical properties of a stimuli (Cherry, 1953; Scharf & Buus, 1986). If this is the mechanism which causes the raised attention reported by Collard et al. (2008) following disfluency, then it is plausible that such attentional changes should also be seen in response to noisy interruptions which disrupt the rhythm of speech similarly.

It would have been satisfying to investigate attention-related ERP effects, such as the MMN and P300 within the data reported here. However, these experiments were not

designed with these effects in mind, and so do not easily lend themselves to making the contrasts necessary to draw out these effects. In contrast to Collard's (2008) work, the target words in these experiments do not differ physically from their surrounding environment, but are rendered surprising or otherwise by the semantic context of the sentences. Having target words appear in physically 'standard' and 'oddball' conditions allowed Collard to investigate attention-related ERP effects within each fluency condition, using an interaction paradigm similar to that described in the experiments in this thesis. Attempting to investigate the MMN and P300 without such an interaction, by directly comparing the amplitude of the ERP waveforms across fluency conditions, raises the same issues of systematic baseline differences which led to the original decision to use an interaction design.

An alternative might be to suggest investigating the ERP waveforms timelocked to the onset of the disfluency or interruption. Whilst possible, the experiments presented in this thesis were not designed specifically to allow this, and so some re-coding of the original EEG files would be necessary to allow data to be processed time-locked to the disfluency or interruption, given that each disfluency or interruption was of a different length. However, given the design of these experiments, analysis of attention-related ERP effects time-locked to disfluency/interruption onset is not entirely free of systematic baseline problems. Although disfluent and interrupted conditions would both have ongoing speech in the pre-epoch baseline it should be remembered that interrupted utterances were created by splicing a noisy interruption onto an otherwise fluently spoken sentence, whereas disfluent *ers* were recorded within the sentence, which contained changes in tempo and prosody consistent with upcoming disfluency. Thus systematic differences in the stimuli during the pre-epoch baseline are inevitable.

10.3.1 Understanding the Lack of N400 Attenuation

The experiments in this thesis have not demonstrated a robust N400 attenuation for items affected by disfluency. Although a failure to find such an attenuation is not strong evidence against the effect, it does raise some questions about why N400 attenuation was not observed.

One area to scrutinise in psycho-linguistic experiments is the stimuli. The stimuli for the experiments reported in this thesis were produced in a carefully controlled manner, and the cloze-testing of prospective stimuli was carried out using participants from the same pool as the auditory experiments. The validity of the predictable/unpredictable targets is demonstrated by the robust N400 effects obtained across experiments. These robust N400 effects also demonstrate that participants were attending to the experiment, and comprehending the stimuli.

If the failure to detect N400 attenuation following disfluency was a product of the quality of the stimuli, then this would suggest that the disfluencies were in some way unbelievable. Arnold et al. (2007) demonstrated that participants need to 'believe' that speaker difficulty underlies disfluency for an effect on prediction to be observed. Whilst great care was taken in the recording and editing to produce utterances which sounded as natural as possible, it is inevitable that artificial, studio-recorded sentences do not sound exactly as they might in spontaneous conversation. In debriefing, there was no indication that participants were aware of the editing, where the target word had been spliced onto each sentence. Whilst considering the effect of editing, and whether this may account for the lack of N400 attenuation, it should also be mentioned that the attenuated N400 effects reported in Corley et al. (2007) and MacGregor et al. (2010) were also elicited using recorded stimuli, and that these two experiments used the same

recordings. If artefacts of recording or editing could be responsible for the failure to find an N400 attenuation, they could similarly be responsible for the reported detection of an N400 attenuation.

If some artefact of the stimuli (such as unbelievable disfluencies) prevented listeners from updating their predictions following disfluent fillers, thus failing to elicit an N400 attenuation, this might suggest that attention-raising, and prediction-altering depend on different mechanisms. Collard et al. (2008), demonstrated that disfluent fillers raised attention to the speech stream. Based on the memory advantage we have demonstrated for words affected by disfluency or delay, it appears that in the experiments reported in this thesis, attention was also raised by both disfluency and delay. This memory advantage has been observed in the absence of detectable changes in prediction strategies. It is possible that whilst attention may be directly sensitive to the temporal and prosodic disruption of disfluency, listeners prediction strategies may depend more on learned patterns and perspective taking in response to perceived speaker difficulty.

It may be useful to repeat one or more of the experiments reported in this thesis with the audio recordings used in Corley et al. (2007), to investigate whether the N400 attenuation they report can be replicated, and particularly whether it can be replicated in the presence of a third, noisy interruption condition. If attenuated N400 effects were revealed, then this would suggest that these effects depend on the specific stimulus recordings used. This would also permit investigation of whether the same attenuated N400s could be elicited by noisy interruptions, and specifically, whether there was any dissociation between raised attention and prediction effects.

Having considered reasons why these experiments may have failed to detect an N400 attenuation, if one exists, and how this could be further investigated, we have to consider

Publication	Disfluency	Evidence for disfluency-attenuated N400 effect?
Corley et al. (2007)	*er*	✓
MacGregor, (2008)	silent pause	✗
MacGregor et al. (2010)	silent pause	✓
MacGregor et al. (2009)	repetitions	✗
MacGregor, (2008)	repairs	✗

TABLE 10.1: Summary table of previous studies investigating N400 attenuation for words affected by disfluency. Although previous authors have concluded that where N400 attenuation was not found, this was a function of the type of disfluency used, it is also possible that N400 attenuation in response to disfluency generally is not always robust.

the notion that the N400 attenuation in response to disfluency may not be robust. As was discussed in Section 7.5.1, attenuated N400 effects following disfluency are not necessarily prevalent in the literature. Where previous studies have failed to find attenuated N400 effects, they have typically assumed that this is due to the type of disfluency used, and so concluded that N400 effects differ between fillers, silent pauses and repetitions (Corley et al., 2007; MacGregor et al., 2009, 2010; MacGregor, 2008). An overview of experiments investigating N400 attenuation and disfluency is given in Table 10.1. An alternative explanation would be to consider that attenuated N400 effects may not respond systematically to disfluency types, but may be variable depending on participant cohort, audio recording quality, or other factors.

It is not clear that disfluency having a variable effect on the N400 effect is necessarily the same as disfluency's effect on prediction not being robust. A number of studies have indicated that listeners' predictions for upcoming items are informed by hearing a speaker become disfluent (*c.f.* Arnold et al., 2004; Arnold et al., 2007; Barr and Seyfeddinipur, 2010). It may be that the N400 effect is simply the wrong measure, particularly if the size of the N400 effects is simultaneously influenced by changing prediction and changing attention. It seems possible that while disfluency may alter the processing of subsequent words, leading to the N400 attenuation reported in some studies

(Corley et al., 2007; MacGregor et al., 2010), the increase in attention engendered by disfluency may increase the amplitude of the N400 effect (Otten et al., 1993). Thus it is possible that the effects of disfluency appear to cancel each other out, if the N400 is used as a dependent measure. Alternatively, the N400 may simply not be sensitive to cues communicated in disfluency.

A further option to consider is that disfluency may not affect lexical prediction in a reliable and robust manner. Whilst it might be said that the majority of experiments studying disfluency and the N400 have focussed on lexical prediction, some other disfluency experiments have focussed on referential prediction. This is perhaps most apparent when eye- and mouse-tracking studies (e.g. Arnold et al., 2007, 2004; Barr, 2001) are considered, in which participants respond to picture sets or visual world images. It might be suggested that this type of referential prediction is more sensitive to disfluency than lexical prediction, and that this difference in study design should explain the consistent findings of disfluency effects in studies which have not relied on the N400 as a dependent measure.

However, before moving to conclude that the failure to find attenuated N400 effects following disfluency was because the studies in this thesis did not take referential prediction into account, it should be remembered that the unpredictable words were not 'alternative names' for items, in which the lexical choice would be surprising to the listener, but instead were referents not predictable from the previous discourse. Thus it is not clear that the studies in this experiment have relied entirely on lexical prediction. Listeners maintaining a model of discourse keep a record of previously mentioned items and build expectations for upcoming referents and topics, even if these referents are not given lexical names. They may suppress these expectations following disfluency. This seems to reflect closely the way participants were expected to behave in the current

experiments; although the sentences used in the current experiments led to clear lexical predictions about "the most likely word" to complete the sentence, it was not only the lexical prediction that was violated by the unpredictable targets, but also referential predictions about the kind of topic that would be mentioned. If it is not the case that the current experiments relied entirely on lexical prediction, and that this can explain the difference in the apparently robust findings of visual world studies and N400 based studies, then the possibility remains that disfluency effects are only strong enough to be reliably detected when the set of possible referents is closed, as discussed in Section 7.5.1.

Effect Size and the Risk of Publication Bias

One discussion which cannot be overlooked in the context of the current findings is the risk of publication bias within the field of disfluency. Table 7.4 makes it painfully apparent that in the present experiments we are looking at very tiny effects within relatively low-powered experiments. This raises two primary concerns. Firstly, if the effect of disfluency on prediction is really as tiny as it appears, then it is not clear that it should be considered to be important.

Secondly, there is an inherent danger in seeking tiny effects within low-powered experiments with complicated designs, especially when corrections for multiple comparisons are generally not carried out. In this environment, there is a significant danger of unreplicable positive findings heavily influencing the field. This is particularly problematic given that researchers are often reluctant to publish findings which fail to replicate previous experiments, and furthermore, that journal editors and reviewers will often consider such findings to be unpublishable null-results. As discussed by Ferguson and Keene (2012), this tendency to publication bias appears to be stronger in psychology than in

'hard' sciences, and may be stronger still in smaller or newer fields. Verbal and informal discussions at conferences throughout the course of this work have revealed a number of researchers within the field of disfluency who have failed to find expected disfluency effects, and quietly swept these studies into the proverbial filing cabinet. It is entirely possible that reported disfluency effects in the literature do not reliably replicate, but that studies questioning the validity of these effects have simply not been published.

10.3.2 How Might Attention Affect Speech Comprehension?

Could raised attention to the speech stream facilitate listeners in comprehending difficult material? If attention, which has been succinctly defined as the allocation of cognitive resources (Pashler, 1998a, 1998b), is increased, then this may increase the ease of accessing the meaning conveyed by the sentence-final unpredictable word, particularly if one thinks in terms of a model of lexical access by spreading activation. Increased allocation of cognitive resources may assist listeners in comprehending the broad spectrum of problematic material which may be heralded by disfluency, including not only semantically incongruous words, but also low frequency words, mis-pronounciations, false starts, and grammatical anomalies. Increased attention may also, for example, facilitate reassessment of the original meaning of sentences. Attentional modulation may underlie other, related findings, such as that of Bailey and Ferreira (2003), in which listeners heard garden path sentences, requiring a degree of re-parsing to create an acceptable representation. It seems possible here that increased attention, and so an increased allocation of resources, would facilitate successful re-parsing. This would provide a context for their finding that parsing was more successful for sentences containing either a disfluency or a noisy interruption — two conditions which, based on the work of this thesis, appear to raise listeners' attention to the speech stream.

The exact mechanism by which attention may or may not facilitate the integration of unpredictable material is unclear. In order to speculate on this, it is helpful to consider how comprehenders respond when presented with an unpredictable word. If lexical access is automatic, and not directed by semantic context (Swinney, 1979), then it is not accessing the meaning of unpredictable words which causes more effortful comprehension for listeners, although it is possible that increased attention would aid listeners in more quickly selecting a nuance of the target word which fits with the prior context, and rejecting irrelevant meanings.

If raised attention can help facilitate the integration of difficult material, then there is still a disfluency-as-signal question to be asked: Do listeners, hearing a disfluency, 'deliberately' increase the attention paid to the speech stream in preparation for upcoming difficult material? The results in this thesis would suggest not. It appears that a noisy interruption, which should not necessarily be predictive of upcoming difficulty, can also raise attention to the same extent. In this case, then it appears that it is the temporal and prosodic disruption introduced by the disfluency which causes an automatic attentional change, rather than a controlled process.

10.3.3 What Can We Conclude About Prediction?

Corley et al. (2007) found an attenuated N400 for words which had been preceded by disfluency, along with a memory advantage for the same words, particularly those words which had been predictable at presentation. They interpreted this as indicating that the presence of a disfluency had made the predictable words more difficult to integrate into the semantic context, rather than making the integration of the unpredictable words easier. They posited that the more effortful processing required to integrate the predictable words which had been affected by disfluency led to a long-term memory benefit.

The experiments reported in this thesis have demonstrated memory benefits in the absence of evidence of processing changes, as evidenced by the N400 magnitude. It is worth mentioning here that a failure to find an effect is not strong evidence against it's existence, but may simply reflect weaknesses of the experiment or stimuli. However, if there is a true dissociation between prediction, as indexed by the N400 magnitude, and subsequent memory, then it would mean that subsequent memory performance cannot necessarily be used to make inferences about online processing.

Returning then, to the question of how disfluency affects prediction, we cannot necessarily infer that the processing predictable words is made more effortful by the presence of a disfluency, as suggested by Corley et al. (2007). Eye- and mouse- tracking studies (Arnold et al., 2003, 2004; Barr, 2001; Barr & Seyfeddinipur, 2010) have pointed to active prediction of unexpected referents following disfluency (albeit in closed referent sets). Although this indicates a preference for the unpredictable, it does not clearly demonstrate whether this result depends on facilitated access for unpredictable referents, or rejection of the most likely referent, and this mechanism remains a subject for further work.

10.4 Investigating Linear Mixed-Effects Models for ERP Analysis

A second aim of this thesis was to investigate how linear mixed-effects modelling might be used to allow analysis of ERP data in psycho-linguistic experiments. Linear mixed-effects models allow the incorporation of subjects and stimulus items as fully crossed random effects. Whilst this in itself is desirable within psycho-linguistic experiments, it is not clear that the approach taken in this thesis has been entirely successful. Of the

four data-sets analysed using mixed-effects modelling, none allowed convergence with a fully specified random effects structure, and so a forward model selection procedure was necessary. Barr et al. (2013) warn that such model selection can lead to anti-conservativity, and that within the vast number of possible model structures, it is easy to design a model-selection procedure which does not lead to the best fitting model for the data.

Despite the risk of anti-conservativity described by Barr et al. (2013), the forward model-selection procedure described in Chapter 9 arrived at final models which, for at least one dataset, appear to have been more conservative that the traditional ANOVA analysis pursued in Chapters 5-8. Whereas the analysis of the N400 data in Chapter 5 revealed a significant interaction of predictability and fluency, no such interaction was detected using mixed-effects modelling. This may be an example of the incorporation of stimulus items as random effects guarding against Type I errors.

The difficulty of achieving convergence for the models attempted here may point towards a poor fit between the method used and the data obtained in the experiments. The models described in this thesis sought to fit data in a linear fashion. It is possible that a linear fit is not appropriate, and if this is the case, then it may lead to convergence failures as the model is insufficiently able to describe the data submitted to it. In principle, it would be possible to create non-linear models, for example, fitting polynomial functions. However, given the complexity of the data described, and the severe convergence difficulties encountered in fitting linear models, it does not seem likely that fitting more complex models would be practical, within current computing limitations.

Linear mixed-effects models are not the only type of mixed-effects models which may offer solutions to ERP researchers looking to model crossed random effects. Further

research could focus on how other types of models, such as Generalised Additive Mixed Models (GAMMs) (Wood, 2006), whose flexibility may offer more appropriate tools to fit the data generated in psycho-linguistic ERP experiments.

Alternatives are also currently being investigated elsewhere. Smith and Kutas (in press) have recently developed an analysis framework using linear regression rather than averaging to estimate the ERP at each time point and each electrode. They argue that the traditional method of averaging ERPs is really performing least-squares regression, and that by extending this to a full linear regression approach, a wider range of situations can be handled, opening the door to a wider range of potential paradigms, including both linear and categorical predictors. Thus Smith and Kutas' regression ERP (rERP) framework fulfils some of the aims of the mixed-effects modelling approach attempted in this thesis. Smith and Kutas' method is similar to Janssen's (Janssen et al., 2013) approach, in so far as he also modelled the data at each time-point in the EEG recording. It seems likely that, in common with Janssen's approach, Smith and Kutas's rERP framework would require access to powerful computing resources to be viable. However, one of the key outcomes of Chapter 9 was the finding that even a 'lightweight' approach to mixed-models ERP analysis was somewhat impractical without access to powerful cloud-based computing, and so it is worth considering whether the regression approach outlined by Smith and Kutas may give better value in terms of detail in results for computing power.

To summarise the findings of this thesis with regards to using linear mixed-effects models to analyse ERP data, the approach used here has given mixed sucess. Whilst it may have exposed anti-conservativity in ANOVA analysis, the model selection procedure demanded by the convergence failure of maximal models does not guard against anti-conservativity as fully as might be desired. Thus it is still far from an ideal analysis

structure, and whilst it may offer some benefits over traditional ANOVA analysis, it is not clear that this benefit justifies the increase in computing demands, and time — from under one second to run ANOVA analysis, to 8-9 weeks to obtain a result using mixed models.

10.5 Summary

The experiments reported in this thesis have sought to systematically investigate how delay may contribute to the effects of disfluency on comprehension and subsequent memory. Whilst we have not revealed evidence for a robust effect of disfluency or delay on the N400 effect, memory advantages for words affected by disfluency and delay suggest that attention may have been raised. This would support an account of disfluency in which attention is raised in response to the temporal and prosodic interruption, whereas predictive effects, if discernible in the N400, may be more sensitive to naturalistic cues and a sympathetic response to speaker difficulty. As such, different aspects of disfluency may affect different parts of the listener's comprehension processes.

Appendix A

List of Stimuli

Below are listed all of the experimental stimulus sentences used in the experiments reported in this thesis. The final word in each utterance comprises the target word. Predictable words are presented in bold font, unpredictable counterparts in italics. Sentences are presented in pairs, with the predictable target of one sentence being the unpredictable target for the other. If a sentence was presented in a disfluent or interrupted condition, the disfluency (*er*) or interruption (beep, cough) was inserted directly before the utterance final target word.

List of Stimuli

Jack's already left for Rome He's always wanted to visit the capital of **Italy**/*Greece*

I mix up the Mediterranean Countries But I'm fairly sure Athens is the capital of **Greece**/*Italy*

Everybody knows that Harry Potter is a young **wizard**/*pig*

I thought everybody knew that a piglet is a young **pig**/*wizard*

I think a lot of people go to university because they like the student **lifestyle**/*discount*

If you take your student card, you'll get a ten percent student **discount**/*lifestyle*

I'm looking forward to my sister's birthday but I want to get a new dress for the **party**/*funeral*

I've got to go shopping on Wednesday, my grandma died and I need a black dress for the **funeral**/*party*

Tomorrow is my aunt's birthday, and I want to buy her a **present**/*ticket*

She'd love to go to that concert, so I decided to buy her a **ticket**/*present*

That shop is filthy They should think about hiring a better **cleaner**/*excuse*

He'll never believe the dog ate my homework I need a better **excuse**/*cleaner*

We've not booked any accommodation so when we arrive we need to find somewhere to **stay**/*play*

I've brought a football so we just need to find somewhere to **play**/*stay*

Ive been at the gym and I feel all sweaty I'm going to have a **shower**/*Kit-kat*

Well, you know what they say - have a break, have a **Kit-kat**/*shower*

Well, you can retire when you want, but you have to be sixty to get a **pension**/*job*

There are so many well qualified graduates it's hard to get a **job**/*pension*

Before you start baking, remember to warm up your **oven**/*muscles*

They say to stretch before exercise you should warm up your **muscles**/*oven*

Last Thursday was our anniversary you know, so Tim and I went to the new Italian restaurant for a nice **meal**/*walk*

Yesterday afternoon was beautiful, and I fancied getting out of the house, so I took my dog out for a **walk**/*meal*

Jimmy's parents have been told he can't hear well - he's going to a school for the **deaf**/*blind*

Apparently Toni is partially sighted, she's going to a school for the **blind**/*deaf*

To wipe our fingers on at mealtimes we always use **napkins**/*conditioner*

To keep my hair soft, after shampooing, I always use **conditioner**/*napkins*

We climbed to the top of the hill to see the **view**/*dentist*

Have you still got toothache? I think you should see the **dentist**/*view*

Margaret is always invited for Christmas Its like she's a member of the **family**/*team*

Josh is so proud of his new football strip he loves being a member of the **team**/*family*

I wanted to wash my hands but there wasn't even a bar of **soap**/*chocolate*

I'm a terrible snacker Just the other day I ate a whole bar or **chocolate**/*soap*

I'm looking forward to lying and soaking in a long hot **bath**/*summer*

It's been quite warm so we're hoping for a long hot **summer**/*bath*

I wanted to wash my hands but they were sticky and I didn't know how to turn on the **tap**/*heating*

A lot of student flats are cold and damp because they can't afford to turn on the **heating**/*tap*

Everyone's got bad habits, I'm always biting my **nails**/*tongue*

That coffee was too hot - I've burnt my **tongue**/*nails*

They say that to enjoy wine, you need to pour it into a proper **glass**/*gentleman*

Aaron has grown up to be very courteous! He's turned into a proper **gentleman**/*glass*

I spent ages in the library, but I couldn't really find anything I wanted to **read**/*wear*

I took ages getting ready to go out cos I couldn't decide what I wanted to **wear**/*read*

Are you excited about blowing out the candles on your birthday? You know, Granny Sue's made your **cake**/*appointment*

It's all been sorted out and the doctor will see you tomorrow I rang them this morning and made your **appointment**/*cake*

She's got high heeled ones, flat ones, boots, wellies, really every kind of **shoe**/*dog*

I can't remember what it looked like It barked so I guess it was some kind of **dog**/*shoe*

Since the accident he's had such a stiff neck - he can scarcely move his **head**/*car*

I don't think I'm going to be able to get out of this parking space Perhaps we can ask Niall to move his **car**/*head*

I couldn't find my classes cos they had been moved to different lecture **theatres**/*notes*

I can't remember that formula, I need to check in my lecture **notes**/*theatres*

The court found him guilty He'll be going to **jail**/*school*

Phillip's nearly five It won't be long before he's going to **school**/*jail*

None of them have made any films before They're a bunch of **amateurs**/*flowers*

Mrs Houghton is leaving, so P6 have bought her a bunch of **flowers**/*amateurs*

To add a final touch to my armchair, I'm going to buy a **cushion**/*dress*

She's looking forward to her wedding, but she still needs to buy a **dress**/*cushion*

I think he'll be in prison a long time with such a harsh **sentence**/*weather*

We couldn't leave the house for days due to the harsh **weather**/*sentence*

I really like indoor raquet sports, that's why I love **squash**/*baking*

I'm very good at making cakes and bread, that's why I love **baking**/*squash*

We're not allowed much in the exam - pen, pencil, and a bottle of **water**/*wine*

We ought to take something to the Morris' - it's not every day you get invited for dinner Perhaps a bottle of **wine**/*water*

If everyone has their coffee white, we'll need some more **milk**/*bread*

I can make everyone sandwiches, but we'll need some more **bread**/*milk*

Jillie used to think there were fairies at the bottom of the **garden**/*pile*

I was looking for the report with my name on it, it was right at the bottom of the **pile**/*garden*

After work on Fridays we usually head for the nearest **pub**/*exit*

If you hear the fire alarm, please move to the nearest **exit**/*pub*

We crossed the lake with Steve in his **boat**/*office*

I had a meeting with the director in his **office**/*boat*

I could tell the men had been in a serious meeting, cos they were wearing **suits**/*wellies*

The kids kept going in the puddles but their feet were dry cos they were wearing **wellies**/*suits*

I went to see Sam's team play but they lost the **game**/*ring*

They got engaged last week but she's already lost the **ring**/*game*

Dave brought the cake in, and Millie blew out all the **candles**/*windows*

The bomb wrecked the shop and blew out all the **windows**/*candles*

Its been really busy today, and the phones haven't stopped **ringing**/*working*

I thought the boiler had been fixed but now its stopped **working**/*ringing*

I hate it when the dentist has to hollow out my tooth with the **drill**/*flu*

Wendy has been really ill She's been in bed for three days with the **flu**/*drill*

This meat is tough, please could you pass me a **knife**/*tissue*

I think I'm going to sneeze Please could you pass me a **tissue**/*knife*

When you get there, don't forget to wipe your feet on the **mat**/*shelf*

In my office the books are arranged alphabetically on the **shelf**/*mat*

When I got in he'd fallen asleep in front of the telly Just sprawled out on the **couch**/*path*

You should be able to find your way through the woods, so long as you stay on the **path**/*couch*

I love it when I wake up and the birds are **singing**/*talking*

Its distracting in lectures when the others are **talking**/*singing*

Its getting late, so one of us had better take your granny home, or at least walk her down the **road**/*aisle*

It was such a beautiful wedding, but I could scarcely stop myself crying when Steve was walking her down the **aisle**/*road*

How can they imagine that makes you smell good? Smells like there's all kinds of chemicals in that **perfume**/*magazine*

There's not much intellectual content Just all kinds of celeb gossip in that **magazine**/*perfume*

I've got a deadline to meet for tomorrow, so I can't afford to waste **time**/*paper*

I always reuse envelopes cos I can't bear to waste **paper**/*time*

Apparently Barbara's husband died last week We ought to send a **card**/*bill*

Lawyers don't work for nothing you know - he'll certainly send a **bill**/*card*

Well I love hearing the news first thing when I get up, so I always listen to the **radio**/*teacher*

Now, what are you going to do in your lesson? You're going to sit still and listen to the **teacher**/*radio*

They've been living together for ages so it was no surprise when they decided to get **married**/*divorced*

I think it was really hard for the kids when they heard their parents were going to get **divorced**/*married*

It was to noisy to have a conversation on the dance floor - we had to shout to be **heard**/*seen*

Its dark - you'll need to make sure you wear good reflective clothing to be **seen**/*heard*

They have a very rare Dickens book, a first **edition**/*choice*

I've decided to go to Aberdeen, though its not my first **choice**/*edition*

When I visited Cambridge, I got shouted at for walking on the **grass**/*pavement*

I don't understand how that car hit him He was walking on the **pavement**/*grass*

Watching that cooking programme has made me feel **hungry**/*sad*

Seeing her just leave like that has made me feel **sad**/*hungry*

My sister had a skiing accident and she broke her **leg**/*promise*

She said she wouldn't cheat on him but she broke her **promise**/*leg*

We always paid the cash for the rent direct to our **landlord**/*teachers*

Our school was great We really do owe our success to our **teachers**/*landlord*

Kids love cartoons where the hero needs to defeat an evil **villain**/*eye*

I must have done something to annoy her She's giving me the evil **eye**/*villain*

Terry's bored of being ill, he says there's nothing to watch on the **tv**/*wall*

Some of these paintings are amazing - I can't stop looking at that picture on the **wall**/*tv*

I love going to the seaside, especially having fish and **chips**/*ice-cream*

Do you remember birthdays when we were little? I used to love jelly and **ice-cream**/*chips*

Mum hated the band, but then she's never like my taste in **music**/*clothes*

David thought the dress was too short, but he's never liked my taste in **clothes**/*music*

I want to travel and see new places, meet new **people**/*animals*

They're extending Edinburgh zoo and getting in some new **animals**/*people*

It wasn't til I'd put on my hat and scarf I realised I lost one of my **gloves**/*wheels*

My bike got wrecked in the accident I lost one of my **wheels**/*gloves*

I got a card with a picture of a lucky black **cat**/*tie*

Apparently the dress code is black **tie**/*cat*

I had lunch quickly on my way to work All I ate was a small **sandwich**/*desk*

My office is really bare There's just a chair and a small **desk**/*sandwich*

If your face feels dry from the cold, you should apply a **moisturiser**/*bandage*

If its a serious open wound, you ought to apply a **bandage**/*moisturiser*

If you're coming home late then stay in a group and keep **together**/*fit*

The doctor's advised me to lose weight, so I've got to go to the gym and keep **fit**/*together*

Marion started crying about her haircut as soon as she left the **hairdressers**/*vets*

Poor Sue she's upset about her dog! She burst into tears as soon as she left the **vets**/*hairdressers*

I love the feel of this restaurant - especially the fresh flowers on all the **tables**/*graves*

Its nice to visit the churchyard at Easter, when there's flowers on all the **graves**/*tables*

Ben's so stubborn, he'll never admit that he's **wrong**/*drunk*

Jamie's only had two beers, but I'm pretty sure he's **drunk**/*wrong*

I need to get the washing machine fixed so I can do some proper **washing**/*exercise*

I've realised that to really get fit I need to do some proper **exercise**/*washing*

Before I go to South America, I really want to learn to speak **spanish**/*properly*

When my sister puts on a silly voice mum always tells her to speak **properly**/*spanish*

At Christmas, mum cooked us a huge **turkey**/*screen*

Max wants to go to that cinema because they have a huge **screen**/*turkey*

At weddings, everyone wants to kiss the **bride**/*ground*

When prisoners arrive home after years of exile, they often kiss the **ground**/*bride*

I'm worried about the move to Paris, mostly 'cos I don't speak **french**/*chinese*

I don't think they'll have difficulty settling in Hong Kong - he and his wife both speak **chinese**/*french*

None of his brothers like football, they prefer **rugby**/*blondes*

Well, you know what they say, gentlemen prefer **blondes**/*rugby*

She's been having driving lessons for months, but she's still nervous about taking the **test**/*class*

It's hard for the teachers when there are so many children in the **class**/*test*

Don't go running off as soon as you've eaten At least offer to help load the **dishwasher**/*van*

We're moving at the weekend, but the men are coming early Friday morning to load the **van**/*dishwasher*

In the maths exam on Thursday we're not allowed to use a **calculator**/*shovel*

There was so much snow that to clear the drive we had to use a **shovel**/*calculator*

I can't wait to go to Crete I just keep thinking about sitting on the **beach**/*floor*

There were no more chairs in the lecture so we ended up sitting on the **floor**/*beach*

I stood up to speak but then my mind went totally **blank**/*random*

The list had no order and seemed totally **random**/*blank*

She was by far the best in her age group and broke loads of **records**/*plates*

I was a rubbish waitress I was too clumsy, I broke loads of **plates**/*records*

It was pouring - I borrowed her umbrella so I wouldn't get **wet**/*burnt*

It was so hot yesterday - I wore a hat so I wouldn't get **burnt**/*wet*

The sitting room is really cold so I think you should light a **fire**/*cigarette*

We're in a smoking area so now you can light a **cigarette**/*fire*

The meeting was pretty boring Rob kept looking at his **watch**/*reflection*

He was pleased how he looked, he kept looking in the mirror at his **reflection**/*watch*

I was so nervous about the exam that my heart was pounding in my **chest**/*garage*

I was so happy I kept going to look at the new car in my **garage**/*chest*

My translation into French wasn't very good It had lots of **mistakes**/*food*

I hope everyone turns up for Christmas dinner this year, otherwise we're going to be left with lots **food**/*mistakes*

My grandad used to keep all his cricket trophies in a glass **cabinet**/*bowls*

My granny always used to serve ice cream in really small glass **bowls**/*cabinet*

This Christmas I want to decorate my living room with a **tree**/*roll*

At the restaurant I ordered a bowl of soup with a **roll**/*tree*

I went to open the front door, but I couldn't find my **keys**/*purse*

Sarah offered to pay for my lunch cos I couldn't find my **purse**/*keys*

I'm gasping for a pint Lets go to the pub for a **drink**/*burger*

I'm starving - why don't we go to MacDonalds for a **burger**/*drink*

I'm scared of big animals, so I've never learnt to ride a **horse**/*motorbike*

I can't afford a car, so I'm going to get a licence to ride a **motorbike**/*horse*

To go out in the snow you need to keep your head warm. Put on a **hat**/*suit*

I think it's going to be a formal meeting You should probably wear a **suit**/*hat*

Don't ask me! If you want to know the time, look at the **clock**/*calendar*

I'm not sure what we're doing on Thursday - I'll look at the **calendar**/*clock*

There's a shortage of good houses in my parents' village because so many people have second **homes**/*degrees*

A BA or BSc doesn't count for that much because so many people have second **degrees**/*homes*

I fancied soup for lunch but couldn't find anything to open the **tin**/*curtains*

First thing when I get up I always make the bed and open the **curtains**/*tin*

After we moved, it took ages to unpack all the **boxes**/*children*

When I worked as a dinner lady, I had to make sure there was enough to feed all the **children**/*boxes*

Maria is constantly ill with something She always has a **cold**/*banana*

Clare loves her fruit, after lunch she always has a **banana**/*cold*

When the children come in from school, they love to have a glass of **juice**/*whiskey*

No trip to a Highland distillery is complete without a glass of **whiskey**/*juice*

What grade you get at the end of the year depends on how well you do in your **exams**/*room*

A parcel arrived for you I put it in your **room**/*exams*

The couple finally decided to fight out the divorce settlement in the **court**/*sunlight*

According to the myths, vampires can't go out in the **sunlight**/*court*

I'm trying to keep mum's birthday present a surprise I wish my sisters would stop dropping **hints**/*litter*

After all the effort the children went to to tidy up the park! I wish people would stop dropping **litter**/*hints*

I left the ice cream out of the freezer too long now its/The ice-cream's been out of the freezer too long, now its **melted**/*frozen*

The milkman came really early even though it was so cold, now the milk's **frozen**/*melted*

Birds wouldn't be able to fly if they didn't have a pair of **wings**/*trainers*

Jo's a keen runner so he spent most of his birthday money on a new pair of **trainers**/*wings*

At the end of break Mrs Beeston will choose someone to ring the **bell**/*Fire-brigade*

I don't think that smoke is from a bonfire - get someone to ring the **Fire-brigade**/*bell*

Your staircase has a beautifully carved **bannister**/*throne*

The king was sitting on a beautifully carved **throne**/*bannister*

Was that your phone? I think you've got a **text**/*dictionary*

When I'm not sure how to spell a word, I always check in the **dictionary**/*text*

You've been working really hard for hours Perhaps you should take a **break**/*tablet*

Well I first noticed it yesterday I had a huge headache so I decided to take a **tablet**/*break*

I never really got on with French, I'm a bit slow when it comes to **languages**/*numbers*

I'm not much good at maths, I get a real blank when it comes to **numbers**/*languages*

My granny used to love drinking tea from a beautiful porcelain **cup**/*sink*

We're having a new bathroom - with an antique porcelain **sink**/*cup*

I'm a bit of a lightweight I can never drink more than two **pints**/*options*

We don't have to take either of those solutions There are more than two **options**/*pints*

To make a glass of fresh juice for breakfast you'll need a big **orange**/*extension*

The house is too small and needs a big **extension**/*orange*

It will be lovely to leave the car and explore the area on a **bike**/*show*

He's always wanted to be on telly so he was pleased to be interviewed on a **show**/*bike*

I wonder if its true what they say, that when someone's really terrified you can smell the **fear**/*coffee*

He's living in a dream he needs to wake up and smell the **coffee**/*fear*

I thought it would be a nice day, but now the sky's full of **clouds**/*sun*

I had to move to the shade, cos it was too hot to sit in the **sun**/*clouds*

They wanted to keep an eye on everyone involved in the accident, so we had to spend the night in a **hospital**/*dark*

Make sure Emma eats all her carrots Tell her they'll help her see in the **dark**/*hospital*

There was a big mess and we all ran around like headless **chickens**/*neighbours*

We can't come, tomorrow we're going to pop next door and visit the **neighbours**/*chickens*

I've got to remember to return that book when I go to the **library**/*police*

Look, I think someone's broken in next door I think we should call the **police**/*library*

Emma will probably want you to read her the story about the princess trapped at the top of the **tower**/*plane*

I hate long flights, you spend all day on the **plane**/*tower*

Our stop is the one before the bridge Remember to tell the driver when you want to get off the **bus**/*record*

I agreed to talk to the journalist but there were some things I had to say off the **record**/*bus*

If you listen in the woods at night you might hear the hooting **owls**/*butterflies*

I don't like insects, but I'm not afraid of moths or colourful **butterflies**/*owls*

Pauline had geography today, and learnt that Great Britain is a big **island**/*egg*

I'm really looking forward to the holidays and eating my Easter **egg**/*island*

I'm exhausted I think I'll have an early **night**/*income*

Since I was made redundant, we've had to survive on one **income**/*night*

I could only just get it all in the wheelie bin - and there was no way I could shut the **lid**/*door*

It really annoys me when he just walks in without knocking and then doesn't even shut the **door**/*lid*

Have you seen the pictures from Dave and Sarah's wedding? I had no idea Alan was such a talented **photographer**/*footballer*

They say Ronaldinho is the world's most skillful **footballer**/*photographer*

No, that band was pretty terrible Seriously, after two songs they'd completely emptied the **place**/*bin*

He makes so much fuss about helping around the house All I did was ask him to empty the **bin**/*place*

Put the trap behind the kitchen sink Maybe you'll catch another **mouse**/*fish*

He's been fishing again today - he's hoping to catch another **fish**/*mouse*

We both love art, so when we were in Paris, we spent the whole day at the **gallery**/*airport*

The flight was delayed by hours - we spent the whole day at the **airport**/*gallery*

Even now she's grown up, the bit where Bambi's mother dies always makes her **cry**/*laugh*

She loves comedy shows, and stuff, anything that makes her **laugh**/*cry*

We wanted a souvenir of our trip, so we asked a tourist to take a picture of **us**/*himself*

He spent ages rewatching the scene he was in, trying to spot **himself**/*us*

You've cut your finger! Wait a minute I've got a **plaster**/*camera*

If we're going to take photos, we need to remember to take a **camera**/*plaster*

I heard him in the corridor and I recognise his **voice**/*smell*

The house has been empty for years, as soon as I got in I noticed a strange **smell**/*voice*

I always carry so many things in it, I'd feel lost without my **bag**/*lorry*

All the fairground rides arrived in a big **lorry**/*bag*

When I take notes in lectures I always use a nice **pen**/*chalk*

In the victorian days, teachers used to write on the blackboard with **chalk**/*pen*

Land is at a premium in Manchester city centre, that's why the council have started building sky **scrapers**/*light*

We get pretty good views from the attic cos of the sky **light**/*scrapers*

My boyfriend took flight lessons, now he can fly small **planes**/*kites*

It was windy, so we took the kids up to the park to fly a **kites**/*planes*

If the mash is going to be ready at the same time as the meat you'd better hurry up and boil the **potatoes**/*equipment*

After doing an operation, surgeons sterilise and boil the **equipment**/*potatoes*

I can't get it to work I think there's something wrong with the keyboard on the **computer**/*piano*

John takes music seriously. Today, he's got someone coming to tune the **piano**/*computer*

I think a really good typist can manage over a hundred words a **minute**/*week*

They say that to stay fit, you should exercise three times a **week**/*minute*

I didn't realise that when you had a PhD you'd be called a **doctor**/*cheat*

She always tries to copy in exams cos she's a **cheat**/*doctor*

I love Colin Firth He's my favourite British **actor**/*passenger*

The plane was delayed, cos we had to wait for one late **passenger**/*actor*

I'm going to be late this morning, could you pass on a **message**/*baton*

In relay races, runners have to pass on a **baton**/*message*

I saw a photo in the paper, and I recognised his **face**/*knee*

Walking downhill hurts, I think I must have damaged my **knee**/*face*

We went to the beach but forgot to take buckets and **spades**/*sand*

The beach had miles and miles of golden **sand**/*spades*

Ollie wanted to give Grannie a present, so he drew a **picture**/*sword*

The knight killed the dragon with his **sword**/*picture*

I think you have to order online You need to go on their **website**/*honeymoon*

Venice is a romantic place for newlyweds to go on their **honeymoon**/*website*

I need some help sorting out documents for Moscow I need someone who speaks fluent **russian**/*German*

Sylvia used to live in Germany, so she speaks fluent **German**/*russian*

I can't really afford a new kitchen, I haven't got enough **money**/*evidence*

They've already arrested someone, I think they'll charge him when they've got enough **evidence**/*money*

I wanted to post the letter but I haven't got a **stamp**/*clue*

My supervisor wants to know what the results mean but I haven't got a **clue**/*stamp*

I do like living in town, but I'd love to move to the **country**/*city*

My brother is tired of living somewhere so quiet, he wants to move to the **city**/*country*

Honestly I couldn't believe the way they all turned on him at the meeting It was like a pack of **wolves**/*cards*

Take something to do on the journey, maybe a game or a pack of **cards**/*wolves*

I'm not sure if we've got any more yoghurt Check in the **fridge**/*microwave*

If you're in a rush, it only takes a few minutes to warm up soup in the **microwave**/*fridge*

Do you think they do ice cream? Let's ask the waiter for the **menu**/*stapler*

I need to attach these sheets to my notes Could you pass me the **stapler**/*menu*

Its like when they tried to close the pit and all the miners went on **strike**/*holiday*

The kids were on half-term last week so we all went on **holiday**/*strike*

My skirt is too big I'll have to nip it in with a few safety **pins**/*goggles*

Make sure you protect your eyes when you're working in the chemical lab - wear your safety **goggles**/*pins*

Even though he moved from Scotland years ago he's very proud of his roots, so at weddings and things he always wears his **kilt**/*jumper*

In winter its nice to feel cosy and wear a warm woolly **jumper**/*kilt*

Him and Maria spotted a market for it, and decided to start a small **business**/*trend*

Now everyone's wearing their hair up like that. She seems to have started a **trend**/*business*

Every few years you might see a solar **eclipse**/*panel*

There's no generator but I can see a solar **panel**/*eclipse*

Well the conference starts on Monday, and you need to be there at least the night before You probably need to arrive by **Sunday**/*ferry*

I don't really like flying, so I usually get to France by **ferry**/*Sunday*

I guess she must be engaged, now that she's started wearing a ring on that **finger**/*tail*

Even when my dog's barking he'll still wag his **tail**/*finger*

Danielle is easy when it comes to Christmas - she loves reading I always get her a **book**/*vase*

What beautiful flowers! As soon as we get home remind me and I'll put them in a **vase**/*book*

The kids have got a swimming lesson, so I've got to drop them at the **pool**/*shops*

Could you get some milk and a packet of toilet roll on your way past the **shops**/*pool*

Most fluent English speakers speak at a rate of three words a **second**/*month*

Cos of my asthma, I've got to go for a check-up once a **month**/*second*

He's going pretty bald - he's started using a special shampoo to stop him losing any more **hair**/*weight*

She's been working so hard and getting really thin - the doctor's put her on a special diet to stop her losing any more **weight**/*hair*

I think the chef has forgotten to season the pasta, please pass me the **salt**/*sugar*

This coffee's not sweet enough. It needs some more **sugar**/*salt*

I love having my sister's kids over, but it seems I'm always picking up **toys**/*girls*

He's a massive flirt. He's always going out and picking up **girls**/*toys*

It can be hard when you're doing something simple, to think outside of the **box**/*army*

My brother really like the idea of a military life, so he decided to go into the **army**/*box*

Gemma's on monitor duty tomorrow. She's going to spend the morning helping with P1 and sharpening **pencils**/*knives*

Lisa loves cooking. Before she starts making anything she always sharpens her **knives**/*pencils*

I'll make you something quick to eat. Beans on **toast**/*cereal*

We use so much milk at breakfast, Bradley has about half a pint on his **cereal**/*toast*

I tried speaking to him but he didn't hear me, he was listening to music through his **headphones**/*ears*

I wanted to talk to my boss about it but I think it fell on deaf **ears**/*headphones*

Rory wants me to take him to the lake to feed the **ducks**/*tigers*

I'm not sure, but I would guess lions must be fairly closely related to **tigers**/*ducks*

I wonder if Neil Armstrong really was the first man on the **moon**/*stage*

I'm really nervous about singing in the musical tonight. It'll be my first time on the **stage**/*moon*

We'll book it as soon as possible. How about after dinner this **evening**/*morning*

Sorry I'm a bit late. I overslept this **morning**/*evening*

I found out why the kitchen floor was wet, there was a leak behind the washing **machine**/*basket*

I remember once we were playing pirates, so for a boat we used mum's washing **basket**/*machine*

Sam really wants a traditional front door, made of **wood**/*metal*

Cars need to be made really strong, thats why they're made of **metal**/*wood*

I would have turned on the light but I couldn't find the **switch**/*bulb*

The light's stopped working. We probably just need to change the **bulb**/*switch*

its great having a sat-nav for getting places, we don't argue about reading the **map**/*label*

Before you start taking your new pills, remember to read the **label**/*map*

I hoped she would remember how many there were, but she's already **forgotten**/*left*

I wanted to speak to Lucy before she went away but she's already **left**/*forgotten*

I ran to the shop but it was too late, it was **closed**/*dirty*

I only just cleaned the kitchen, and look at it, its already **dirty**/*closed*

It was a formal dinner, so I was pretty surprised to be on the top **table**/*soil*

The fields are very poor because the rain's washed away the top **soil**/*table*

The holiday's all booked - I just need to collect the **tickets**/*rent*

We've signed a tenancy agreement The landlord'll come on Friday's to collect the **rent**/*tickets*

Appendix B

Mixed-Effects Models Tables

Over the next few pages follow linear mixed-effects model output tables for the data described in Chapter 9.

B.1 Experiment 1 - N400 effect data (midline)

Fixed effects:				
	Estimate	std error	t value	p
(Intercept)	-0.1674	0.3395	-0.493	0.622
s(Predictability)	-2.4638	0.5085	-4.845	0.000
FluencyFluent_v_er	0.1616	0.3363	0.481	0.631
Fluencyer_v_beep	-1.1685	0.2991	-3.907	0.000
LocationCP	-0.5721	0.1691	-3.382	0.001
LocationF	1.1991	0.2357	5.087	0.000
LocationFC	0.8406	0.2075	4.051	0.000
LocationP	-0.8422	0.1809	-4.655	0.000
LocationPO	-1.0674	0.1970	-5.419	0.000
s(Predictability):FluencyFluent_v_er	-0.5367	0.5742	-0.935	0.350
s(Predictability):Fluencyer_v_beep	0.2312	0.5980	0.387	0.699
s(Predictability):LocationCP	-0.0421	0.3383	-0.124	0.901
s(Predictability):LocationF	0.6978	0.3528	1.978	0.048
s(Predictability):LocationFC	0.2897	0.3481	0.832	0.405
s(Predictability):LocationP	0.0450	0.3484	0.129	0.897
s(Predictability):LocationPO	0.3210	0.3606	0.890	0.373
FluencyFluent_v_er:LocationCP	0.4786	0.2395	1.998	0.046
Fluencyer_v_beep:LocationCP	0.5917	0.2395	2.470	0.014
FluencyFluent_v_er:LocationF	-1.6076	0.2538	-6.334	0.000
Fluencyer_v_beep:LocationF	-1.2504	0.2926	-4.274	0.000
FluencyFluent_v_er:LocationFC	-0.9641	0.2661	-3.623	0.000
Fluencyer_v_beep:LocationFC	-0.8519	0.2681	-3.177	0.001
FluencyFluent_v_er:LocationP	0.8555	0.2694	3.176	0.001
Fluencyer_v_beep:LocationP	1.0978	0.2470	4.445	0.000
FluencyFluent_v_er:LocationPO	1.0911	0.3021	3.612	0.000
Fluencyer_v_beep:LocationPO	1.4283	0.2617	5.459	0.000
s(Predictability):FluencyFluent_v_er:LocationCP	0.0349	0.4790	0.073	0.942
s(Predictability):Fluencyer_v_beep:LocationCP	-0.1174	0.4791	-0.245	0.806
s(Predictability):FluencyFluent_v_er:LocationF	-0.1085	0.4893	-0.222	0.824
s(Predictability):Fluencyer_v_beep:LocationF	0.0602	0.5147	0.117	0.907
s(Predictability):FluencyFluent_v_er:LocationFC	-0.0259	0.4840	-0.054	0.957
s(Predictability):Fluencyer_v_beep:LocationFC	0.0038	0.4938	0.008	0.994
s(Predictability):FluencyFluent_v_er:LocationP	0.0238	0.4921	0.048	0.962
s(Predictability):Fluencyer_v_beep:LocationP	-0.2681	0.4883	-0.549	0.583
s(Predictability):FluencyFluent_v_er:LocationPO	0.1505	0.4957	0.304	0.761
s(Predictability):Fluencyer_v_beep:LocationPO	-0.2703	0.4971	-0.544	0.586

TABLE B.1: Fixed effects output from the linear mixed-effects model described in Section 9.7.1 (mid-line electrodes).

B.2 Experiment 1 - N400 effect data (CPz)

Fixed effects:				
	Estimate	std error	t value	p
(Intercept)	-0.6643	0.3133	-2.120	0.034
s(Predictability)	-2.5320	0.3952	-6.407	<0.0001
FluencyFluent_v_er	0.6122	0.2915	2.1	0.0358
Fluencyer_v_beep	-0.5002	0.2198	-2.275	0.0229
s(Predictability):FluencyFluent_v_er	-0.6959	0.4287	-1.623	0.105
s(Predictability):Fluencyer_v_beep	-0.0619	0.4718	-0.131	0.896

TABLE B.2: Fixed effects output from the linear mixed-effects model described in Section 9.7.1 (CPz electrode).

B.3 Experiment 2 - LPONE data

Fixed effects:				
	Estimate	std error	t value	p
(Intercept)	0.6989	0.5435	1.286	0.198450215
StringFL	-0.3814	0.5277	-0.723	0.469683946
StringFR	-0.376	0.4827	-0.779	0.435984143
StringPR	-0.9761	0.3083	-3.166	0.001546622
s(Predictability)	-0.7524	0.3856	-1.951	0.051063745

TABLE B.3: Fixed effects output from the linear mixed-effects model described in Section 9.7.2. Data were averaged over four regions of interest, referred to here as 'string'. FL refers to the front left quadrant, FR the front right, PL the posterior left, and PR the posterior right. See Section 9.7.2 for further details.

B.4 Experiment 3 - LPONE data

Fixed effects:				
	Estimate	std error	t value	p
(Intercept)	-5.0813	2.5925	-1.960	0.050
StringFL	5.3745	1.6515	3.254	0.001
StringFR	6.7888	1.9210	3.534	0.000
StringPR	-0.3326	1.3184	-0.252	0.801
s(Predictability)	0.4607	0.7446	0.619	0.536
Fluencyer	0.0461	0.4957	0.093	0.926
Fluencyinterruption	-0.4566	0.5831	-0.783	0.434
StringFL:s(Predictability)	-0.8384	0.4504	-1.861	0.063
StringFR:s(Predictability)	-0.6511	0.4504	-1.446	0.148
StringPR:s(Predictability)	0.3547	0.4504	0.788	0.431
StringFL:Fluencyer	-0.1046	0.3154	-0.332	0.740
StringFR:Fluencyer	-0.4995	0.3154	-1.584	0.113
StringPR:Fluencyer	0.1688	0.3154	0.535	0.593
StringFL:Fluencyinterruption	-0.3893	0.3156	-1.233	0.218
StringFR:Fluencyinterruption	-0.0795	0.3156	-0.252	0.801
StringPR:Fluencyinterruption	0.1540	0.3156	0.488	0.626
s(Predictability):Fluencyer	-0.0676	0.8225	-0.082	0.935
s(Predictability):Fluencyinterruption	0.6104	0.7379	0.827	0.408
StringFL:s(Predictability):Fluencyer	2.3731	0.6315	3.758	0.000
StringFR:s(Predictability):Fluencyer	1.2742	0.6315	2.018	0.044
StringPR:s(Predictability):Fluencyer	-0.8725	0.6315	-1.382	0.167
StringFL:s(Predictability):Fluencyinterruption	0.7795	0.6320	1.233	0.218
StringFR:s(Predictability):Fluencyinterruption	0.7247	0.6320	1.147	0.251
StringPR:s(Predictability):Fluencyinterruption	-0.0716	0.6320	-0.113	0.910

TABLE B.4: Fixed effects output from the linear mixed-effects model described in Section 9.7.3. Data were averaged over four regions of interest, referred to here as 'string'. FL refers to the front left quadrant, FR the front right, PL the posterior left, and PR the posterior right.

B.5 Experiment 3 - N400 effect data

Fixed effects:				
	Estimate	std error	t value	p
(Intercept)	-0.8470	0.5654	-1.498	0.134
s(Predictability)	-1.6168	0.8427	-1.919	0.055
FluencyFvD	0.1375	0.3242	0.424	0.671
FluencyIvD	-1.8774	0.5412	-3.469	0.001
LocationCP	-0.8406	0.5014	-1.676	0.094
LocationF	1.6524	0.5445	3.035	0.002
LocationFC	1.1032	0.5273	2.092	0.036
LocationP	-1.3793	0.5138	-2.684	0.007
LocationPO	-1.5481	0.5328	-2.906	0.004
Contextcough	-2.0397	0.6093	-3.348	0.001
Trial	0.0016	0.0030	0.519	0.604
s(Predictability):FluencyFvD	-0.1978	0.5602	-0.353	0.724
s(Predictability):FluencyIvD	0.0046	0.9609	0.005	0.996
s(Predictability):LocationCP	-0.1392	0.9979	-0.140	0.889
s(Predictability):LocationF	0.9053	0.9979	0.907	0.364
s(Predictability):LocationFC	0.3709	0.9979	0.372	0.710
s(Predictability):LocationP	0.0640	0.9979	0.064	0.949
s(Predictability):LocationPO	0.2934	0.9979	0.294	0.769
FluencyFvD:LocationCP	0.4834	0.3546	1.363	0.173
FluencyIvD:LocationCP	-0.1931	0.6078	-0.318	0.751
FluencyFvD:LocationF	-0.9886	0.3547	-2.787	0.005
FluencyIvD:LocationF	1.2957	0.6079	2.132	0.033
FluencyFvD:LocationFC	-0.5793	0.3547	-1.634	0.102
FluencyIvD:LocationFC	0.7717	0.6079	1.269	0.204
FluencyFvD:LocationP	0.8502	0.3546	2.397	0.017
FluencyIvD:LocationP	-0.0669	0.6078	-0.110	0.912
FluencyFvD:LocationPO	1.1279	0.3547	3.180	0.001
FluencyIvD:LocationPO	0.0883	0.6079	0.145	0.884
s(Predictability):Contextcough	-1.0376	1.1005	-0.943	0.346
FluencyFvD:Contextcough	1.1969	0.3973	3.012	0.003
FluencyIvD:Contextcough	0.6266	0.6615	0.947	0.344
LocationCP:Contextcough	0.2530	0.6885	0.367	0.713
LocationF:Contextcough	-0.1174	0.7537	-0.156	0.876
LocationFC:Contextcough	-0.3163	0.7278	-0.435	0.664
LocationP:Contextcough	0.4980	0.7047	0.707	0.480
LocationPO:Contextcough	0.5256	0.7277	0.722	0.470
s(Predictability):Trial	-0.0005	0.0041	-0.126	0.900
FluencyFvD:Trial	-0.0048	0.0016	-3.025	0.002
FluencyIvD:Trial	0.0062	0.0027	2.348	0.019
LocationCP:Trial	0.0015	0.0026	0.572	0.567
LocationF:Trial	-0.0024	0.0028	-0.857	0.391
LocationFC:Trial	-0.0020	0.0027	-0.726	0.468
LocationP:Trial	0.0032	0.0027	1.195	0.232
LocationPO:Trial	0.0040	0.0027	1.480	0.139
Contextcough:Trial	0.0030	0.0034	0.878	0.380
s(Predictability):FluencyFvD:LocationCP	-0.0512	0.7093	-0.072	0.943
s(Predictability):FluencyIvD:LocationCP	0.0949	1.2156	0.078	0.938
s(Predictability):FluencyFvD:LocationF	0.1611	0.7094	0.227	0.820
s(Predictability):FluencyIvD:LocationF	-0.2954	1.2157	-0.243	0.808
s(Predictability):FluencyFvD:LocationFC	0.1270	0.7094	0.179	0.858
s(Predictability):FluencyIvD:LocationFC	-0.1024	1.2157	-0.084	0.933
s(Predictability):FluencyFvD:LocationP	-0.2117	0.7094	-0.298	0.765
s(Predictability):FluencyIvD:LocationP	0.0699	1.2156	0.057	0.954
s(Predictability):FluencyFvD:LocationPO	0.0141	0.7094	0.020	0.984
s(Predictability):FluencyIvD:LocationPO	0.3390	1.2157	0.279	0.780
s(Predictability):FluencyFvD:Contextcough	-0.7440	0.7659	-0.971	0.331
s(Predictability):FluencyIvD:Contextcough	0.1936	1.3197	0.147	0.883
s(Predictability):LocationCP:Contextcough	0.4261	1.3701	0.311	0.756
s(Predictability):LocationF:Contextcough	-0.6382	1.3702	-0.466	0.641
s(Predictability):LocationFC:Contextcough	-0.3314	1.3702	-0.242	0.809
s(Predictability):LocationP:Contextcough	0.2518	1.3701	0.184	0.854
s(Predictability):LocationPO:Contextcough	0.2952	1.3701	0.215	0.829
FluencyFvD:LocationCP:Contextcough	-0.2967	0.4850	-0.612	0.541
FluencyIvD:LocationCP:Contextcough	0.0151	0.8379	0.018	0.986
FluencyFvD:LocationF:Contextcough	0.3469	0.4850	0.715	0.474
FluencyIvD:LocationF:Contextcough	-0.2118	0.8380	-0.253	0.801
FluencyFvD:LocationFC:Contextcough	0.3359	0.4850	0.693	0.489

FluencyIvD:LocationFC:Contextcough	-0.2793	0.8380	-0.333	0.739
FluencyFvD:LocationP:Contextcough	-0.6018	0.4850	-1.241	0.215
FluencyIvD:LocationP:Contextcough	-0.2494	0.8380	-0.298	0.766
FluencyFvD:LocationPO:Contextcough	-1.0273	0.4850	-2.118	0.034
FluencyIvD:LocationPO:Contextcough	-0.4430	0.8380	-0.529	0.597
s(Predictability):FluencyFvD:Trial	0.0028	0.0029	0.967	0.334
s(Predictability):FluencyIvD:Trial	-0.0017	0.0050	-0.345	0.730
s(Predictability):LocationCP:Trial	0.0004	0.0052	0.076	0.939
s(Predictability):LocationF:Trial	-0.0019	0.0052	-0.354	0.723
s(Predictability):LocationFC:Trial	-0.0008	0.0052	-0.160	0.873
s(Predictability):LocationP:Trial	0.0000	0.0052	0.009	0.993
s(Predictability):LocationPO:Trial	0.0001	0.0052	0.014	0.989
FluencyFvD:LocationCP:Trial	-0.0007	0.0019	-0.361	0.718
FluencyIvD:LocationCP:Trial	-0.0003	0.0032	-0.101	0.920
FluencyFvD:LocationF:Trial	0.0017	0.0019	0.909	0.363
FluencyIvD:LocationF:Trial	0.0003	0.0032	0.080	0.936
FluencyFvD:LocationFC:Trial	0.0009	0.0019	0.501	0.616
FluencyIvD:LocationFC:Trial	-0.0007	0.0032	-0.232	0.816
FluencyFvD:LocationP:Trial	-0.0012	0.0019	-0.665	0.506
FluencyIvD:LocationP:Trial	-0.0016	0.0032	-0.494	0.622
FluencyFvD:LocationPO:Trial	-0.0017	0.0019	-0.921	0.357
FluencyIvD:LocationPO:Trial	-0.0026	0.0032	-0.820	0.412
s(Predictability):Contextcough:Trial	0.0022	0.0062	0.361	0.718
FluencyFvD:Contextcough:Trial	0.0008	0.0022	0.369	0.712
FluencyIvD:Contextcough:Trial	-0.0014	0.0037	-0.386	0.700
LocationCP:Contextcough:Trial	-0.0005	0.0038	-0.126	0.899
LocationF:Contextcough:Trial	-0.0005	0.0042	-0.129	0.898
LocationFC:Contextcough:Trial	0.0008	0.0040	0.187	0.852
LocationP:Contextcough:Trial	-0.0010	0.0039	-0.255	0.799
LocationPO:Contextcough:Trial	-0.0003	0.0040	-0.066	0.947
s(Predictability):FluencyFvD:LocationCP:Contextcough	-0.2350	0.9700	-0.242	0.809
s(Predictability):FluencyIvD:LocationCP:Contextcough	-0.4563	1.6759	-0.272	0.785
s(Predictability):FluencyFvD:LocationF:Contextcough	0.6220	0.9701	0.641	0.521
s(Predictability):FluencyIvD:LocationF:Contextcough	0.5844	1.6761	0.349	0.727
s(Predictability):FluencyFvD:LocationFC:Contextcough	0.2685	0.9701	0.277	0.782
s(Predictability):FluencyIvD:LocationFC:Contextcough	0.4793	1.6760	0.286	0.775
s(Predictability):FluencyFvD:LocationP:Contextcough	-0.3250	0.9701	-0.335	0.738
s(Predictability):FluencyIvD:LocationP:Contextcough	-0.7332	1.6759	-0.437	0.662
s(Predictability):FluencyFvD:LocationPO:Contextcough	-0.6640	0.9701	-0.685	0.494
s(Predictability):FluencyIvD:LocationPO:Contextcough	-1.5352	1.6760	-0.916	0.360
s(Predictability):FluencyFvD:LocationCP:Trial	0.0007	0.0037	0.182	0.856
s(Predictability):FluencyIvD:LocationCP:Trial	-0.0002	0.0064	-0.029	0.977
s(Predictability):FluencyFvD:LocationF:Trial	-0.0026	0.0037	-0.698	0.485
s(Predictability):FluencyIvD:LocationF:Trial	0.0014	0.0064	0.211	0.833
s(Predictability):FluencyFvD:LocationFC:Trial	-0.0014	0.0037	-0.380	0.704
s(Predictability):FluencyIvD:LocationFC:Trial	0.0002	0.0064	0.033	0.974
s(Predictability):FluencyFvD:LocationP:Trial	0.0016	0.0037	0.422	0.673
s(Predictability):FluencyIvD:LocationP:Trial	0.0000	0.0064	0.002	0.998
s(Predictability):FluencyFvD:LocationPO:Trial	0.0006	0.0037	0.166	0.868
s(Predictability):FluencyIvD:LocationPO:Trial	-0.0012	0.0064	-0.186	0.853
s(Predictability):FluencyFvD:Contextcough:Trial	0.0033	0.0043	0.765	0.444
s(Predictability):FluencyIvD:Contextcough:Trial	0.0018	0.0074	0.248	0.805
s(Predictability):LocationCP:Contextcough:Trial	-0.0041	0.0075	-0.551	0.582
s(Predictability):LocationF:Contextcough:Trial	0.0059	0.0075	0.792	0.429
s(Predictability):LocationFC:Contextcough:Trial	0.0034	0.0075	0.457	0.647
s(Predictability):LocationP:Contextcough:Trial	-0.0042	0.0075	-0.566	0.571
s(Predictability):LocationPO:Contextcough:Trial	-0.0046	0.0075	-0.615	0.538
FluencyFvD:LocationCP:Contextcough:Trial	0.0000	0.0026	-0.013	0.989
FluencyIvD:LocationCP:Contextcough:Trial	-0.0011	0.0046	-0.246	0.806
FluencyFvD:LocationF:Contextcough:Trial	-0.0009	0.0026	-0.330	0.741
FluencyIvD:LocationF:Contextcough:Trial	0.0011	0.0046	0.239	0.811
FluencyFvD:LocationFC:Contextcough:Trial	-0.0009	0.0026	-0.359	0.720
FluencyIvD:LocationFC:Contextcough:Trial	0.0019	0.0046	0.425	0.671
FluencyFvD:LocationP:Contextcough:Trial	0.0005	0.0026	0.202	0.840
FluencyIvD:LocationP:Contextcough:Trial	-0.0006	0.0046	-0.134	0.893
FluencyFvD:LocationPO:Contextcough:Trial	0.0022	0.0026	0.845	0.398
FluencyIvD:LocationPO:Contextcough:Trial	-0.0006	0.0046	-0.130	0.896
s(Predictability):FluencyFvD:LocationCP:Contextcough:Trial	0.0002	0.0053	0.037	0.971
s(Predictability):FluencyIvD:LocationCP:Contextcough:Trial	0.0034	0.0091	0.367	0.713
s(Predictability):FluencyFvD:LocationF:Contextcough:Trial	-0.0019	0.0053	-0.363	0.717
s(Predictability):FluencyIvD:LocationF:Contextcough:Trial	-0.0029	0.0091	-0.313	0.754

s(Predictability):FluencyFvD:LocationFC:Contextcough:Trial	-0.0008	0.0053	-0.147	0.883
s(Predictability):FluencyIvD:LocationFC:Contextcough:Trial	-0.0028	0.0091	-0.302	0.763
s(Predictability):FluencyFvD:LocationP:Contextcough:Trial	-0.0002	0.0053	-0.029	0.977
s(Predictability):FluencyIvD:LocationP:Contextcough:Trial	0.0066	0.0091	0.723	0.470
s(Predictability):FluencyFvD:LocationPO:Contextcough:Trial	0.0007	0.0053	0.138	0.890
s(Predictability):FluencyIvD:LocationPO:Contextcough:Trial	0.0118	0.0091	1.291	0.197

TABLE B.5: Fixed effects output from the linear mixed-effects model described in Section 9.7.4.

References

Adrian, E. D., Matthews, B. H., & C. (1934). The Berger rhythm: potential changes from the occipital lobes in man. *Brain*, *57*, 355–385.

American Electroencephalographic Society. (1994). Guideline thirteen: Guidelines for standard electrode position nomenclature. *Journal of Clinical Neurophysiology*, *11*, 111–113.

Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003, January). Disfluencies signal theee, um, new information. *Journal of psycholinguistic research*, *32*(1), 25–36. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/12647561`

Arnold, J. E., Hudson-Kam, C. L., & Tanenhaus, M. K. (2007, September). If you say thee uh you are describing something hard: the on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *33*(5), 914–930. doi: 10.1037/0278-7393.33.5.914

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004, September). The old and thee, uh, new: disfluency and reference resolution. *Psychological science*, *15*(9), 578–82. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/15327627` doi: 10.1111/j.0956-7976.2004.00723.x

Arnold, J. E., Wasow, T., Losongco, A., & Ginstrom, R. (2000). Heaviness vs. newness: The effects of complexity and information structure on constituent ordering. *Language*, *76*, 28–55.

Baayen, R., Davidson, D., & Bates, D. (2008, November). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0749596X07001398` doi: 10.1016/j.jml.2007.12.005

Baayen, R., & H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bagiella, E., Sloan, R., & Heitjan, D. (2000, January). Mixed-effects models in psychophysiology. *Psychophysiology*, *37*(1), 13–20. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/10705763`

Bailey, K. G. D., & Ferreira, F. (2003, August). Disfluencies affect the parsing of garden-path sentences. *Journal of Memory and Language*, *49*(2), 183–200. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0749596X03000275` doi: 10.1016/S0749-596X(03)00027-5

Barr, D. (2001). Trouble in mind: Paralinguistic indices of effort and uncertainty in communication. In S. Santi, I. Guaïtella, C. Cave & G. Konopczynsk (Eds.), *Oralité et gestualité, communication multimodale, intéraction* (pp. 597–600). Paris: LHarmattan.

Barr, D., Levy, R., Scheepers, C., & Tily, H. J. (2013, April). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and*

Language, *68*(3), 255–278. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0749596X12001180 doi: 10.1016/j.jml.2012.11.001

Barr, D., & Seyfeddinipur, M. (2010, May). The role of fillers in listener attributions for speaker disfluency. *Language and Cognitive Processes*, *25*(4), 441–455. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/01690960903047122 doi: 10.1080/01690960903047122

Bates, D., Maechler, M., & Bolker, B. (2013). *lme4.0: Linear mixed-effects models using S4 classes. R package version 0.999999-4/r1876.* Retrieved from http://r-forge.r-project.org/projects/lme4/

Beattie, W., & Butterworth, B. L. (1979). Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech*, *22*(3), 201–211.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, *113*(2), 1001. Retrieved from http://scitation.aip.org/content/asa/journal/jasa/113/2/10.1121/1.1534836 doi: 10.1121/1.1534836

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials associated with semantic priming. *Electroencephalography and Clinical Neurophysiology*, *60*, 343–355.

Berger, H. (1929). Ueber das Elektrenkephalogramm des Menschen. *Archives fur Psychiatrie Nervenkrankheiten*, *87*, 527–570.

Besson, M., Faita, F., Czternasty, C., & Kutas, M. (1997). Whats in a pause: Event-related potential analysis of temporal disruptions in written and spoken sentences. *Biological Psychology*, *16*, 3–23.

Bicknell, K. (2014a). *More on Old and New lme4.* Retrieved from https://hlplab.wordpress.com/2014/06/24/more-on-old-and-new-lme4/

Bicknell, K. (2014b). *Old and New lme4.* Retrieved from https://hlplab.wordpress.com/2014/03/17/old-and-new-lme4/

Boomer, D. S. (1965). Hesitation and grammatical encoding. *Language and Speech*, *8*(3), 148–158.

Brennan, S. E., & Schober, M. F. (2001, February). How Listeners Compensate for Disfluencies in Spontaneous Speech. *Journal of Memory and Language*, *44*(2), 274–296. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0749596X00927530 doi: 10.1006/jmla.2000.2753

Brennan, S. E., & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive state of speakers. *Journal of Memory and Language*, *34*, 383–398.

Bridson, N. C., Fraser, C. S., Herron, J. E., & Wilding, E. L. (2006, October). Electrophysiological correlates of familiarity in recognition memory and exclusion tasks. *Brain research*, *1114*(1), 149–60. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16934780 doi: 10.1016/j.brainres.2006.07.095

Cajal, S. R. (1909). *Histologie du systeme nerveux de lhomme et des vertebres.* Paris: Maloine.

Carlson, N. A. (1992). *Foundations of Physiological Psycholgy.* Needham Heights, Massachusetts: Simon & Schuster.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, *25*(975-979).

Christenfeld, N. (1995). Does It Hurt to Say Um? *Journal of Nonverbal Behaviour*, *19*(3), 171–186.

Chun, M. M., & Turk-Browne, N. B. (2007, April). Interactions between attention and memory. *Current opinion in neurobiology*, *17*(2), 177–84. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/17379501` doi: 10.1016/j.conb.2007.03.005

Clark, H. H. (1973). The Language-as-Fixed-Effect Fallacy : A Critique of Language Statistics in Psychological Research. *Journal of Learning and Verbal Behaviour*, *12*, 335–359.

Clark, H. H. (1994). Managing problems in speaking. *Speech Communication*, *15*, 243–250.

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*, 73–111.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. New York: Academic Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Coleman, E. B. (1964). Generalizing to a language population. *Psychological Reports*, *14*, 219–226.

Collard, P. (2009). *Disfluency and listeners attention : An investigation of the immediate and lasting effects of hesitations in speech* (Unpublished doctoral dissertation). University of Edinburgh.

Collard, P., Corley, M., MacGregor, L. J., & Donaldson, D. I. (2008). Attention orienting effects of hesitations in speech: Evidence from ERPs. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *34*, 696–702.

Corley, M., & Hartsuiker, R. J. (2003). Hesitation in speech can. . . um. . . help a listener understand. In *Proceedings of the 25th meeting of the cognitive science society* (pp. 276–281).

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, *105*, 658–668.

Coulson, S., King, J., & Kutas, M. (1998, January). Expect the Unexpected: Event-related Brain Response to Morphosyntactic Violations. *Language and Cognitive Processes*, *13*(1), 21–58. Retrieved from `http://www.informaworld.com/openurl?genre=article&doi=10.1080/016909698386582&magic=crossref||D404A21C5BB053405B1A640AFFD44AE3` doi: 10.1080/016909698386582

Coulson, S., & Wu, Y. C. (2005, March). Right hemisphere activation of joke-related information: an event-related brain potential study. *Journal of cognitive neuroscience*, *17*(3), 494–506. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/15814008` doi: 10.1162/0898929053279568

Craik, F. I., Govoni, R., Naveh-Benjamin, M., & Anderson, N. D. (1996, June). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of experimental psychology. General*, *125*(2), 159–80. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/9747524`

Curran, T. (2000, September). Brain potentials of recollection and familiarity. *Memory & Cognition*, *28*(6), 923–38. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/11105518`

Curran, T., & Dien, J. (2003). Differentiating amodal familiarity from modality-specific memory processes : An ERP study. *Psychophysiology*, *40*(6), 979–988. Retrieved from `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1413574`

Curran, T., Schacter, D. L., Johnson, M. K., & Spinks, R. (2001, February). Brain Potentials Reflect Behavioral Differences in True and False Recognition. *Journal of Cognitive Neuroscience*, *13*(2), 201–216. Retrieved from `http://www`

.mitpressjournals.org/doi/abs/10.1162/089892901564261 doi: 10.1162/089892901564261

Curran, T., Tepe, K. L., & Piatt, C. (2006). Event-related potential explorations of dual processes in recognition memory. In H. Zimmer, A. Mecklinger, & U. Lindenberger (Eds.), *Binding in human memory: A neurocognitive approach* (pp. 467–492). Oxford: Oxford University Press.

Davidson, D. J. (2009, June). Functional Mixed-Effect Models for Electrophysiological Responses. *Neurophysiology*, *41*(1), 71–79. Retrieved from http://link.springer.com/10.1007/s11062-009-9079-y doi: 10.1007/s11062-009-9079-y

Deterding, D. (2001). Letter to the Editor The measurement of rhythm : a comparison of Singapore and British English. *Journal of Phonetics*, *29*, 217–230.

Duez, D. (1985). Perception of Silent Pauses in Continuous Speech. *Language and Speech*, *28*(4), 377–389.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods,*, *39*, 175–191.

Federmeier, K. D., & Kutas, M. (1999, November). A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. *Journal of Memory and Language*, *41*(4), 469–495. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0749596X99926608 doi: 10.1006/jmla.1999.2660

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007, May). Multiple effects of sentential constraint on word processing. *Brain research*, *1146*, 75–84. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2704150&tool=pmcentrez&rendertype=abstract doi: 10.1016/j.brainres.2006.06.101

Ferguson, C. J., & Heene, M. (2012). A Vast Graveyard of Undead Theories: Publication Bias and Psychological Science's Aversion to the Null. *Perspectives on Psychological Science*, *7*, 555–561. Retrieved from http://pps.sagepub.com/content/7/6/555.abstract doi: 10.1177/1745691612459059

Ferreira, F., & Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language*, *25*, 725–745.

Fox Tree, J. E. (1995). The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech. *Journal of Memory and Language*, *34*, 709–738.

Fox Tree, J. E. (2001). Listeners' uses of um and uh in speech comprehension. *Memory & Cognition*, *29*(2), 320–326.

Fox Tree, J. E. (2002). Interpreting Pauses and Ums at Turn Exchanges. *Discourse Processes*, *34*(1), 37–55.

Fox Tree, J. E., & Clark, H. H. (1997, February). Pronouncing "the" as "thee" to signal problems in speaking. *Cognition*, *62*(2), 151–67. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9141905

Fraundorf, S. H., & Watson, D. G. (2008). Dimensions of Variation in Disfluency Production in Discourse. In J. Ginzburg, P. Healey, & Y. Sato (Eds.), *Proceedings of londial 2008, the 12th workshop on the semantics and pragmatics of dialogue* (pp. 131–138). London: King's College London.

Fraundorf, S. H., & Watson, D. G. (2011, April). The disfluent discourse: Effects of filled pauses on recall. *Journal of Memory and Language*, *65*(2), 161–175. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0749596X11000234 doi: 10.1016/j.jml.2011.03.004

Gibbs, F. A., Davis, H., & Lennox, W. G. (1935). The electo-encephalogram in epilepsy and in conditions of impaired consciousness. *Archives of neurolgy and psychiatry*, *34*(6), 1133–1148.

Goffman, E. (1981). Radio talk. In E. Goffman (Ed.), *Forms of talk* (pp. 197–327). Philadelphia, PA: University of Pennsylvania Press.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, *56*, 208–216.

Hawkins, R. R. (1971). The syntactic location of hesitation pauses. *Language and Speech*, *14*, 277–288.

Heike, A. E. (1981). A content-processing view of hesitation phenomena. *Language and Speech*, *24*, 147–160.

Holcomb, P. J., & Neville, H. J. (1991). Natural speech processing : An analysis using event-related brain potentials. *Psychobiology*, *19*(4), 286–300.

Hosman, L. A., & Wright II, J. W. (1987). The effects of hedges and hesitations on impression formation in a simulated courtroom context. *Western Journal of Speech Communication*, *51*(2), 173–188.

Irwin, R. S., Rosen, M. J., & Braman, S. S. (1977). Cough: A Comprehensive Review. *Arch Intern Med*, *137*(9), 1186–1191.

Janssen, N., van der Meij, M., & Barber, H. A. (2013). 11th Syposium of Psycholinguistics. In *11th syposium of psycholinguistics* (p. 15). Tenerife, Spain.

Jasper, H. H., & Carmichael, L. (1958). The ten-twenty electrode system of the International Federation. *Electroencephalography and Clinical Neurophysiology*, *10*, 371–375.

Joyce, C. A., Paller, K. A., Schwartz, T. J., & Kutas, M. (1999, September). An electrophysiological analysis of modality-specific aspects of word repetition. *Psychophysiology*, *36*(5), 655–65. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/10442034`

Kaan, E., Harris, A., Gibson, E., & Holcomb, P. J. (2000, April). The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes*, *15*(2), 159–201. Retrieved from `http://www.informaworld.com/openurl?genre=article&doi=10.1080/016909600386084&magic=crossref|D404A21C5BB053405B1A640AFFD44AE3` doi: 10.1080/016909600386084

Kutas, M., & Dale, A. (1997). Electrical and magnetic readings of mental function. In M. D. Rugg (Ed.), *Cognitive neuroscience studies in cognition* (pp. 197–241). Cambridge, Massachusetts: MIT Press.

Kutas, M., & Federmeier, K. D. (2009, December). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual review of psychology*, *62*(August), 621–47. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/20809790` doi: 10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1980). Reading Senseless Sentences: Brain Potentials Reflect Semantic Incongruity. *Science*, *207*, 203–205.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectnacy and semantic association. *Nature*, *307*, 161–163.

Kutas, M., Neville, H. J., & Holcomb, P. J. (1987). A preliminary comparison of the N400 response to semantic anomalies during reading, listening and signing. *Electroencephalography and Clinical Neurophysiology Supplement 39*, 325–330.

Lawrence, M. A. (2013). *ez: Easy analysis and visualization of factorial experiments.* R package version 4.2-0. Retrieved from `http://cran.r-project.org/package=ez`

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, *14*, 41–104.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, Massachusetts: MIT Press.

Luck, S. J. (2005). *An introduction to the Event Related Potential Technique.* Cambridge, Massachusetts.

MacGregor, L. J. (2008). *Disfluencies affect language comprehension : evidence from event-related potentials and recognition memory* (Unpublished doctoral dissertation). University of Edinburgh.

MacGregor, L. J., Corley, M., & Donaldson, D. I. (2009, October). Not all disfluencies are are equal: The effects of disfluent repetitions on language comprehension. *Brain and language*, *111*(1), 36–45. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/19700188` doi: 10.1016/j.bandl.2009.07.003

MacGregor, L. J., Corley, M., & Donaldson, D. I. (2010, December). Listening to the sound of silence: disfluent silent pauses in speech have consequences for listeners. *Neuropsychologia*, *48*(14), 3982–3992. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0028393210004148http://www.ncbi.nlm.nih.gov/pubmed/20950633` doi: 10.1016/j.neuropsychologia.2010.09.024

Maclay, H., & Osgood, C. E. (1959). Hesitation Phenomena in Spontaneous English Speech. *Word - Journal of the International Linguistic Association*, *15*(1), 19–44.

Marslen-Wilson, W., & Tyler, L. K. (1980, March). The temporal structure of spoken language understanding. *Cognition*, *8*(1), 1–71. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/7363578`

Martin, J. G. (1967). Hesitations in the Speaker's Production and Listener's Reproduction of Utterances. *Journal of Verbal Learning and Verbal Behavior*, *909*, 903–909.

Martin, J. G., & Strange, W. (1968, November). The perception of hesitation in spontaneous speech. *Perception & Psychophysics*, *3*(6), 427–438. Retrieved from `http://www.springerlink.com/index/10.3758/BF03205750` doi: 10.3758/BF03205750

Maxfield, N. D., Lyon, J. M., & Silliman, E. R. (2009, November). Disfluencies along the garden path: brain electrophysiological evidence of disrupted sentence processing. *Brain and language*, *111*(2), 86–100. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/19765813` doi: 10.1016/j.bandl.2009.08.003

McCallum, W. C., Farmer, S. F., & Pocock, P. V. (1984, November). The effects of physical and semantic incongruites on auditory event-related potentials. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, *59*(6), 477–488. Retrieved from `http://www.sciencedirect.com/science/article/pii/0168559784900066` doi: 10.1016/0168-5597(84)90006-6

McCarthy, G., & Wood, C. C. (1985). Scalp Distributions of Event-Related Potentials: An Ambiguity Associated With Analysis of Variance Models. , *62*, 203–208.

Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, *11*(56-60).

Mulligan, N. W. (1998, January). The role of attention during encoding in implicit and explicit memory. *Journal of experimental psychology. Learning, memory, and cognition*, *24*(1), 27–47. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/9438952`

Näätänen, R. (2001, January). The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology*, *38*(1), 1–21. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/11321610`

Näätänen, R., Gaillard, A. W. K., & Mäntysalo, S. (1978). Early Selective-attention Effect on Evoked Potential Reinterpreted. *Acta Psychologica*, *42*, 313–329.

Näätänen, R., & Kreegipuu, K. (2012). The Mismatch Negativity (MMN). In S. J. Luck & E. S. Kappenman (Eds.), *The oxford handbook of event related potential components* (pp. 143–157). New York: Oxford University Press.

Näätänen, R., & Winkler, I. (1999, November). The concept of auditory stimulus representation in cognitive neuroscience. *Psychological bulletin*, *125*(6), 826–59. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/10589304`

Nessler, D., Mecklinger, A., & Penney, T. B. (2001, January). Event related brain potentials and illusory memories: the effects of differential encoding. *Cognitive brain research*, *10*(3), 283–301. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/11167052`

NHS. (2013). *Cough - NHS Choices.* Retrieved from `http://www.nhs.uk/conditions/cough/pages/introduction.aspx`

Osterhout, L., & Holcomb, P. J. (1992). Event-Related Brain Potentials Elicited by Syntactic Anomaly , Journal of Memory and Language, 31:6 (1992:Dec.) p.785. *Journal of Memory and Language*, *31*, 785–806.

Otten, L. J., Rugg, M. D., & Doyle, M. C. (1993). Modulation of event-related potentials by word repetition: the role of selective attention. *Psychophysiology*, *30*, 559571.

Oviatt, S. (1995). Precicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*, *9*(19-35).

Ozuru, Y., & Hirst, W. (2006). Surface features of utterances, credibility judgments, and memory. *Memory & cognition*, *34*(7), 1512–1526. doi: 10.3758/BF03195915

Paller, K. A., Voss, J. L., & Boehm, S. G. (2007, June). Validating neural correlates of familiarity. *Trends in cognitive sciences*, *11*(6), 243–50. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/17475539` doi: 10.1016/j.tics.2007.04.002

Pashler, H. E. (1998a). *Attention* (H. E. Pashler, Ed.). Psychology Press.

Pashler, H. E. (1998b). *The psychology of attention.* MIT Press.

Polich, J. (2004). Neuropsychology of P3a and P3b: A theoretical overview. In N. Moore & K. Arikan (Eds.), *Brainwaves and mind: Recent developments* (p. 1529). Wheaton, IL: Kjellberg Inc. Retrieved from `http://books.google.co.uk/books?hl=en&lr=&id=XYkQ-u5A54oC&oi=fnd&pg=PA15&dq=Polich+2004&ots=pmw1Rdo7Zp&sig=9TVpyD0s3OdzEd3erTP1kz4PEPU#v=onepage&q&f=false`

Polich, J., & Criado, J. R. (2006, May). Neuropsychology and neuropharmacology of P3a and P3b. *International journal of psychophysiology : official journal of the International Organization of Psychophysiology*, *60*(2), 172–85. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/16510201` doi: 10.1016/j.ijpsycho.2005.12.012

R Core Team. (2013). *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* Vienna, Austria. Retrieved from `www.R-project.org`

Rugg, M. D., Mark, R. E., Walla, P., Schloerscheidt, A. M., Birch, C. S., & Allan, K. (1998). Dissociationof the neural correlates of implicit and explicit memory. *Nature*, *392*(April), 595–598.

Rugg, M. D., & Nagy, M. E. (1989, May). Event-related potentials and recognition memory for words. *Electroencephalography and Clinical Neurophysiology*, *72*(5), 395–406. Retrieved from `http://www.sciencedirect.com/science/article/pii/001346948990045X` doi: 10.1016/0013-4694(89)90045-X

Schachter, S., Christenfeld, N., Ravina, B., & Bilous, F. (1991). Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology*, *60*,

362–267.

Schachter, S., Rauscher, F., Christenfeld, N., & Tyson Crone, K. (1994). The vocabularies of academia. *Psychological science*, *5*, 37–41.

Scharf, B., & Buus, S. (1986). Handbook of perception and human performance, volume i. In K. Boff, L. Kaufman, & J. Thomas (Eds.), (pp. 14.1–14.71). New York: Wiley.

Schnadt, M. J., & Corley, M. (2006). The Influence of Lexical , Conceptual and Planning Based Factors on Disfluency Production. In *Proceedings of the twenty-eighth meeting of the cognitive science society.* Vancouver, Canada.

Schroger, E. (1997). On the detection of auditory deviations: A pre-attentive activation model. *Psychophysiology*, *34*, 245–257.

Smith, N. J., & Kutas, M. (n.d.-a, October). Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. *Psychophysiology*. Retrieved from `http://doi.wiley.com/10.1111/psyp.12320` doi: 10.1111/psyp.12320

Smith, N. J., & Kutas, M. (n.d.-b). Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology*.

Smith, V., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language*, *32*, 25–38.

Squires, N. K., Squires, K. C., & Hillyard, S. A. (1975, April). Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalography and clinical neurophysiology*, *38*(4), 387–401. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/46819`

Swinney, D. A. (1979). Lexical Access during Sentence Comprehension: (Re)Consideration of Context Effects. *Journal of Verbal Learning and Verbal Behavior*, *18*, 645–659.

Tanenhaus, M. K., & Trueswell, J. C. (1995). Sentence comprehension. In J. Miller & P. Eimas (Eds.), *The handbook of perception and cognition: Vol 11.* San Diego, CA: Academic Press.

Urbach, T. P., & Kutas, M. (2002, November). The intractability of scaling scalp distributions to infer neuroelectric sources. *Psychophysiology*, *39*(6), 791–808. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/12462507` doi: 10.1017/S0048577202010648

Van Berkum, J., Hagoort, P., & Brown, C. M. (1999). Semantic Integration in Sentences and Discourse : Evidence from the N4 00. *Journal of Cognitive Neuroscience*, *11*(6), 657–671.

Van Petten, C. (1993, November). A comparison of lexical and sentence-level context effects in event-related potentials. *Language and Cognitive Processes*, *8*(4), 485–531. Retrieved from `http://www.tandfonline.com/doi/abs/10.1080/01690969308407586` doi: 10.1080/01690969308407586

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *25*(2), 394–417.

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory & Cognition*, *18*(4), 380–393.

Van Petten, C., Kutas, M., Kluender, R., Mitchiner, M., & McIsaac, H. (1991). Fractionating the word repetition effect with event-related potentials. *Journal of Cognitive Neuroscience*, *3*(2), 131–150.

Voss, J. L., & Paller, K. A. (2006, January). Fluent conceptual processing and explicit memory for faces are electrophysiologically distinct. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *26*(3), 926–33. Retrieved from

http://www.ncbi.nlm.nih.gov/pubmed/16421312 doi: 10.1523/JNEUROSCI.3931-05.2006

Watanabe, M., Hirose, K., Den, Y., & Minematsu, N. (2008, February). Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech Communication*, *50*(2), 81–94. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0167639307001033 doi: 10.1016/j.specom.2007.06.002

Wilding, E. L. (2000, February). In what way does the parietal ERP old/new effect index recollection? *International journal of psychophysiology : official journal of the International Organization of Psychophysiology*, *35*(1), 81–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10683669

Wilding, E. L. (2006, June). The practice of rescaling scalp-recorded event-related potentials. *Biological psychology*, *72*(3), 325–32. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16442202 doi: 10.1016/j.biopsycho.2005.12.002

Wilding, E. L., Doyle, M. C., & Rugg, M. D. (1995, June). Recognition memory with and without retrieval of context: an event-related potential study. *Neuropsychologia*, *33*(6), 743–67. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7675165

Wilding, E. L., & Rugg, M. D. (1996, June). An event-related potential study of recognition memory with and without retrieval of source. *Brain : a journal of neurology*, *119 (Pt 3*, 889–905. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8673500

Wilding, E. L., & Rugg, M. D. (1997, September). An event-related potential study of memory for words spoken aloud or heard. *Neuropsychologia*, *35*(9), 1185–95. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9364489

Wood, S. N. (2006). *Generalized Additive Models: an introduction with R*. Fort Lauderdale: Chapman & Hall/CRC.