In M. Carreiras & C. Clifton, Jr. (2004) (Eds.), The on-line study of sentence comprehension: Eyetracking, ERPs and beyond (pp. 229-270). New York: Psychology Press. This book has several other overview chapters on on-line methodology, including contributions from Don Mitchell, Julie Boland, Martin Pickering, Lee Osterhout, and Mike Tanenhaus. Check it out!

Chapter 13
Sentence Comprehension in a Wider Discourse: Can We Use ERPs To Keep Track of Things?

JOS J. A. VAN BERKUM

13.1 INTRODUCTION

13.1.1 ERPs and Sentence Comprehension

It has been known for a long time that event-related brain potentials (ERPs) can provide valuable information about the nature and time course of sentence comprehension. Brain potential research on sentence comprehension took off in the late 1970s, when Marta Kutas and Steve Hillyard discovered that semantically anomalous words at the end of a sentence, as in *He spread the warm bread with socks,* elicited a conspicuous negative deflection in the ERP at about 400 ms after the offending word (Kutas & Hillyard, 1980). Because this so-called N400 effect was not elicited by a typographic anomaly, as in *He spread the warm bread with BUTTER,* Kutas and Hillyard took the effect to reflect something about the semantic processing of words in relation to the sentence-semantic context. Follow-up experiments soon confirmed this observation. It also became clear that N400 effects reflected graded modulations of an underlying N400 *component*, elicited by every content word, but with its amplitude increasing to the extent that the word was somehow less expected given the sentence-semantic context (see Kutas & Van Petten, 1994, for a review).

The significance of these N400 findings was enhanced by the end of the 80s, when syntactically anomalous or unexpected words were found to elicit a qualitatively very different ERP effect, the so-called P600/SPS effect (Osterhout & Holcomb, 1992; Hagoort, Brown, & Groothusen, 1993). The discovery of these two very different ERP "signatures" raised interesting theoretical questions about the architecture of the sentence comprehension system and the types of representations it computed. However, it also implied that ERPs could be used as a tool to *selectively* keep track of specific aspects of sentence comprehension as a sentence unfolded in real time.

Of course, a process as complex as sentence comprehension was bound to generate more than just two ERP effects. Several other language-relevant ERP

phenomena were soon discovered, including another short-lived effect associated with syntactic analysis (the so-called Left-Anterior Negativity or LAN; e.g., Neville, Nicol, Barss, Forster, & Garrett, 1991; Friederici, Hahne, & Mecklinger, 1996), slow ERP shifts associated with verbal working memory (e.g., Kluender & Kutas, 1993; Mueller, King, & Kutas, 1997), and effects that seem to reflect the extent to which the phonology of a word matches some sentence-based expectation (e.g., Connoly & Phillips, 1994; Van den Brink, Brown, & Hagoort, 2001). With this small but growing repertoire of relatively selective ERP effects in hand, EEG researchers have begun to explore the architecture of the sentence comprehension system (see Osterhout, McLaughlin, Kim, Greenwald, & Inoue, chapter 14, this volume, for review; see also Brown & Hagoort, 2000; Brown, Hagoort, & Kutas, 2000; Friederici, 1998, 2002; Hagoort, Brown, & Osterhout, 1999; Kutas & Federmeier, 2000; Kutas, Federmeier, Coulson, King, & Münte, 2000; Kutas & Schmitt, 2003; Osterhout, McLaughlin, & Bersick, 1997).

In the typical sentential ERP experiment, subjects read or listen to a sequence of totally unrelated sentences, such as *The red-white rocket safely landed on the moon./He spread the warm bread with socks./Max shot at Onno as he jumped over the fence./The broker persuaded to sell the stock was sent to jail./* etc. For many of the issues addressed in this research, this is by all means good enough—after all, people can parse and make sense of a sentence presented in isolation, and there is no *a priori* reason to assume that the basic processes involved here are radically different from those involved in processing sentences in context (but see Clark, 1997, for a different view). On the other hand, of course, this assumption needs to be checked at some point in time. Moreover, if we want to understand how language users integrate their comprehension of an unfolding sentence with their knowledge of the wider discourse, such as a conversation or a piece of written text, we need to go beyond the isolated sentence. The purpose of this chapter is to see whether we can take ERPs up to the level of discourse-dependent comprehension, and what we might gain by doing so.

13.1.2 Taking ERPs Beyond the Single Sentence

Discourse-level comprehension has been studied with a wide range of behavioral measures, including eyetracking, self-paced reading, and concurrent probe response tasks (see Graesser, Millis, & Zwaan, 1997; Kintsch, 1998; and Myers & O'Brien, 1998, for reviews). To date, however, there is very little research in which discourse-level processing has been examined by means of ERPs—or any other neuroimaging measure, for that matter. Of the 130 citations in the Graesser et al. review of discourse comprehension research, for instance, none is to a neuroimaging paper. It is not that Graesser et al. missed an entire body of research. A PsychInfo search on "discourse" ("stories," "text") and "ERP" (or other potentially relevant neuroimaging terms) yields only a handful of studies, most of which appeared after 1997.

One reason why ERP and other neuroimaging researchers may have stayed away from discourse-level comprehension is that working with neuroimaging methods is difficult enough as it is. On the practical side of things, for example, the use of EEG imposes rather severe constraints on one's experiment, one of which is the need to

have a much larger number of critical trials per condition (at least ~30–40) than is common in most behavioral designs. With a story in every trial, this can get rather awkward, an issue to which I will return below. However, there is probably more to the story than just practical hurdles. Discourse-level processing is often associated with "everything affecting everything else," a form of intractability that might unfavorably interact with the perceived intractability of ERP waveform interpretation. Also note that compared to, say, parsing, discourse-level comprehension also comes awfully close to the comprehender's "central system," the neurocognition of which was declared to be doomed by Jerry Fodor some 20 years ago in a monograph that was extremely influential amongst psycholinguists (Fodor, 1983).

The few ERP studies that have used discourse-level materials come in several varieties. One of the earliest ERP experiments on the N400 in sentence comprehension (Kutas & Hillyard, 1983) actually used prose passages taken from children's books, and the semantically incongruent words were deliberately made incongruous with respect to the particular sentence in which it occurred as well as with the gist of the entire story. This approach made sense given the research goals at hand, but it also made it impossible to disentangle the respective contribution of the local sentence and the wider discourse. Several other studies have used multisentence text stimuli simply because prose passages happened to provide a natural and convenient way to present large amounts of content and function words (Osterhout, Bersick, & McKinnon, 1997; Brown, Hagoort, & Ter Keurs, 1999).

More relevant here are the handful of ERP studies that were designed to examine comprehension as a function of the wider discourse context. To my knowledge, St. George, Mannes, and Hoffman (1994) were the first to do so, in an elegant ERP experiment. St. George et al. recorded ERPs as subjects were reading passages taken from earlier behavioral research (Bransford & Johnson, 1972; Dooling & Lachman, 1971) designed to make only limited sense when read without their title but perfect sense when the title was given with the passage (e.g., *The procedure for washing clothes*). The words in these passages generated smaller N400s when the title had been given than when the title had not been presented, allowing St. George et al. to infer that the N400 is not only sensitive to local sentence-semantic context, but also to global, discourse-level context.

Several studies have since then supported and extended this N400 claim in a variety of ways (Federmeier & Kutas, 1999a, 1999b; Federmeier, McLennan, De Ochoa, & Kutas, 2002; Van Berkum, Hagoort, & Brown, 1999c; Van Berkum, Zwitserlood, Brown, & Hagoort, 2003b; see also Van Petten, Kutas, Kluender, Mitchiner, & McIsaac, 1991 for indirect evidence). Other recent studies have examined how and when comprehenders establish reference to entities mentioned in the earlier discourse (Van Berkum, Brown, & Hagoort, 1999a; Van Berkum, Brown, Hagoort, & Zwitserlood, 2003a; Streb, Rösler, & Hennighausen, 1999); the extent to which discourse context affects syntactic ambiguity resolution (Van Berkum et al., 1999a, 1999b), and the degree to which comprehenders can use discourse-level information to make inferences (St. George, Mannes, & Hoffman, 1997), or to predict specific upcoming words (Van Berkum, Brown, Hagoort, Zwitserlood, & Kooijman; 2002a; Van Berkum, Kooijman, Brown, Zwitserlood, & Hagoort, 2002b). Finally, in an interesting twist, recent research has examined the

ERP correlates of discourse-level semantic integration by means of stories relayed not in the form of text, but in the form of entirely non-textual cartoons (West & Holcomb, 2002) or videos (Sitnikova, Kuperberg, & Holcomb, 2003).

13.1.3 Goals and Plan of This Paper

The main goal of this paper is to assess whether, in the light of the limited experience gained so far, it makes sense to use ERPs as a tool to track the processes involved in relating a sentence to the wider discourse on-line, as these processes occur. Of course, whether the use of a tool makes sense depends on what people want to do with it. For example, using ERPs to locate a specific discourse-relevant neural generator with 5 mm precision in the brain would not be a good idea. After briefly reviewing the purpose and nature of on-line measurement, I will therefore examine the utility of ERPs for *each of several different types of inferences* one might wish to draw from ERPs, illustrating each with recent discourse-level ERP research. After this relatively detailed analysis, I evaluate the specific pros and cons of using ERPs to track discourse-level comprehension relative to the other measures avaiable in this domain. In a final section, I briefly look at some new EEG-related developments, as well as at what other neuroimaging methods might bring.

13.2 ON-LINE MEASUREMENT—WHAT IS IT AND WHO CARES?

One can learn a great deal about how language comprehension actually works by looking at how the final interpretation changes as a function of the input. For example, a comparison of the ease with which language users arrive at a correct final interpretation of sentences with center-embedded versus right-branching phrase structures tells us that the comprehension process operates in a way that makes the latter easier to process than the former. Unfortunately, such off-line findings have as yet not provided sufficient constraints to pin down the exact nature of the language comprehension process. And, given the complexity of this process, it is unlikely that they ever will. It is generally accepted, therefore, that research on language comprehension also requires so-called on-line measures, which allow the researcher to track the comprehension process *as it unfolds*.

Processes by definition unfold in time but sentential input itself does, too. This confounding can make it hard to see what on-line measurement actually amounts to in the sentence comprehension domain. In the hypothetical example depicted in Figure 13.1, I try to disentangle the two confounding factors. Suppose we are interested in the comprehension processes that deliver a final interpretation for the spoken sentence *SnowWhite kissed a dwarf*. Furthermore, assume for sake of the argument that this sentential input unfolds *instantaneously*, taking 0 ms instead of the usual 1 or 2 s. Keeping the rest of the universe as it is, the sentence comprehension process that takes this "sentence impulse" as its input and delivers an interpretation as its result will, by definition, extend over time. To understand this unfolding comprehension process or "*impulse response*,"[1] it should be examined at various moments in time (Figure 13.1a), particularly if we suspect the process to have a complex internal

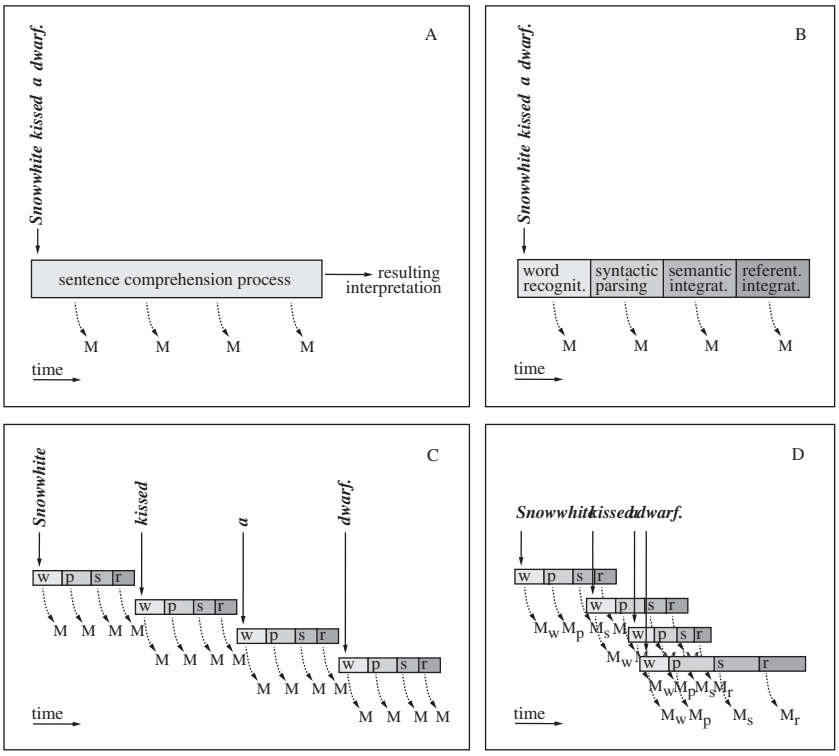

Figure 13.1 On-line measurement in language comprehension. Panel A: Impulse response tracking with instantaneous sentential input and a homogeneous response of the system (M = on-line measurement). Panel B: Impulse response tracking with instantaneous sentential input and an interestingly structured response (hypothetically assumed to consist of word recognition, parsing, semantic integration, and referential integration). Panel C: Incremental impulse response tracking in some ideal world, with word-driven impulse responses that are strictly sequential, nonoverlapping, of the same duration and internal structure, and unambiguously tied to their respective instantaneous word inputs. Panel D: Incremental impulse response tracking in the real world with word-driven impulse responses that are overlapping, of potentially different duration and internal structure, and (for spoken-language input) tied to word input that itself also unfolds and overlaps. Things would become more tractable if the on-line measurements themselves tell us what type of subprocess they are tapping in to (indicated by subscripted Ms).

structure (Figure 13.1b).[2] Note that the assumption of instantaneous sentential input reveals an aspect of on-line measurement that is solely related to *processes* unfolding in time and not to a *sentence* unfolding in time. I refer to this aspect of on-line measurement as *impulse response tracking*.

Of course, we know sentential input is not instantaneous, in that its words come in one after the other for spoken and written sentences alike (ignoring, for the moment, that not every word is fixated in reading, and that more than one word may be processed in a particular fixation; Rayner, 1998). We also know that the sentence

comprehension system is highly incremental, in that it processes every in-coming word to at least a considerable extent. Thus, we can remodel the problem of on-line measurement in sentence comprehension as one of *incremental impulse response tracking*, with every word eliciting its own little incremental impulse response, including, say, some word recognition, some parsing, some semantic, and some referential interpretation (Figure 13.1c).

Unfortunately, if the language comprehension system's impulse response to a word takes longer than the unfolding word itself (an empirical issue), the impulse responses to consecutive words of a sentence are going to overlap. To make things worse, the incoming words are usually not driving the system at fixed onset asynchronies, neither in natural reading (where fixation durations differ substantially) nor in listening (where acoustic word durations differ substantially). Moreover, in a spoken sentence, every word itself also unfolds in time, adding deep uncertainty as to how much of the unfolding lexical signal is enough to trigger (specific aspects of) a lexically driven impulse response. In all, measuring on-line in sentence comprehension can be characterized as tracking a sequence of potentially overlapping incremental impulse responses (Figure 13.1d). In plain English: a real mess.[3]

In an attempt to make the situation more tractable, researchers who study sentence comprehension on-line usually focus their attention on specific critical words, for example, one at which some syntactic garden path becomes apparent. However, relative to the situation depicted in Figure 13.1d, things would also become more tractable if the particular on-line measure that is used to track the system's incremental impulse response would tell us whether we are looking at, for instance, some aspect of parsing or some aspect of interpretation, as illustrated schematically by different subscripts in Figure 13.1d. One of the central claims developed in this chapter (see also Osterhout et al., chapter 14, this volume) is that this is precisely what ERPs are good at.

13.3 TYPES OF INFERENCES ONE CAN DRAW FROM ERPS

When comparing ERPs to other neuroimaging methods, the former is often said to have excellent temporal resolution but rather poor spatial resolution. As will be seen below, there is some truth to this. However, the adagium might also incorrectly be taken to suggest that ERPs can *only* be used to make timing inferences. A more careful analysis reveals that ERPs can actually be used to support some very different types of inferences about the language comprehension system—or, for that matter, any other cognitive subsystem (see also Rugg & Coles, 1995). Researchers usually look at the ERP response to some particular manipulation X to draw one or more of the following inferences:

1. *Sensitivity inferences.* Does anything at all happen in response to X, at whatever level of the comprehension system?
2. *Timing inferences.* At what moment in time is the comprehension system sensitive to X, that is, when does it "know" about X?

3. *Identity inferences.* Is whatever happens in the system in response to X, indexed by ERP effect E_X, the same as or qualitatively different from whatever happens in response to Y?
4. *Magnitude inferences.* How exactly does the manipulation X determine the size of the ERP effect E_X?
5. *Neuronal generator inferences.* What specific areas of the brain are involved in the processing of X, that is, where are the neural generators of ERP effect E_X?

To see how ERPs might contribute to the study of discourse-level comprehension, I will illustrate each of these types of inferences with specific ERP experiments on sentence comprehension in discourse. Most examples will be drawn from my own work, conducted in collaboration with Peter Hagoort, Colin Brown and, for the spoken language research, Pienie Zwitserlood. However, other relevant work will be drawn in whenever appropriate. Bear in mind that the purpose of this section is to examine ERPs as a tool and to have a closer look at the methodological issues that come up. For each of the example experiments, therefore, the specific theoretical debate that motivated it will be touched upon only briefly; interested readers are referred to the original literature.

13.3.1 Sensitivity Inferences

By far the simplest thing one can show with ERPs is that the comprehension system is sensitive to some manipulation X. We recently conducted a discourse-level ERP experiment that primarily used ERPs for this purpose (Van Berkum, Brown, et al., 2002a, 2004; Van Berkum, Kooijman, et al., 2002b). The goal of the study was to find out whether people can use their knowledge of the wider discourse to rapidly *predict* specific upcoming words while a sentence is unfolding. For example, do listeners anticipate the word *painting* by the time they have heard the mini-story in (1) up to the indefinite article?

(1) *The burglar had no trouble whatsoever locating the secret family safe. Of course, it was situated behind a*

Note that various phenomena suggest that they might indeed be able to do this. One is that in natural conversation, people can "take over" and finish each other's sentence quite successfully. Another is that when subjects are asked to complete a truncated story like the above in a so-called story completion or *cloze* test, they tend to come up with the same word (in this case, *painting*). Both observations suggest that, in at least some circumstances, people can indeed use their knowledge of the wider discourse to predict specific upcoming words. However, one might object that people may only be able to do this when given ample time, e.g., because the other speaker hesitates, or, in the paper-and-pencil cloze test, because subjects can essentially take all the time they want. The issue is whether people can use their knowledge of the wider discourse *rapidly enough* to predict specific upcoming words "on the fly," as the current sentence is unfolding.

To examine this issue, we created 120 Dutch two-sentence mini-stories like the above example. Each story was relatively predictable in that, in a written cloze pretest, at least 50% of the subjects had used the same noun to complete the story. The predictability of this noun always hinged on the discourse context sentence, as revealed by the fact the same noun was practically never generated if subjects had only seen the incomplete second sentence (*Of course, it was situated behind a …*).

As in German and French, every Dutch noun has a fixed and essentially arbitrary syntactic *gender* feature, which in indefinite (but not definite) NPs controls an inflectional suffix on the adjective:

(2) een groot schilderij	*a big$_{NEU}$ painting$_{NEU}$*	*neuter gender*	*"zero"suffix*
*een grot**e** boekenkast*	*a big$_{COM}$ bookcase$_{COM}$*	*common gender*	***-e** suffix*

In the ERP experiment, we used this fact to probe for discourse-based prediction of a noun by *first* continuing the story with an adjective whose inflectional suffix was either congruent or incongruent with the gender of the predictable noun. To make sure we would not be confounding the critical inflection *mismatch* effect with a mere inflection effect (e.g., "e" vs. zero inflection), we counterbalanced the latter across critical condition. Subjects were merely asked to listen to the stories (amid many fillers) as we recorded their EEG. The research logic was simple: If listeners indeed expect a specific noun by the time they have heard the indefinite article, an incongruently gender-inflected adjective should be an unpleasant surprise. The processing consequences of this perturbation might show up as an ERP effect at the adjective.

As can be seen in Figure 13.2, gender-incongruent adjectives indeed elicited a small but reliable ERP effect right at the inflection. Also, the effect disappears when the same critical sentences are heard in isolation, that is, without the discourse context that supports the lexical prediction (see Van Berkum, Brown, Hagoort, Zwitserlood, & Kooijman 2004). Because the ERP effect hinges on the (arbitrary) syntactic gender of an expected but not yet presented noun, it suggests that discourse-level information can indeed lead people to anticipate specific upcoming words "on the fly," as a local sentence unfolds. In addition, the effect suggests that the syntactic gender properties of a strongly anticipated noun can immediately begin to interact with local syntactic constraints (such as the gender inflection of a prenominal adjective).

This effect raises many questions (see Van Berkum et al., 2004 for discussion). For one, why is it so early? Also, is it really a message-level effect rather than some lexical priming effect? And, as it hinges on a syntactic gender agreement violation, why does not the effect look like, say, a P600/SPS effect? We are now pursuing these issues in follow-up research.[4] For current purposes, however, note that the research logic only required *a differential* ERP effect at the adjective, and neither required it to have a specific polarity or identity (P600/SPS, LAN, N400, etc.), nor a particular timing (provided that the features of the ERP effect actually observed are "reasonable"). Because all we needed was evidence for some perturbation of the system at the critical adjective, any other on-line measure might have done the job as well.[5]

The burglar had no trouble whatsoever locating the secret family safe. Of course, it was situated behind a... (preferred completion: ***painting-NEU***)

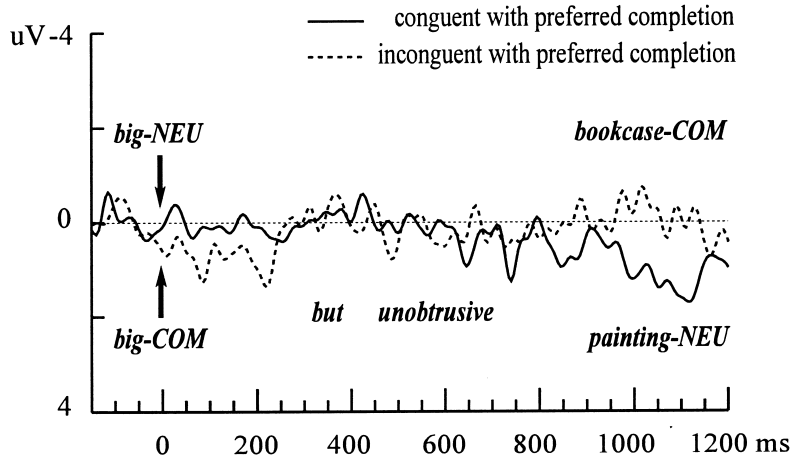

Figure 13.2 ERPs elicited by spoken adjectives whose suffixes are congruent or incongruent with the gender of a predictable noun. Results are shown for a right-temporal site (RT), and for stories with at least 75% cloze probability. Apart from the critical inflection-elicited early positivity, one can also see a discourse-induced N400 effect, elicited if a subsequently presented noun was not the predictable one. Estimated onset of the inflectional suffix is at 0 ms. Negative voltage is up. Data from J. J. A. Van Berkum et al., 2004.

At this point, it is instructive to examine some methodological issues associated with the use of ERPs to assess system sensitivity. First, what if we had not picked up an effect at the critical adjective inflection? As with any other measure, such a null result might indicate that the manipulation did not perturb the comprehension system, either because it *really* did not matter (e.g., people do not predict upcoming nouns), or because it was not strong enough. However, because of several reasons related to the ERP technique and the underlying physiology, there is also a real possibility that the language comprehension system *did* get perturbed by the critical adjective but the ERPs simply did not pick this up. As described in more detail by, for instance, Kutas, Federmeier, and Sereno (1999), the EEG recorded at the scalp reflects tiny changes in postsynaptic potentials that, in order for the associated tiny electrical fields to summate, must occur simultaneously within a very large number of neurons all oriented in the same way. If there are cognitive events—and there might be—for which any of these conditions does not hold, these events will not generate a measurable ERP effect at the scalp.

Furthermore, even cognitive events that do generate a "blip" in the ongoing raw EEG each time they occur may fail to show up in the ERP. The stimulus-tied blips that ERP researchers are after are very small, and therefore usually completely hidden in the much larger fluctuations of the background EEG. They can only be uncovered by computing an ERP, that is, by averaging the raw EEG measured at,

say, 50 critical stimulus events assumed to be equivalent in some way (e.g., all are unexpected inflections). Everything that is *not systematically time-locked* to the critical event (e.g., the random background EEG) is going to cancel out in the averaging, whereas the blips that *are* systematically time-locked to the critical event across trials will survive. And here is the catch: If each of our 50 critical trials does indeed generate a blip, but does so at highly variable latencies ("latency jitter") relative to what we take to be the critical event for our EEG segment alignment, the blips are not going to be fully superimposed in the averaging procedure. The consequence is that the resulting ERP effect may be "smeared," possibly even to such an extent that there is nothing left to see.[6]

A final word of caution: Perhaps, in part, because working with EEG is such a bother, those who make the effort can easily be tempted to believe (and suggest) that the EEG is per definition a more sensitive on-line measurement instrument than, say, self-paced reading, eyetracking, or cross-modal probe response time. In fact, and because of all of the reasons listed above, this is an entirely empirical issue, to be assessed anew for every new domain of inquiry.

13.3.2 Timing Inferences

For some issues, it is important to move beyond demonstrating any ERP effect at any (reasonable) time, and to more closely examine *when* the comprehension system is sensitive to some manipulation X. As an example, we recently used ERPs to establish how quickly listeners relate the meaning of incoming words to their knowledge of the wider discourse (Van Berkum et al., 2003b). Following up on a written-language study (Van Berkum et al., 1999c), we presented listeners with the Dutch equivalent of mini-stories such as in (3).

(3) *As agreed upon, Jane was to wake her sister and her brother at five o'clock in the morning. But the sister had already washed herself, and the brother had even got dressed. Jane told the brother that he was exceptionally* ***quick/slow****.*

These stories had been designed such that the two alternative critical words (in boldface in the example) were equally coherent with respect to the local sentence context. However, the wider discourse context rendered one of these words semantically incoherent (*slow* in the example), while leaving the other word (*quick*) perfectly acceptable.

When presented in discourse in a spoken-language ERP experiment, discourse-incoherent words elicited a differential ERP effect relative to their discourse-coherent counterparts (Figure 13.3a), which at some electrode sites emerged as early as 150 ms relative to the acoustic onset of the critical word. Also, when the critical sentences were presented in their carrier sentence but without the wider discourse, the differential effect disappeared (Figure 13.3b), showing that it, indeed, hinged on how the critical words related to the discourse. As revealed by a re-analysis of the ERP data for just those stories in which, according to a paper-and-pencil pretest, the *coherent* word had not been expected either (Figure 13.3c), the present discourse-dependent

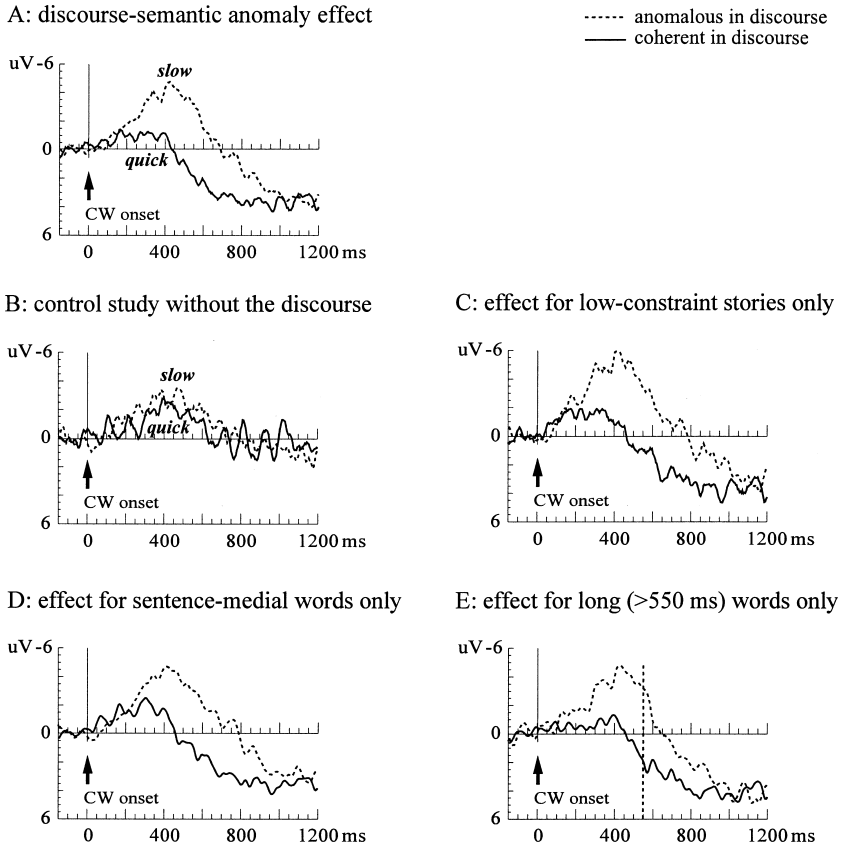

Figure 13.3 Panel A: ERPs to spoken words that were semantically coherent or anomalous with respect to the wider discourse. Panel B: ERPs to the same spoken words, now presented in their local carrier sentence only. Panel C: Discourse-semantic anomaly effect for low-constraint stories in which, next to the anomalous word, the *coherent* word had not been expected either (mean cloze probability of .01, and all below .05). Panel D: Discourse-semantic anomaly effect for sentence- (and clause-) medial words only. Panel E: Discourse-dependent anomaly effect for long words only, with the minimum word duration indicated by a vertical bar. Estimated acoustic onset of the critical words is at 0 ms. Data at Pz, from "When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect," by J. J. A. Van Berkum, P. Zwitserlood, C. M. Brown, and P. Hagoort, 2003b, *Cognitive Brain Research*, 17, 701–718.

negativity did *not* depend on the disconfirmation of a strong lexical prediction. Finally, subanalyses revealed that discourse-incoherent words elicited this early ERP effect regardless of whether they were at the end of a main or subordinate clause (not shown) or right in the middle of it (Figure 13.3d).

Note that the ERP data in Figure 13.3 afford two different kinds of inferences about how quickly listeners relate the meaning of incoming words to their knowledge of the wider discourse. One is *how soon in the sentence*, i.e., at which word the

effects show up. The fact that a discourse-dependent ERP effect shows up right in the middle of a clause confirms what intuition, earlier written-language ERP work (e.g., St. George et al., 1994; Van Berkum et al., 1999c), and earlier behavioral work (e.g., Marslen-Wilson & Tyler, 1980) already suggested, namely, that the system *incrementally* relates the incoming words to a representation of the wider discourse, as the sentence is unfolding.

In addition, ERPs are particularly good at also simultaneously supporting a second timing inference, which is *how soon after onset of the critical word* the processing consequences of a manipulation show up. The ERP elicited by discourse-incoherent words begins to diverge from that elicited by discourse-coherent words at about 150 ms after acoustic onset of the critical word. The average critical word in our study, however, took about 550 ms to unfold. This suggests that the language comprehension system seems to have detected that a particular word is not going to fit the discourse *long before the offending word itself has been fully heard*.

This is where several cautions should be made. First, remember that ERPs are averages, and not necessarily representative of each of the single-trial blips going into the average. In particular, these little blips need not be completely aligned in time ("latency jitter"). This means that for the result at hand, the early part of the ERP effect might in principle hinge on a few critical words of very short (say, below 150 ms) duration only, in which case our second timing inference would perhaps not be legitimate. To address this issue, we recomputed the ERPs for only those items in which the critical words were of *at least* 550 ms duration. The outcome, displayed in Figure 13.3e, unequivocally confirms that at least some incoming words are really related to the wider discourse well before their offset.

A second caution is related to the possibility that interesting cognitive events may fail to show up in ERPs for reasons discussed before. As noted many times before (e.g., Rugg & Coles, 1995; Kutas et al., 1999), the consequence for timing inferences is that the earliest evidence for a differential effect to manipulation X can in principle only be taken as an *upper-bound* estimate of the time it takes the system to discover about X. For example, we cannot rule out the possibility that the comprehension system already detected a problem with discourse-incoherent words at, say, 100 ms from word onset, and that these earlier processing consequences simply failed to show up in ERPs as they occurred. In such a case, the later effect that we do observe could either be a direct reflection of some later process that *independently* detected the problem or of some downstream process whose operation is changed by the earlier detection.

A final caution relates to the temporal resolution of EEG and its derived measures (e.g., ERPs). EEG is usually advertized as providing continuous measurement with millisecond precision. This is entirely correct, and it is one of the things that give EEG its power as an on-line measure. However, although EEG can be recorded continuously throughout the unfolding sentence, the interpretation of these signals requires them to be anchored to (i.e., averaged relative to) the timing of some predefined critical event. In many ERP studies on sentence comprehension, there is only one such event per critical trial, usually a specific critical word—or morpheme therein—at which something might become apparent to the comprehension system:

a syntactic dead-end, a discourse-semantic anomaly, a lexical prediction shown wrong. Of course, there is the possibility to examine the impulse responses of the system at other words than the critical one, and sometimes this makes sense. In addition, one can examine slow shifts building up in the EEG across the sentence, relative to some critical word therein or the onset of the sentence (e.g., Kutas, 1997; Münte, Schiltz, & Kutas, 1998). But in all but the latter case, the continuous EEG is chopped into discrete segments to study the impulse response of the comprehension system at just one or a few word positions per critical sentence.

Moreover, the extent to which our ERP findings inherit the millisecond precision of EEG depends on how accurately we can define the critical events that elicit the blips going into an ERP average. In most research on sentence comprehension, including the example just discussed, ERPs are computed relative to a word's visual or acoustic onset. With written words flashed on the screen, word onset is not difficult to determine. However, in research with fully connected speech stimuli, there is usually considerable uncertainty as to where a spoken word begins, sometimes up to several tens of milliseconds for a particular word onset. Furthermore, with a spoken word it may take some—and, across items, some variable amount of—time before the truly critical information comes along in the acoustic signal. Researchers can try to locate this critical bit of information in the acoustic signal (e.g., an inflectional suffix) and reaverage relative to this, but if this is not realistic, they will need to accept some additional latency jitter as well as, potentially, some latency bias.[7]

Because coherent ERP effects can be found with spoken language, there is no reason to be overly concerned about the temporal uncertainties introduced by acoustic onset measurement. Also, the problem can sometimes be attenuated by, for example, using the same critical words and the same recordings in two different context conditions. However, it will be clear that the millisecond precision of EEG cannot always be exploited to the maximum.

13.3.3 Identity Inferences

To address the above timing question, all that mattered was the *onset* of the differential ERP effect elicited by words that did not fit the wider discourse, relative to the ERP elicited by fully acceptable words. Nothing in the discussion hinged on the *identity* of the effect, for instance, whether it was an N400 effect, a P600/SPS effect, or something else. However, one of the major strengths of ERPs is that every ERP effect has not only a time course, but also a polarity (Is it a positive or negative deflection?), a morphology (What is the shape of the waveform?), and a scalp distribution (the size of the effect at various recording locations over the scalp). In addition, some ERP effects (e.g., the N400 effect) can be characterized as modulations of some known underlying ERP component.

This rich multidimensional "signature" can provide cues as to the identity of the cognitive event at hand. If the signatures of two observed ERP effects are identical, the most parsimonious inference is that the cognitive processes that gave rise to these two effects are the same. Conversely, if two observed ERP effects have nonidentical signatures, the most parsimonious inference is that the cognitive processes that gave rise to these two effects are different in some way. There is a lot

to be said about these two inferences, most of which goes beyond the scope of this chapter (see Rugg & Coles, 1995; Urbach and Kutas, 2002). However, it is important to know that, by and large, the ERP research community interprets nonidentical polarity, morphology, and scalp distribution as reliable cues to the presence of *qualitatively* different cognitive processes (the reason being that if two ERP effects differ in any of these ways, the neural generators that gave rise to them must be at least partially nonoverlapping; Urbach & Kutas, 2002). Furthermore, if two ERP effects are deemed identical in terms of these three features, then (reasonable) differences in their time course or overall magnitude or both are most commonly interpreted as reflecting *quantitative* variations in when and to what extent a single underlying cognitive process is engaged.

The above identity logic can be used to make inferences about the cognitive processes underlying *any* two ERP effects E_X and E_Y observed in response to manipulations X and Y, within or across experiments. However, and quite fortunately, research has revealed that at least some observed ERP effect tokens tend to cluster in *types* such as the N400 effect or the P600/SPS effect, which, in turn, rather reliably map onto interestingly different classes of language comprehension events. Thus, in addition to comparing any two observed ERP effect tokens to each other, we can also compare a single ERP effect token to any of the known effect *types* and draw inferences on the basis of that.

Virtually all of the ERP research on the comprehension of sentences in discourse has posed the identity question in some way. For example, the ERP effect of a discourse-dependent semantic anomaly in Figure 13.3a has been compared to the effect elicited by *sentence*-dependent semantic anomalies, as well as to the effects of discourse-dependent *referential* and *syntactic* problems. I discuss each of these below.

13.3.3.1 Comparing Discourse- and Sentence-Dependent Semantic ERP Effects

One of the reasons for conducting an ERP study with discourse-dependent semantic anomalies was to see whether or not such anomalies would engage the same comprehension subsystem as so-called sentence-dependent semantic anomalies. We know that the latter elicit an N400 effect in ERPs, but would the former also do so? The St. George et al. (1994) research discussed before had already clearly suggested that the N400 is sensitive to global coherence. In our experiments (Van Berkum et al., 2003b) we wanted to have a closer look at this issue in a way that would be more directly comparable to how most of the sentence-semantic N400 research is conducted.

The discourse-dependent ERP effect whose timing we examined in the previous section had all the features of a classic *sentence*-dependent N400 effect (see Kutas & Van Petten, 1994, for review): a monophasic and relatively peaked negative deflection which emerged in the grand average at about 150–200 ms after acoustic word onset, peaked at about 400 ms, lasted for about 800–1000 ms, and reached its maximum over centro-parietal scalp sites. To support the comparison, we obtained a sentential N400 effect in the same lab under similar conditions. Figure 13.4 redisplays the ERP effect of a discourse-dependent semantic anomaly next to the

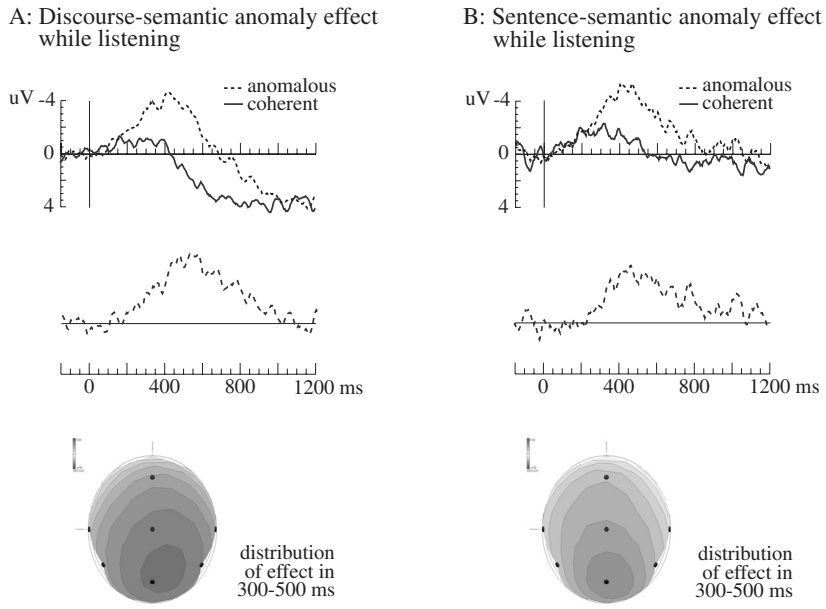

Figure 13.4 Discourse- and sentence-dependent semantic anomaly effects in spoken language comprehension. Top: Grand-average waveforms at Pz for anomalous and coherent words respectively. Middle: Corresponding anomalous–coherent difference waves. Bottom: spline-interpolated scalp distribution of the anomaly effect, based on mean difference-wave amplitude in the 300–500 ms latency range at each of 13 electrodes (6 of which are below the "equator" and therefore not visible; data from "When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect," by J. J. A. Van Berkum, P. Zwitserlood, C. M. Brown, and P. Hagoort, 2003b, *Cognitive Brain Research*, 17, 701–718).

N400 effect elicited by sentence-dependent semantic anomalies. As revealed most clearly in the difference waveforms and the associated scalp distributions, the ERP effect of a discourse-dependent semantic anomaly is identical to the classic sentence-dependent N400 effect in polarity, morphology, scalp distribution, and coarse timing. As shown in Figure 13.5, this equivalence was also observed in experiments with *written* language (Van Berkum et al., 1999c).

The equivalence of discourse- and sentence-dependent N400 effects suggests that within the domain of spoken as well as written language processing, the semantic comprehension process indexed by the N400 is *indifferent* to where the semantic constraints originally came from ("local" or "global" context), and simply evaluates the incoming words relative to the widest interpretive domain that is available. We briefly return to this in a later section, but refer to Van Berkum et al. (2003b) for a more detailed discussion.

13.3.3.2 Comparing Discourse-Dependent Semantic, Referential, and Syntactic ERP Effects

Whereas the ERP effect to discourse-dependent semantic problems turned out to be identical to the classic N400 effect observed for sentence-dependent semantic

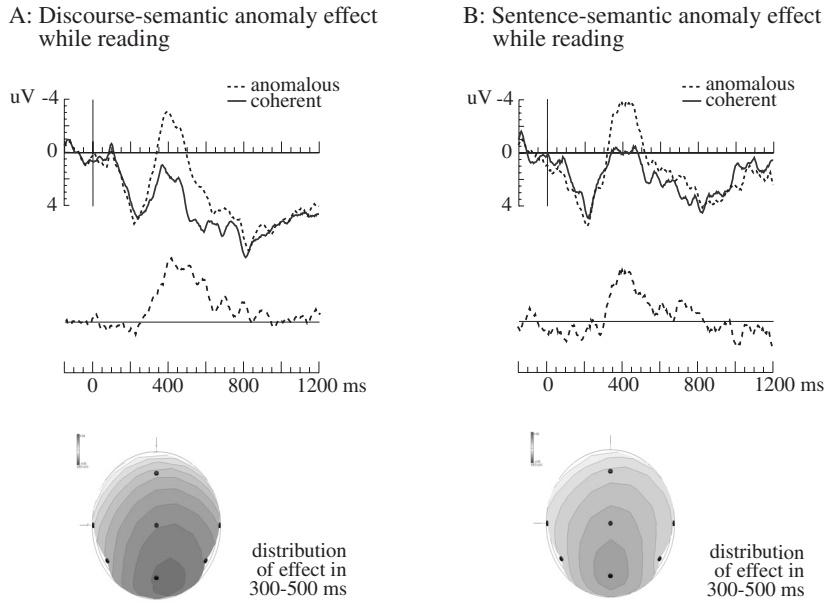

Figure 13.5 Discourse- and sentence-dependent semantic anomaly effects in written language comprehension. (Interpretation as in Figure 13.4; data at Pz, from "Semantic integration in sentences and discourse: Evidence from the N400," by J. J. A. Van Berkum, P. Hagoort, & C. M. Brown, 1999c, *Journal of Cognitive Neuroscience, 11*(6), 657–671).

problems, discourse-dependent *referential* problems elicited a very different ERP effect (Van Berkum, Brown, et al., 1999a; Van Berkum, Zwitserlood, et al., 2003A). The specific problem involved was a referential ambiguity, illustrated in (4), an example translated from Dutch.

(4a) *Just as the elderly hippie had lit up a joint, he got a visit from a friend and a nephew. Even though his friend had had quite a few drinks already, and the nephew had just smoked quite a lot of pot already, they insisted on smoking along. The hippie warned the* ***friend*** *that there would be some problems soon.*

(4b) *Just as the elderly hippie had lit up a joint, he got a visit from two friends. Even though one of his friends had had quite a few drinks already, and the other one had just smoked quite a lot of pot already, they insisted on smoking along. The hippie warned the* ***friend*** *that there would be some problems soon.*

Whereas in (4a), the critical noun "friend" was referentially unique, *two* equally eligible candidate referents had been supplied by the wider discourse context in (4b). As shown in Figure 13.6a for written-language input (Van Berkum et al., 1999a),

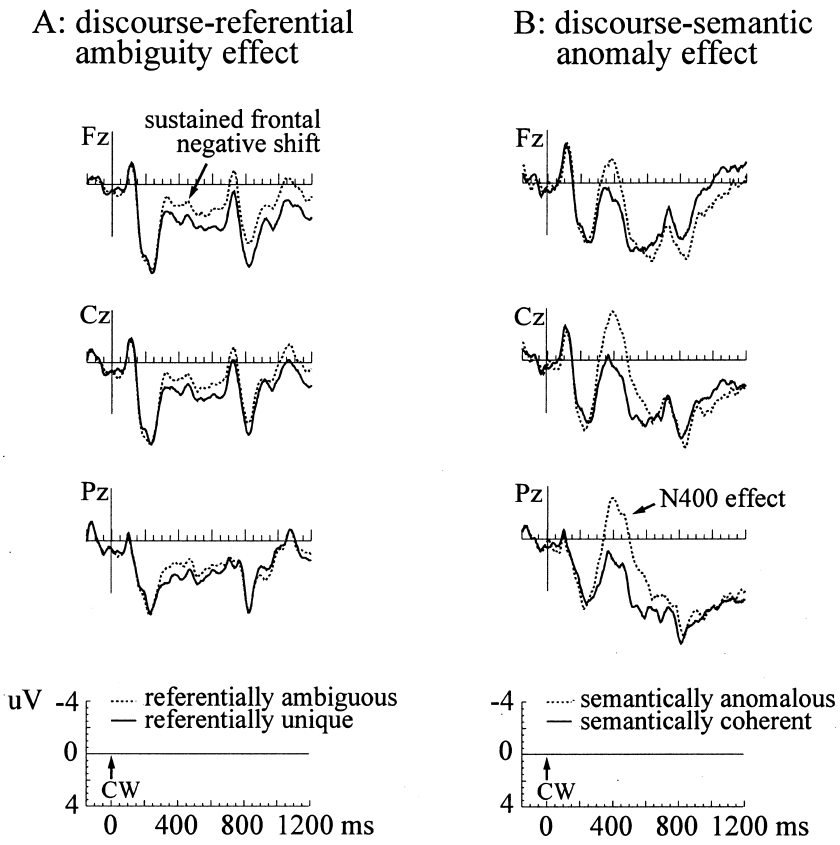

Figure 13.6 Panel A: ERPs to written words that were referentially unique or ambiguous with respect to the wider discourse. Data from "Early referential context effects in sentence processing. Evidence from Event-Related Brain Potentials," by J. J. A. Van Berkum, C. M. Brown, and P. Hagoort, 1999a, *Journal of Memory and Language*, 41, 147–182. Panel B: ERPs to written words that were semantically coherent or anomalous with respect to the wider discourse. Data from "Semantic integration in sentences and discourse: Evidence from the N400," by J. J. A. Van Berkum, P. Hagoort, and C. M. Brown, 1999c, *Journal of Cognitive Neuroscience, 11*(6), 657–671. Effects are shown for a frontal, central, and parietal midline site.

the processing consequences of such referential ambiguity showed up immediately in ERPs, right at the critical noun.

The early onset of the referentially induced ERP effect reveals that people rapidly discover whether a singular definite NP is referentially ambiguous or not. Moreover, whereas semantically problematic words elicit an N400 effect (redisplayed in Figure 13.6b for the same three midline sites), referentially ambiguous nouns elicited a *qualitatively different* effect: a frontally dominant and sustained negative shift in ERPs, starting at about 300 ms from word onset during reading (Van Berkum et al., 1999a), and about 300–400 ms from acoustic word onset during

listening (Van Berkum et al., 2003A). As discussed in more detail elsewhere, this nonidentity suggests that the processing implications associated with difficulties in making sense and establishing reference are at least partially distinct, and that, perhaps not too surprisingly, the N400 effect therefore does not exhaustively reflect *all* aspects of conceptual integration.

In the same experiment, discourse-induced referential ambiguity also had an immediate impact on how the parser analyzed a subsequent local *syntactic* ambiguity. For instance, if *the friend* in *The hippie warned the friend* ***that*** … was referentially ambiguous, the parser was more inclined to take the syntactically ambiguous word *that* as the beginning of a referentially restrictive *relative clause* (*… that had had quite a few drinks*) than as the beginning of a complement clause (*… that there would be some problems soon*). This discourse-induced preference for a relative-clause analysis was revealed by the fact that in relative-supporting contexts like (4b), a subsequent critical word that ruled out this analysis by signalling a *complement-clause* continuation (*there* in the example) elicited a P600/SPS effect, indicating that the parser had been led down a garden path by the two-referent discourse context. Conversely, if the discourse context favored a complement-clause analysis at the word *that* (the one-referent context in 4b), a P600/SPS effect was instead elicited by a word that signalled a *relative-clause* continuation. This crossover pattern of P600/SPS results is depicted in Figure 13.7 (see Van Berkum et al., 1999a, for details and two additional conditions).

Words at which the syntactic analysis currently pursued by the parser runs into trouble reliably elicit a P600/SPS effect (see Hagoort et al., 1999; Osterhout & Hagoort, 1999, for a review). Hence, if one predicts—as we did—that certain features of the discourse context will lure the parser into an analysis that runs into a dead end at particular critical words, then one should find P600/SPS effects at those words, and at those words only.[8] This is precisely what happened. Obtaining an *N400* instead of a P600/SPS effect at these critical words would have forced us to revise our interpretation. The *identity* of these ERP effects therefore provides an important constraint on interpretation.

As illustrated by Figure 13.8 for written-language input, the three discourse-dependent ERP effects depicted in Figures 13.5a, 13.6a, and 13.7 were actually obtained in a single ERP experiment (see Van Berkum et al., 1999a, 1999b). ERPs thus allow us to simultaneously yet *selectively* examine how prior discourse modulates the referential, syntactic, and semantic analyses as a local sentence unfolds within the same subjects and the same texts. Given that discourse-level comprehension is bound to involve a complex dynamic interplay of many different processes overlapping with each other as well as with the input (Figure 13.1d), this ability of ERPs to simultaneously yet selectively track at least some of those processes with high temporal precision can be of considerable importance.

Several important cautions remain to be made. If a particular ERP effect observed in some experiment is labeled as, say, a "P145," and there have been no earlier reports using the same label, then the term is probably intended as descriptive of the effect token, meaning no more than a positivity peaking at 145 ms. However, if an observed ERP effect is labeled with the name of a *familiar effect type*, such as an N400 effect, then the researcher usually claims that the observed effect token

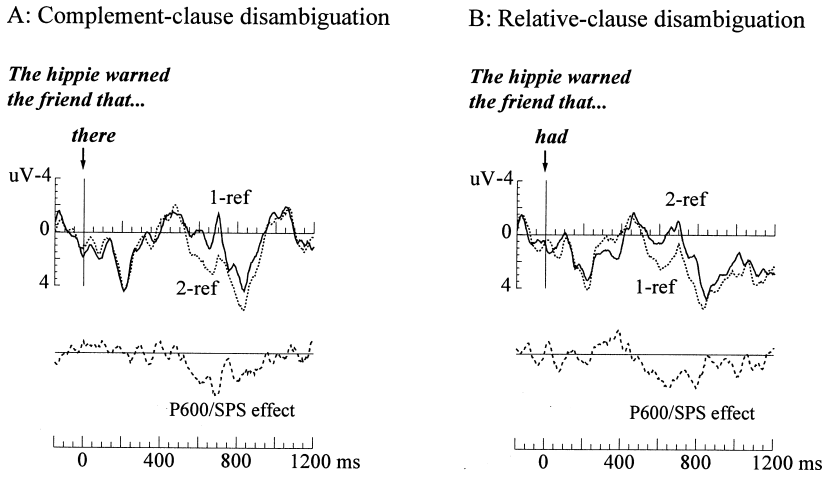

Figure 13.7 Panel A: P600/SPS effect to written complement-clause disambiguation in a two-referent context. Panel B: P600/SPS effect to written relative-clause disambiguation in a one-referent context. Data, at Pz from "Early referential context effects in sentence processing: Evidence from event-related brain potentials," by J. J. A. Van Berkum, C. M. Brown, and P. Hagoort, 1999a, *Journal of Memory and Language, 41,* 147–182.

matches the well-known effect type to a sufficient degree to justify *classifying* the token as an instance of this type. Within a particular domain of inquiry, ERP researchers will often agree on whether the claim is reasonable or not, given the effect type at hand and the variability observed across other accepted instances of the type. However, it is important to realize that the *ERP waveforms themselves do not come with labels*, and that it is the researcher that provides them. Even when backed up by the entire ERP community, the claim that is implicit in such classification may turn out to be wrong, either because the specific effect was misclassified ("It looked exactly like an N400 effect, but it turned out to be a perfect look-alike generated by a completely different process"), or because later research reveals that our partitioning of known effects needs reconsideration.

Related to the latter, note that it is almost never the case that two ERP effect tokens classified as identical *are* in fact completely identical. Many of the differences can be reasonably viewed as "noise" (e.g., residual noise that failed to cancel out in the averaging process), but many others reflect potentially interesting variations that the research community has decided to ignore "for the time being" to keep things tractable.

Finally, clustering ERP effects is more than sorting one's stamp collection on shape and color, for ERP effects that are considered to be equivalent are also considered to reflect the same underlying functional process. Effect classification, hence, always involves a claim about the functional interpretation of the ERP effect (or component) at hand. Those who follow the psycholinguistic ERP literature will have noted that, even for extensively studied phenomena like the N400 and the P600 effect, the exact functional interpretation is still under debate. The fact that one

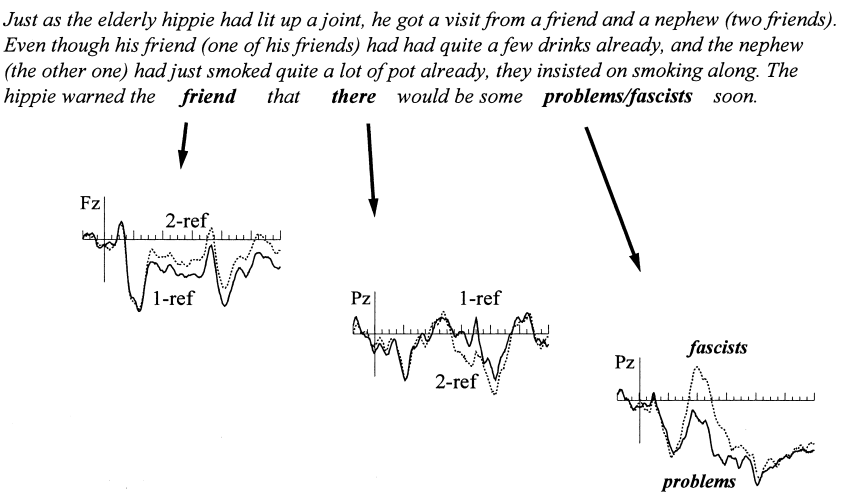

Figure 13.8 From left to right: A sustained frontal negative shift to a discourse-induced referential problem ("friend" is referentially ambiguous in the 2-referent context), a P600/SPS effect to a discourse-induced *syntactic* problem ("there" rules out the provisional relative-clause analysis pursued at "that" in the 2-referent context), and an N400 effect to a discourse-induced *semantic* problem ("fascists" does not fit the wider story context); see text for explanation. The example item is shown here in several variants (1- and 2- referent contexts, coherent/anomalous ending), but any one subject saw only a single variant. Data from "Early referential context effects in sentence processing: Evidence from event-related brain potentials." by J. J. A. Van Berkum, C. M. Brown. & P. Hagoort, 1999a, *Journal of Memory and Language*, 41, 147–182. and from "Semantic integration in sentences and discourse: Evidence from the N400." by J. J. A. Van Berkum, P. Hagoort, & C. M. Brown (1999a). *Journal of Cognitive Neuroscience*, 11(6), 657–671.

cannot as yet *exhaustively* characterize the system that generates, say, the P600 effect, does *not* mean that research cannot exploit the effect at hand to address some psycholinguistically relevant issue (see also note 8). However, our limited understanding of even the "well-known" N400 and P600 effects serves as a reminder that the linking of ERP effects to underlying functional processes is difficult business. In all, the search for meaningful (and manageable) clusters in the set of observable ERP effects is a nontrivial job. The quality of the clusters we have defined at any given point in time will ultimately be judged by whether we make any scientific progress with it or not.

13.3.4 Magnitude Inferences

In none of the examples of ERP research on discourse-level comprehension discussed so far did the *size* of a particular ERP effect matter to the inferences we wanted to draw. So, even if the discourse-dependent N400 effect would have been bigger than the sentence-dependent N400 effect, this would not have prevented us from classifying both as N400 effects, and as such license our inference that both types

of anomalies engage the same processing system. However, some inferences do critically depend on the magnitude of an effect. A recent discourse-level ERP study by Federmeier and Kutas (1999a) exemplifies this. Federmeier and Kutas asked subjects to read stories such as the one in (5).

(5) *They wanted to make the hotel look more like a tropical resort.*
So along the driveway they planted rows of ***palms/pines/tulips***.

Like the example story, each critical story either ended with a highly expected word (*palms*), a "within-category violation" that belonged to the same semantic category as the expected control word (e.g., *pines*, a tree as well), and a "between-category violation" coming from a different semantic category (e.g., *tulips*). Relative to the expected control word *palms*, both types of anomalous words were semantically incoherent with respect to the wider discourse. And, in line with other work (St. George et al., 1994; Van Berkum et al., 1999c), both elicited a clear discourse-dependent N400 effect. However, the N400 effect elicited within-category violations (e.g., *pines*, a tree as well) turned out to be *smaller* than the N400 effect elicited by semantically incoherent words coming from a different semantic category (e.g., *tulips*). Because in off-line pretests both types of anomalous words had on average been rated as *equally* implausible and unpredictable, Federmeier and Kutas argued that the differential N400 effect could not solely derive from differences in how the words related to the dynamic discourse context, and therefore had to be interpreted as evidence for the context-independent involvement of permanent semantic memory structure. Regardless of whether their interpretation is correct,[9] the work by Federmeier and Kutas illustrates how the magnitude of an ERP effect can support particular inferences about discourse-dependent comprehension.

Whereas the above inferences require only coarse-grained ordinal differences in the magnitude of an ERP effect, it is sometimes relevant to plot the size of an ERP effect in a more fine-grained manner, as a function of some interval-scale factor (e.g., 10 equidistant increasing levels of cloze probability). Unfortunately, within the domain of discourse-level comprehension, such parametric designs are very difficult to realize. The reason is that to get a relatively noise-free ERP average for conditions X_1 … X_N, one needs the raw EEG of a large number of critical trials (say, ~30 at least) *per condition*. Furthermore, if every critical condition needs its own control, the number of required trials doubles. If every trial requires a *piece of discourse*, things really get out of hand. For some issues, the problem can perhaps be attenuated by including multiple critical trials within a single discourse (cf. St. George et al., 1994). In general, however, it will not be easy to use fine-grained parametric factors in a discourse-level ERP experiment.[10]

Interval-level effect size also matters in another type of ERP design. If the processes that respond to manipulations X and Y are functionally independent, the corresponding ERP effects E_x and E_y should be strictly additive, that is, the effect E_{xy} to a *combined* XY manipulation should equal $E_x + E_y$. Among other things, this additivity logic has been used to study the functional independence of sentence-level semantic and syntactic processing (e.g., Osterhout & Nicol, 1999). Because it often requires a mere 2 x 2 design, it can also be used to study functional independence

issues in discourse comprehension. However, as with the classification of ERP effect tokens, there is some uncertainty as to how strict one needs to be in applying the logic to actual ERP datasets (notably, how much divergence from perfect additivity can we do away with as "noise?"). Although this by no means invalidates the approach, it does sometimes make it a little harder to use.

13.3.5 Neuronal Generator Inferences

So far, we have discussed inferences based on the presence, timing, identity, and magnitude of an ERP effect. However, one additional question has been conspicuously absent from the discussion until now: What specific areas of the brain generate a given ERP effect E_x? In order arrive at a complete understanding of language-relevant ERP effects, we need to know what the underlying neural generators are. Unfortunately, it has turned out to be very difficult to determine the neuronal sources of a given scalp-recorded ERP effect from the nature of the effect itself. The reason is that any given ERP effect can be explained by a principally unbounded number of different generator configurations (see Kutas et al., 1999 for a good explanation of this so-called inverse problem; see also Fabiani, Gratton, & Coles, 2000; Urbach & Kutas, 2002). As a consequence, for example, an ERP effect that is largest over Pz need *not* have its generator right below (or even close to) Pz. The N400 effect is a case in point. Within the domain of language comprehension, this effect is usually largest at midline centro-parietal sites (say, close to Pz). Intracranial recordings, however, suggest that the N400 effect is generated by a number of brain areas, bilaterally distributed, and none of which being particularly close to Pz (see Kutas & Federmeier, 2000, and references therein). On their own, therefore, scalp-recorded ERPs do not straightforwardly inform us about which areas of the brain gave rise to them.

One way to use ERPs to make (coarse) spatial inferences without tackling the inverse problem is illustrated a recent discourse-level experiment on *hemispheric processing asymmetries* (Federmeier & Kutas, 1999b). Following up on the ERP study that I used to illustrate magnitude inferences with, Federmeier and Kutas now presented the critical words of stories such as in (5) in either the left or right visual hemifield, causing these words to be initially processed by the right and left hemisphere, respectively. Interestingly, only for words that were initially presented to the *left* hemisphere did within-category violations (*pines)* elicit a smaller N400 effect than between-category violations (*tulips*). For words initially presented to the *right* hemisphere, the N400 effects elicited by within-category violations (*pines*) and between-category violations (*tulips*) were of equal size. Based on these and other aspects of their ERP data, Federmeier and Kutas have proposed that whereas the right hemisphere would engage in discourse-driven plausibility-based integration, the left hemisphere might operate in a more predictive mode of processing, using the ever-changing discourse context as well as the more permanent structure of semantic memory as its source. Whether this specific interpretation is correct remains to be seen (cf. note 9), but these ERP results do suggest an asymmetry in how the two hemispheres contribute to discourse-level comprehension. Note, though, that the critical spatial inference hinges on a differential effect of left/right hemifield

presentation, which happened to show up in ERPs but might have shown up in some other measure as well.

13.3.6 What Can We Infer About Discourse-Level Comprehension?

It is beyond the scope of this chapter to discuss the specific theoretical implications of the ERP work on discourse-level comprehension reviewed so far. However, before evaluating the utility of ERPs relative to other available on-line measurement tools, let me just lay out some of the most obvious implications. In a nutshell, the work reviewed suggests that the impact of discourse-level information is immediate, penetrating, proactive and—perhaps, surprisingly, in a chapter like this—maybe not so special. As for the first, the ERP data depicted in Figure 13.8 (as well as all the other discourse-dependent ERP effects observed so far, e.g., St. George et al., 1994) suggest that incremental interpretation holds "all the way up." The words of an unfolding sentence are immediately mapped onto a representation of the wider discourse in terms of their syntactic, their semantic, and their referential implications. Not only does the impact of discourse show up at the very first word where it can be expected in each case, but it also shows up very rapidly, within only a few hundred milliseconds after onset of the word at hand.

Because our P600/SPS data (cf. Figure 13.7) suggest that the analysis of a local syntactic ambiguity is not insensitive to discourse-level referential constraints, the ERP data also suggest that discourse-based comprehension is *penetrating* lower levels of analysis. One may debate about whether the ERP research at hand has tapped the earliest possible stage of parsing or just missed it (see the exchange between Brysbaert & Mitchell, 2000, and Van Berkum, Hagoort, & Brown, 2000 for an example), but the fact remains that discourse extremely rapidly affects the parsing process, being able to affect the provisional resolution of a syntactic ambiguity right at the word at which it emerges and, as discussed in detail elsewhere (Van Berkum et al., 1999b; see also Brown, Van Berkum, & Hagoort, 2000) even before a locally available syntactic gender cue is brought to bear on the analysis. Because the explanation of these findings involves a discourse-based anticipation of plausible syntactic structures, they can also be taken to suggest that discourse-level comprehension is *proactive*. Recent ERP evidence for discourse-based anticipation of specific upcoming words, shown in Figure 13.2, points in the same direction. The ERP effect at hand is relatively fragile and needs to be consolidated in follow-up experiments. However, with the same materials, we have already found that unexpectedly inflected adjectives also slow down the reader in a self-paced reading task (albeit only at a second adjective; see Van Berkum et al., 2002a; 2004). Thus, we have converging evidence from different methods that both listeners and readers can routinely exploit their knowledge of the wider discourse to anticipate upcoming language.

Finally, the equivalence of sentence- and discourse-dependent N400 effects, illustrated in Figures 13.4 and 13.5, can be taken to suggest that, in a sense, discourse-level comprehension *may not be all that special*. The comprehension

process indexed by the N400 does not seem to care about whether the semantic constraints come from the first few words of a single unfolding sentence or from some wider discourse. What the data suggest is that it simply evaluates the incoming words relative to the widest interpretive domain available. One way to account for this is to recognize that, although sentence-semantic violations are local or sentence-internal in that it takes a single sentence to elicit them, the unfolding sentence itself draws in a much wider interpretive context, involving knowledge of the world, the language at hand, the present situation, and possibly even the absent speaker (see Van Berkum et al., 2003b , for a discussion in terms of "common ground"; Clark, 1996). If prior discourse simply adds a few more bits of information to this ever-present vast interpretive context, the equivalence of sentence- and discourse-dependent semantic anomaly effects should come as no surprise.

13.4 SO, CAN ERPS BE USED TO STUDY DISCOURSE-LEVEL COMPREHENSION?

One reason why ERP researchers may so far have stayed away from discourse-level comprehension is that working with ERPs is difficult enough as it is. With so many word- and sentence-level issues unresolved, why not get a firm grip on these first before getting into the complexity and potential intractability of discourse-level comprehension? This is a reasonable consideration. However, the work I reviewed shows that ERP research with discourse-level materials is feasible and can lead to systematic, tractable, and interpretable effects which support all of the various types of inferences that ERPs can be used for. Of course, the fact that we can do it does not necessarily mean we should. Working with ERPs is really difficult, and if we can get all the "goodies" with self-paced reading in 6 weeks instead of 6 months, why bother with ERPs? In the end, therefore, the utility of ERPs to track discourse-level comprehension depends on the pros and cons of using this method relative to the other tools available. The purpose of this section is to examine some of these trade-offs.

13.4.1 Advantages

13.4.1.1 Sensitivity Without Artificial Response Task

The language comprehension system did not develop to support timed responses to a concurrent probe word or control the index finger in a self-paced reading task. Of course, artificial response tasks by no means necessarily generate bad data. For example, many of the findings initially obtained with the self-paced reading task turned out to generalize to less artificial reading situations (Mitchell, chapter 2, this volume). However, every artificial response task does carry a risk of bringing artificial response strategies along with it. With ERPs, one can study the processes involved in discourse-level language comprehension by giving subjects just one simple task: read or listen for comprehension. Note that this happens to be what the

system is for. This seems like as good a reason as any to consider ERPs as a tool to study discourse-level language comprehension.

As it happens, not everybody agrees that the absence of a secondary task is a virtue. ERP research sometimes faces skepticism because, without having assessed the quality of comprehension (e.g., through comprehension questions), how can we know what the subjects are doing? There may well be situations in which the concern is real. For example, if one's aim is to compute ERPs as a function of whether some difficult sentence was interpreted correctly or not, then this must be independently assessed at every trial. However, the how-do-we-know-what-they-are-doing? concern is often rather overstated. For one, it is not as though *comprehension questions* provide privileged access to what subjects are doing while they work on the input—if so, why need on-line measures to begin with? Also, to the extent that secondary tasks tell us what subjects are doing because *we told them what to do* (a nonnegligible risk, for example, in some types of probe tasks), then what they are doing may not be so interesting.[11]

One advantage of research with discourse-level input is that it is possible to have relatively interesting stimuli (as opposed to, say, a list of unrelated words) that can be used to draw subjects into the natural task of language comprehension. Furthermore, in a well-designed study with non-zero results, the ERP data *themselves* can be used to verify that subjects were drawn into this task to a sufficient degree. An N400 effect elicited by discourse-dependent semantic anomalies, for example, tells us that subjects were interpreting our stories to a sufficient degree to tell an anomalous word from a coherent one. In all, therefore, there is no principal reason why ERP measurements cannot stand on their own.

13.4.1.2 High Temporal Resolution

The ERP research with discourse-dependent semantic anomalies (see Figure 13.3) showed that ERPs are good at supporting two different kinds of inferences about the timing of a discourse-relevant processing event. One is how soon in the sentence, that is, at which word, the effect at hand shows up. In terms of the analysis of on-line measurement supplied in section 2, this type of temporal sensitivity involves establishing *which word-elicited incremental impulse response* turns out to be sensitive to the manipulation at hand. ERPs can do this well, but so can many behavioral measures (see various chapters in this volume for thorough discussion).[12]

However, bearing the caveats made before in mind, ERPs are particularly good at supporting a second type of timing inference, which is how soon after onset of (or some other relevant point within) the critical word the processing consequences of a manipulation show up. In terms of the earlier analysis, this involves establishing *when in the relevant word-elicited impulse response* things begin to happen. ERP effects are immediately linked to differential activity in the associated neuronal generator, with 0 ms delay (see also Osterhout et al., chapter 14, this volume). Hence, if some ERP effect emerges at time *t*, then we know the manipulation matters to the system at time *t*. Because behavioral measures tap cognition through its consequences for actual behavior, the timing of a cognitive event must in these measures always be inferred by subtracting some approximate correction factor (e.g., ~ 200 ms estimated saccade

programming time). For questions that require precise timing *within* a word's impulse response, therefore, ERPs have a distinct cutting edge over the behavioral alternatives.

13.4.1.3 Selective Tracking and Process Identification

Figure 13.8 reveals what is perhaps the biggest advantage of using ERPs to track discourse-level comprehension as it unfolds: the possibility to *selectively* track specific aspects of the comprehension process. To make the point, consider what would happen if the example item in Figure 13.8 took part in a self-paced reading experiment. If this task is sensitive enough to pick up on referential ambiguity at the noun *friend*, a referentially induced syntactic dead-end at the word *there*, and the semantic problem emerging at *fascists*, we would see three delays in reading time, and that is it. There would be nothing *in the data* telling us that we are looking at three qualitatively different cognitive events. With ERPs, such information comes "for free." Although we do not yet fully understand the nature and the degree of language-specificity of the processes that underlie each of the ERP effects displayed in Figure 13.8, the fact is that within the domain of discourse-dependent language comprehension, ERPs clearly allow one to selectively tap into aspects of the ongoing referential, syntactic, and semantic analyses. If any feature of ERPs is going to help us disentangle and make sense of the many different processes interacting in discourse-level comprehension, it will be this one.

The fact that ERP effects "come in different flavors" also implies that there is information in obtaining *identical* ERP effects. Our finding that sentence- and discourse-dependent anomalies elicit the very same N400 effect (Figures 13.4 and 13.5), for example, clearly suggests that from the perspective of the interpretive process indexed by the N400, there is no functional distinction between a context set up by a single unfolding sentence and one set up by a larger piece of discourse. Also, the equivalent N400 findings with spoken- and written-language input (compare Figures 13.4 and Figure 13.5) suggests that the underlying machinery reflected by this component is basically modality-independent. Now, all this may or may not be considered a big surprise (as it happens, much of this was no surprise to me), but the fact is that the observation of, say, four exactly identical reading time delays would not in itself have licenced these inferences about process equivalence (see Hess, Foss, & Carroll, 1995, for an example of how hard it is to license such inferences by means of response-time methodology).

13.4.1.4 Easy Cross-Modal Investigation

Eyetracking in reading is necessarily limited to the comprehension of written language, and so is self-paced reading. The use of head-mounted eyetracking in a visual-world setting, on the other hand, is limited to the study of spoken-language comprehension. Of all the tools currently used to track sentence comprehension on-line, only ERPs can be used with spoken as well as written language input. In addition, it can be used with sign language, as well as potentially relevant nonlinguistic forms of input (e.g., a cartoon or video sequence; West & Holcomb, 2002; Sitnikova et al., 2003) For issues in which generalizations across modality are important, the possibility to do so within the same measure is highly attractive. Moreover, and

returning to the equivalence of discourse-dependent N400 effects across modality exemplified in Figures 13.4 and 13.5, ERPs provide the means to *directly* test to what extent the processes involved in listening and reading are (non-) identical.

Because eye movements generate deflections in the EEG that are much larger than the ERP effects we are looking for, it is important to avoid them while recording EEG. It is for this reason that virtually all ERP research with written sentences uses Serial Visual Presentation (SVP) in which the words of a sentence are consecutively presented in the center of the screen at some fixed rate (e.g., 600 or 250 ms per word). Psycholinguists who examine reading under more natural conditions are often rather skeptical about this way of presenting written language. And, indeed, it does make sense to be concerned about whether SVP might affect the results obtained. As it happens, however, there is now a substantial database of ERP studies that reveal surprisingly comparable findings for fully natural spoken-language input and written-language input presented with SVP (e.g., for discourse-level N400 effects, compare Figures 13.4 and 13.5, derived from Van Berkum, Zwitserlood, et al., 2003b, and Van Berkum, Hagoort, et al., 1999c; see also Federmeier & Kutas, 1999a; Federmeier et al., 2002; for sentence-dependent N400 effects, see, for instance, Hagoort & Brown, 2000a; for P600/SPS effects, see, for instance, Hagoort & Brown, 2000b; Osterhout & Holcomb, 1992, 1993; for memory-related slow ERP shifts, see, for example, Kutas, 1997). Of course, the fact that *some* sentence-level ERP results generalize from SVP to speech input does not guarantee that *every* new result will do the same (prosodic factors being an obvious complication). However, it *does* mean that we can now no longer discard an ERP result simply because it was obtained with SVP.

13.4.2 Drawbacks

13.4.2.1 Many Constraints on Experimental Design

By far the most important drawback of ERPs is that, as with other neuroimaging measures, they severely constrain the shape of an experiment. One such constraint already mentioned before is the need to have a much larger number of critical trials per condition (at least ~30–40) than is common in most behavioral designs. With a single bleep, flash, or word as critical stimulus, this is not too much of a problem. However, with a 25-s mini-story in every trial and, say, 5 s of overhead around each trial, a simple 2 x 2 design is already looking at 60–80 min pure time-on-task, excluding breaks, practice session, and filler trials. This is a severe constraint on the use of ERPs to study discourse-level comprehension.[13] Having subjects come back to the lab several times is one solution, but because of the time lost in applying and removing electrodes (~1–2 hr per session), this is not very practical. It also leaves subjects with plenty of time to think about the experiment and possibly even discuss it with others. The only other way to attenuate the problem is to make the stories as short as possible, or, if possible, to obtain multiple critical measurements in a single story (cf. St. George et al., 1994).

The limited space for filler trials can be a real concern. However, even a complete lack of fillers does not necessarily invalidate one's design (see Van Berkum et al., 2000). Fillers are usually employed to hide critical features of the design and to

make sure that unwarranted strategic processing does not pay off. However, both functions can also be implemented in other ways. Also, critical features of the design can often be hidden by having more variation in how particular critical stimulus features are realized (e.g., avoid using the same critical verb in all items) by having intrinsically interesting and maximally variable content, by counteracting the regularities of critical sentences in noncritical sentences elsewhere in the story, and by adding a salient but harmless dummy manipulation (e.g., a semantic anomaly in the tails of some stories) to distract the subject. The joint effectiveness of these measures can be assessed in a well-conducted postsession interview. In addition, one can often quite easily use counterbalancing, as well as many of the measures just mentioned, to make sure that even if critical features *do* get noticed, there would be nothing to be gained by using some unnatural comprehension strategy. So, even though a design without filler trials will make many experimentally trained psycholinguists somewhat uneasy (see Brysbaert & Mitchell, 2000, replied to in Van Berkum et al., 2000), there is no particular reason to stick to a simple "if it ain't at least 50-50, it's no good."

ERPs impose several other constraints that complicate discourse-level research. One is the requirement to sit still during recording. With long trials, the problem can sometimes be attenuated by designating a part of each story as non-critical, and by communicating to the subject this opportunity to relax by means of a visual cue. A more serious consequence of this requirement is that it makes it *very* difficult to have the discourse occur as part of some more realistic setting such as a referential communication task. To me, a huge, and by far the biggest, advantage of eyetracking in so-called visual-world experiments over *all* other on-line measures is that in these experiments linguistic utterances can be made relevant to some larger collaborative project ("Now pick up the queen of spades and put it below the ace …"). Although the collaborative task may itself be artificial, once it has been accepted as a legitimate task by the subject, it does provide the latter with a strong natural motivation to process language ("language as action," cf. Clark, 1996). It remains to be seen to what extent ERPs can be taken this far beyond the single sentence.

13.4.2.2 No Intuitively Obvious Interpretation

It is sometimes said that relative to behavioral measures, ERPs provide a more direct window onto cognition because ERPs directly tap into the seat of cognition, the brain. In terms of *timing*, and in terms of whether a behavioral response (e.g., a saccade, a button press) mediates between cognition and its measurement, ERPs are indeed much more immediately related to cognitive events than behavioral measures. However, the general claim that brain measures are somehow privileged can also easily be reversed. Cognition evolved for behavior, and, moreover, for adequate and timely behavior. Thus, although the brain did not specifically evolve to tap a space bar in the self-paced reading task, one could argue that *speed and accuracy measures* are in a sense the more privileged windows onto cognition. Note that related to this, speed and accuracy have an intuitively obvious interpretation: If things go slower or result in more errors, the comprehension system is having a harder time. ERPs lack such an intuitively obvious interpretation. Positive ERP deflections do not necessarily mean the system is doing better, and negative deflections do not tell us

it is in trouble. Moreover, although accepted by many as a reasonable working assumption, it is not even clear to what extent an increase in the amplitude of some particular ERP component should always be interpreted as evidence that the underlying system "has to work harder." I suspect that this lack of intuitively obvious meaning ("It's just squiggles") is one of the main reasons why ERP research is frequently ignored by other researchers in the field (and why fMRI and PET results are more easily accepted). The upside, of course, is that because ERPs are not unidimensionally tied to more or less trouble, we may also be able to go beyond the associated unidimensionality in our thinking about comprehension (see Mitchell, chapter 2, and Boland, chapter 4, both for a similar point).

13.4.2.3 Slow Research Cycle

A final drawback of ERPs is that it is cumbersome research, taking more time to conduct than the average eyetracking experiment, and much more time than an average self-paced reading experiment. Because an ERP design needs many trials, and hence, in this particular field of application, many carefully handcrafted and pretested discourses, ERP research with discourse-level materials can become a real problem (with sometimes over half a year dedicated to the construction and validation of materials alone!). Apart from the obvious drawbacks for individual researchers, as well as for students looking for a small-scale project, the implication is that it takes a much larger research effort to systematically chart some empirical phenomenon. Hence, even if the community has lots of good ideas to track discourse-level comprehension by means of ERPs, progress will be slow.

As with all other tools available to track discourse-level comprehension as it unfolds, there are some very good reasons to use ERPs, as well as some very good reasons to not exclusively rely on them. Only an adequate understanding of each of the available tools can help the researcher decide which (combination) of them is most appropriate, given the issue at hand.

13.5 WHAT IS ON THE HORIZON?

The field of neuroimaging is changing rapidly. Whereas, up until only a decade or so ago, ERPs were just about "the only neuroimaging game in town" (for a psycholinguist, that is), several new methods have since become available (e.g., MEG, fMRI, and TMS). Other innovations are taking place within the EEG research community itself. To my knowledge, and with the exception of fMRI, most of the new tools have not yet been used to investigate discourse-level comprehension, and I will therefore only briefly comment on their potential. However, I believe it is only a matter of time before measures like MEG and TMS, as well as EEG-based measures that complement the simple computation of ERPs, will begin to enrich our understanding of the architecture and neural substrate of discourse-level comprehension.

13.5.1 EEG

As mentioned before, scalp-recorded ERPs cannot straightforwardly inform us on their own about which areas of the brain give rise to them. However, ERPs are no longer on their own. Our knowledge of the brain is increasing, and so are our capabilities to identify specific areas of the brain that are recruited in, say, a language-comprehension task. With intracranial recordings, for example, progress has been made in locating at least some of the generators of the language-relevant N400 effect. In addition, a sizeable research effort currently focuses on how to use data from other imaging techniques (notably fMRI and MEG) to help identify the neuronal generators of particular EEG phenomena (see also Osterhout et al., chapter 14, this volume; Kutas et al., 1999).

Another interesting development concerns the revival of the so-called background EEG. For quite a long time, researchers only looked at this to examine the energetic state of cognition, i.e., whether somebody was highly attentive, relaxed, drowsy, or asleep. Other than that, it was viewed as noise to be canceled away by averaging. Over the past few years, however, EEG researchers have discovered that EEG oscillations ("brain rhythms") can inform us about much more than the energetic state of an individual (e.g., Pfurtscheller & Lopes da Silva, 1999; Tallon-Baudry & Bertrand, 1999) So far, there have been only a few attempts to examine these oscillatory EEG phenomena during language comprehension (e.g., Bastiaansen, Van Berkum, & Hagoort, 2002a, 2002b; Weiss & Mueller, 2003), and, to my knowledge, none of this work has gone beyond the single sentence. However, this is bound to change in the very near future. EEG oscillations are believed to play an important role in binding various sorts of information distributed across the brain together in a single representation (cf. Tallon-Baudry & Bertrand). They have also been associated with operations involved in episodic memory storage and retrieval (see Bastiaansen & Hagoort, 2003, for a language-oriented review). Both of these associations lead one to expect that oscillatory EEG phenomena will turn out to be relevant to understanding the architecture of discourse-level comprehension.

Recent evidence (Makeig et al., 2002) suggests that at least one very early ERP component might actually be the result of a stimulus-induced short-lived synchronization of the oscillatory EEG, rather than the average reflection of a transient "blip" that has nothing to do with the background EEG. This suggests that, in some cases, there may be a much more intimate relationship between ERPs and EEG background oscillations than traditionally assumed. Because of their shape and time course, language-relevant effects like the N400 and P600/SPS are unlikely to receive a similar reinterpretation. However, the finding does illustrate the importance of exploring ERPs and EEG oscillations in closer relationship.

A final development also concerns the interpretation of ERPs. As discussed before, qualitatively different ERPs are assumed to reflect qualitatively different processes. However, as discussed by Rugg and Coles (1995), it is not always obvious what the latter means in terms of an interesting model of cognition. For example, recent findings (Pulvermüller, 2001) suggest that words whose meaning involves movement of different part of the body (*talking*, *walking*, *reaching*) elicit ERPs that differ in scalp distribution in a way that might be associated with different areas of

the human motor cortex that control the movements involved. The example should remind us that the brain's electrophysiology also codes *content*, and that, as a result, ERPs cannot always simply be assumed to reflect processes irrespective of the content involved. In other words, we should not expect a simple mapping between the various different ERPs that we see in our data and the various different "boxes" ("modules," "processors") that we define in our information processing models.

13.5.2 Other Imaging Methods

Over the last few years, researchers have begun to use functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) to examine the neural substrates of discourse-level comprehension (e.g., Ferstl & Von Cramon, 2001; Gallagher et al., 2000; Maguire, Frith, & Morris, 1999; Robertson et al., 2000; St. George, Kutas, Martinez, & Sereno, 1999; Tracy et al., 2003; see also Berthoz, Armony, Blair, & Dolan, 2002). It is beyond the scope of this chapter to review the results of this emerging research area (see Gernsbacher & Kashak, 2003, for a brief review). However, both fMRI and PET examine localized changes in neuronal activity in the brain via the associated changes in regional blood flow. Because such hemodynamic changes are intrinsically rather slow (in the order of seconds), the temporal resolution of measures that depend on these changes is intrinsically limited, perhaps not to the order of seconds but at least to the order of several hundreds of milliseconds (with event-related fMRI; Bandettini, Birn, & Donahue, 2000; Buckner & Logan, 2002). It is for this reason that PET and fMRI are usually referred to as having rather poor temporal resolution. To the extent that the configuration of critical neuronal systems engaged by some unfolding discourse varies more rapidly than a scanner can keep track of—an empirical issue—it will not be easy to relate the functional imaging results to the way comprehension unfolds in time.

The ERP data reviewed in this chapter demonstrate that at least some aspects of discourse-level comprehension occur extremely rapidly. To keep track of these with sufficient temporal resolution, methods that more immediately reflect neuronal activity are to be preferred (cf. Osterhout et al., chapter 14, this volume). The latter include not only measures based on EEG (ERPs and event-related oscillatory changes) but also their magnetic counterparts in magneto-encephalography (MEG). The MEG equivalent of an ERP, a so-called event-related magnetic field or ERF, can support similar types of inferences and can, when combined with ERP data, support somewhat more precise inferences about underlying neuronal generators. Another potentially relevant measure is the so-called event-related optical signal, which reflects changes in the optical scattering properties of brain tissue that are concurrent with neuronal activity (EROS) (Gratton & Fabiani, 2001). The EROS technique is still under development, but if the claims for good temporal *and* spatial resolution turn out to be correct, its benefits will be clear.

All the above-mentioned neuroimaging measures share one very important limitation, which is that regardless of all their sophistication, they can only tell us which neuronal events and systems *correlate* with, say, a discourse-dependent comprehension problem. Of course, what we would really like to know is which neuronal events and systems are *responsible for* (i.e., causally involved in) particular aspects of

discourse-level comprehension. Until recently, the only way to learn anything about the latter was through lesions caused by, say, a bullet, a cerebral vascular accident, or through the electrical stimulation of pieces of cortex exposed during brain surgery. With the development of transcranial magnetic stimulation or TMS, however, we now have a tool that allows us to noninvasively explore which brain systems are critically involved in some aspect of cognition, as well as when such involvement occurs. TMS can be seen as the inverse of MEG in that rather than picking up the magnetic field caused by neuroelectrical events, external magnetic fields are used to temporarily and harmlessly "mess up" neuronal communication at fairly limited regions of the cortex. Psycholinguists have recently begun working with the technique (e.g., Shapiro, Pascual Leone, Mottaghy, Gangitano, & Caramazza, 2001; Knecht et al., 2002), and TMS is ready to be used to explore the causal structure of discourse-level comprehension.

13.5.3 A Gauntlet Picked Up (Briefly): Why Not Just Do Self-Paced Reading Anyway?

In his thought-provoking contribution to this volume (chapter 2), Don Mitchell observes that many sentence-processing phenomena were first established in the self-paced reading task, subsequently confirmed by eyetracking experiments, and then occasionally corroborated once more in ERP studies. This convergence of findings across on-line measurement tools is reassuring, for each then introduces some unnatural features into the language comprehension situation (button pressing, bite bars, electrode caps, etc.). As Mitchell notes, however, such convergence also does raise the question of why we should spend our valuable resources on complex time-consuming measures such as EEG, MEG, and fMRI, instead of running simple self-paced reading studies all the time. In Mitchell's words, "Why use a 3G video-enabled cell-phone when a Post-It Note will do?"

To me, the most obvious reason is that although several decades of self-paced reading research has uncovered lots of phenomena, it has by no means resolved all the debates that came along with them. Like perhaps all other cognitive systems (cf. Newell, 1973; Anderson, 1978), the language comprehension system is a difficult one to pin down by behavioral data alone. The obvious response to this is to look for additional constraints, by linking domain-specific theories to a theory of generic cognitive system architecture (e.g., Newell, 1990; Rumelhart, McClelland, & the PDP Research Group, 1986), by more careful analyses of what the system is supposed to do (e.g., Marr, 1982; Anderson, 1991; Tooby & Cosmides, 2000), and—most relevant here—by gathering other types of data, such as ERPs and fMRI scans. As for the latter, the growing suspicion that the brain is perhaps not an arbitrarily structured device running algorithms that could not care less about their implementation suggests that neuroimaging data may actually provide some rather useful constraints.[14] This does not mean that we should all run for the scanners. At the same time, however, perhaps we should not all go for self-paced reading either. Post-It Notes have become indispensable, and for good reason. After a while, however, they invariably clutter up one's desk and begin to point in different directions.

13.5.4 The Real Challenges

I hope to have shown that at least one neuroimaging technique, the registration of event-related brain potentials, can help keep track of discourse-level comprehension in a way that justifies the effort. Although it takes some creativity to make it work, ERPs and discourse can be combined in a single feasible experiment to explore the existence, precise timing, identity, degree of involvement, and—to some extent—neural substrate of processes involved in discourse-level comprehension. As illustrated in Figure 13.8, ERPs are particularly good at selectively keeping track of specific aspects of comprehension, doing so with high temporal precision as people read or listen for comprehension. Because discourse-level comprehension is no simple task, being able to tell one process apart from another, to identify which of the many processes one is looking at in a particular study, is no trivial benefit. Of course, as with all other measures, this benefit comes at a cost. It depends on the specific issue at hand whether one outweighs the other or whether some other measure should be used instead.

At least three important challenges are ahead of us. One is to see whether we can *really* take ERPs (and other neuroimaging measures) beyond the sentence. Much of the ERP research that I reviewed was framed in terms of whether discourse modulated some aspect of sentence comprehension (e.g., the parsing of a local syntactic ambiguity). However, discourse is obviously more than a contextual factor affecting sentence comprehension. For many interesting discourse-level phenomena, a minimal one- or two-sentence discourse context will not provide enough leverage, and far richer texts will be required. Moreover, for many really interesting discourse-level phenomena, we will have to move away from "textoids" (Graesser et al., 1997) altogether and enter the true arena of language use (Clark, 1996) where language comprehension is relevant to achieving some private or collective goal. In our lab, we are currently exploring new ways to record EEG in a constrained conversational setting, but it remains to be seen whether we can take ERPs this far.

The second challenge is to keep the importance of on-line measurement in perspective. With comprehension and its input unfolding over time, we need tools to keep track of things as they unfold. However, the availability of such tools should not lure us into pursuing timing questions all the time. Comprehension is about things being understood, and the system that does it is doing more than just meeting deadlines. For this reason, measures that are sensitive to the nature of the representations being constructed or to the identity of the processes involved are indispensable. Some of these measures may be more or less on-line (e.g., ERPs, head-mounted eyetracking in visual-world settings, and concurrent probe tasks), but others may be relatively or even completely off-line (e.g., fMRI, PET, as well as paper-and-pencil cloze tests, plausibility ratings, and comprehension questions). If we do not make good use of such content-sensitive measures, we are never going to understand the nature of comprehension.

The final challenge for us psycholinguists, I believe, is to try to be open-minded about other people's tools to track sentence- and discourse-level comprehension. In my opinion, the field of sentence comprehension research suffers from an unhealthy amount of "dependent variable chauvinism," with occasionally very strong biases

towards a single preferred measure (say, eye movements in reading) at the expense of some other measure (say, ERPs). The phenomenon can be seen in polite debates—or, more usually, polite silences—at conferences, in the selective quoting of other people's work, and, more violently, in manuscript and grant proposal reviews. Yes, it is very tempting and very pleasant to believe that one's own method dominates all others, and probably everybody will occasionally fall prey to this, especially if the method involves a lot of hard work that requires justification. However, in view of the rather modest progress we have made with the measures currently at our disposal, there is no reason whatsoever to proclaim any of them as The Measure (or The Invariably Most Cost-Effective Measure; cf. Mitchell, chapter 2, this volume). The problems of discovery that we face are so difficult that it makes sense to welcome *every* feasible and a priori reasonable measure, as long as it provides a new window on the language comprehension system.

13.6 ACKNOWLEDGMENTS

Supported by an NWO Innovation Impulse Vidi grant. I would like to thank Colin Brown and Peter Hagoort for introducing me to the art of sentential ERP research, the research assistants of the Max Planck Institute for Psycholinguistics, the F.C. Donders Centre for Cognitive Neuroimaging, and the University of Amsterdam Psychology Department for helping me out with some very laborious experiments, the AMLaP-2002 organizers for their invitation to give a talk on ERPs and discourse (upon which this chapter is based), and Manuel Carreiras, Chuck Clifton, and Peter Hagoort for comments on an earlier version of this chapter.

References

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review, 85*(4), 249–277.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences, 14,* 471–517.

Bandettini, P. A., Birn, R. M., & Donahue, K. M. (2000). Functional fMRI. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (pp. 978–1014). Cambridge, UK: Cambridge University Press.

Bastiaansen, M. C. M., & Hagoort, P. (2003). Event-induced theta responses as a window on the dynamics of memory./*Cortex*, 39, 967–992.

Bastiaansen, M. C. M., Van Berkum, J. J. A., & Hagoort, P. (2002a). Event-related theta power increases in the human EEG during online sentence processing. *Neuroscience Letters, 323,* 13–16.

Bastiaansen, M. C. M., Van Berkum, J. J. A., & Hagoort, P. (2002b). Syntactic processing modulates the theta rhythm of the human EEG. *NeuroImage, 17,* 1479–1492.

Berthoz, S., Armony, J. L., Blair, R. J. R., & Dolan, R. J. (2002). An fMRI study of intentional and unintentional (embarrassing) violations of social norms. *Brain, 125*(8), 1696–1708.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 717–726.

Brown, C. M., & Hagoort, P. (2000). On the electrophysiology of language comprehension: Implications for the human language system. In M. W. Crocker, M. Pickering, & C. Clifton, Jr. (Eds.), *Architectures and mechanisms for language processing* (pp. 213–237). Cambridge, UK: Cambridge University Press.

Brown, C. M., Hagoort, P., & Kutas, M. (2000). Postlexical integration processes in language comprehension: Evidence from brain-imaging research. In M. S. Gazzaniga (Ed.), *The new cognitive neurosciences* (pp. 881–895). Cambridge, MA: MIT Press.

Brown, C. M., Hagoort, P., & Ter Keurs, M. (1999). Electrophysiological signatures of visual lexical processing: Open- and closed-class words. *Journal of Cognitive Neuroscience, 11*(3), 261–281.

Brown, C. M., Van Berkum, J. J. A., & Hagoort, P. (2000). Discourse before gender: An event-related brain potential study on the interplay of semantic and syntactic information during spoken language understanding. *Journal of Psycholinguistic Research, 29*(1), 53–68.

Brysbaert, M., & Mitchell, D. C. (2000). The failure to use gender information in parsing: A comment on Van Berkum, Brown, and Hagoort (1999). *Journal of Psycholinguistic Research, 29,* 453–466.

Buckner, R. L., & Logan, J. M. (2001). Functional neuroimaging methods: PET and fMRI. In R. Cabeza & A. Kingstone (Eds.), *Handbook of functional neuroimaging of cognition* (pp. 27–48). Cambridge, MA: MIT Press.

Clark, H. H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.

Clark, H. H. (1997). Dogmas of understanding. *Discourse Processes, 23*(3), 567–598.

Connolly, J. F., & Phillips, N. A. (1994). Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. *Journal of Cognitive Neuroscience, 6,* 256–266.

Dooling, D. J., & Lachman, R. (1971). Effects of comprehension on retention of prose. *Journal of Experimental Psychology, 88*(2), 216–222.

Fabiani, M., Gratton, G., & Coles, M. G. H. (2000). Event-related brain potentials. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (pp. 53–84). Cambridge, UK: Cambridge University Press.

Federmeier, K. D., & Kutas, M. (1999a). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language, 41,* 469–495.

Federmeier, K. D., & Kutas, M. (1999b). Right words and left words: Electrophysiological evidence for hemispheric differences in meaning processing. *Cognitive Brain Research, 8,* 373–392.

Federmeier, K. D., McLennan, D. B., De Ochoa, E., & Kutas, M. K. (2002). The impact of semantic memory organization and sentence context information on spoken language processing by younger and older adults: An ERP study. *Psychophysiology, 39*(2), 133–146.

Ferstl, E. C., & Von Cramon, D. Y. (2001). The role of coherence and cohesion in text comprehension: An event-related fMRI study. *Cognitive Brain Research, 11*(3), 325–340.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Friederici, A. D. (1998). The neurobiology of language comprehension. In A. D. Friederici (Ed.), *Language comprehension: A biological perspective* (pp. 263–301). Berlin: Springer.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences, 6*(2), 78–84.

Friederici, A. D., Hahne, A., & Mecklinger, A. (1996). Temporal structure of syntactic parsing: Early and late event-related brain potential effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1219–1248.

Gallagher, H. L., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C.D. (2000). Reading the mind in cartoons and stories: An fMRI study of 'theory of the mind' in verbal and nonverbal tasks. *Neuropsychologia, 38*(1), 11–21.

Gaskell, M. G., & Marslen-Wilson, W. D. (2001). Lexical ambiguity resolution and spoken word recognition: Bridging the gap. *Journal of Memory and Language, 44,* 325–349.

Gernsbacher, M. A., & Kashak, M. P. (2003). Neuroimaging studies of language production and comprehension. *Annual Review of Psychology, 54,* 91–114.

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology, 48,* 163–189.

Gratton, G., & Fabiani, M. (2001). Shedding light on brain function: The event-related optical signal. *Trends in Cognitive Sciences, 5*(8), 357–363.

Hagoort, P., & Brown, C. M. (2000a). ERP effects of listening to speech: Semantic ERP effects. *Neuropsychologia, 38,* 1518–1530.

Hagoort, P., & Brown, C. M. (2000b). ERP effects of listening to speech compared to reading: The P600/SPS to syntactic violations in spoken sentences and rapid serial visual presentation. *Neuropsychologia, 38,* 1531–1549.

Hagoort, P., Brown, C. M., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes, 8,* 439–483.

Hagoort, P., Brown, C. M., & Osterhout, L. (1999). The neurocognition of syntactic processing. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 273–316). Oxford: Oxford University Press.

Hess, D. J., Foss, D. J., & Carroll, P. (1995). Effects of global and local context on lexical processing during language comprehension. *Journal of Experimental Psychology: General, 124*(1), 62–82.

Jongman, A. (1998). Effects of vowel length and syllabic structure on segment duration in Dutch. *Journal of Phonetics, 26,* 207–222.

Kemps, R. J. J. K., Ernestus, M., Schreuder, R., & Baayen, R. H. (2003). *Prosodic cues for morphological complexity*: The case of Dutch plurals. Manuscript submitted for publication.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Kluender, R., & Kutas, M. (1993). Bridging the gap: Evidence from ERPs on the processing of unbounded dependencies. *Journal of Cognitive Neuroscience, 5*(2), 196–214.

Knecht, S., Floël, A., Draeger, B., Breitenstein, C., Sommer, J., Henningsen, H., Ringelstein, E. B., & Pascual Leone, A. (2002). Degree of language lateralization determines susceptibility to unilateral brain lesions. *Nature Neuroscience, 5*(7), 695–699.

Kutas, M. (1997). Views on how the electrical activity that the brain generates reflects the functions of different language structures. *Psychophysiology, 34,* 383–398.

Kutas, M., & Federmeier, K. D. (1998). Minding the body. *Psychophysiology, 35*(2), 135–150.

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. Trends in Cognitive Sciences, 12, 463–470.

Kutas, M., Federmeier, K. D., Coulson, S., King, J. W., & Münte, T. F. (2000). Language. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (pp. 576–601). Cambridge, UK: Cambridge University Press.

Kutas, M., Federmeier, K. D., & Sereno, M. I. (1999). Current approaches to mapping language in electromagnetic space. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 359–387). Oxford: Oxford University Press.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207,* 203–205.

Kutas, M., & Hillyard, S. A. (1983). Event-related brain potentials to grammatical errors and semantic anomalies. *Memory and Cognition, 11,* 539–550.

Kutas, M., & Schmitt, B. M. (2003). Language in microvolts. In M. T. Banich & M. Mack. (Eds.), *Mind, brain, and language*. Hillsdale, NJ: Erlbaum.

Kutas, M., & Van Petten, C. K. (1994). Psycholinguistics electrified: Event-related brain potential investigations. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 83–143). New York: Academic Press.

Maguire, E. A., Frith, C. D., & Morris, R. G. M. (1999). The functional neuroanatomy of comprehension and memory: The importance of prior knowledge. *Brain, 122*(10), 1839–1850.

Makeig, S., Westerfield, M., Jung, T.-P., Enghoff, S., Townsend, J., Courchesne, E., & Sejnowski, T.J. (2002). Dynamic brain sources of visual evoked responses. *Science, 295,* 690–694.

Marr, D. (1982). *Vision*. Freeman.

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. Cognition, 8, 1–71.

Mueller, H. M., King, J. W., & Kutas, M. (1997). Event-related potentials to relative clause processing in spoken sentences. Cognitive Brain Research, 5, 193–203.

Münte, T. F., Schiltz, K., & Kutas, M. (1998). When temporal terms belie conceptual order. *Nature, 395,* 71–74.

Myers, J. L., & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes, 26,* 131–157.

Neville, H., Nicol, J. L., Barss, A., Forster, K. I., & Garrett, M. F. (1991). Syntactically based sentence processing classes: Evidence from event-related brain potentials. *Journal of Cognitive Neuroscience, 3*(2), 151–165.

Newell, A. (1973). You can't play 20 questions with nature and win. In W. G. Chase (Ed.), *Visual information processing* (pp. 283–308). New York: Academic Press.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Osterhout, L., Bersick, M., & McKinnon, R. (1997). Brain potentials elicited by words: Word length and frequency predict the latency of an early negativity. *Biological Psychology, 46*(2), 143–168.

Osterhout, L., & Hagoort, P. (1999). A superficial resemblance doesn't necessarily mean you're part of the family: Counterarguments to Coulson, King, and Kutas (1998) in the P600/SPS debate. *Language and Cognitive Processes, 14*(1), 1–14.

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

Osterhout, L., & Holcomb, P. J. (1993). Event-related potentials and syntactic anomaly: Evidence of anomaly detection during the perception of continuous speech. *Language and Cognitive Processes, 8*(4), 413–437.

Osterhout, L., & Holcomb, P. J. (1995). Event-related potentials and language comprehension. In M. D. Rugg & M. G. H. Coles (Eds.), *Electrophysiology of mind* (pp. 171–215). Oxford: Oxford University Press.

Osterhout, L., McLaughlin, J., & Bersick, M. (1997). Event-related brain potentials and human language. *Trends in Cognitive Sciences, 1*(6), 203–209.

Osterhout, L., & Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language, 34,* 739–773.

Osterhout, L., & Nicol, J. (1999). On the distinctiveness, independence, and time course of the brain response to syntactic and semantic anomalies. *Journal of Psycholinguistic Research, 14*(3), 283–317.

Pfurtscheller, G., & Lopes da Silva, F. H. (1999). Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology, 110,* 1842–1857.

Pulvermüller, F. (2001). Brain Reflections of words and their meanings. *Trends in Cognitive Sciences*, 5, 517–524.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124,* 372–422.

Robertson, D. A., Gernsbacher, M. A., Guidotti, S. J., Robertson, R. R. W., Irwin, W., Mock, B. J., & Campana, M. E. (2000). Functional neuroanatomy of the cognitive process of mapping during discourse comprehension. *Psychological Science, 11*(3), 255–260.

Rugg, M. D., & Coles, M. G. H. (1995). The ERP and cognitive psychology: Conceptual issues. In M. D. Rugg & M. G. H. Coles (Eds.), *Electrophysiology of mind: Event related brain potentials and cognition*. (Vol. 25, pp. 27–39). Oxford: Oxford University Press.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). *Parallel distributed processing*. Cambridge, MA: MIT Press.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension, *Cognition*, 90, 51–89.

Shapiro, K. A., Pascual Leone, A., Mottaghy, F. M., Gangitano, M., & Caramazza, A. (2001). Grammatical distinctions in the left frontal cortex. *Journal of Cognitive Neuroscience, 13*(6), 713–720.

Sitnikova, T., Kuperberg, G. & Holcomb, P. J. (2003). Semantic integration in videos of real-world events: An electrophysiological investigation. *Psychophysiology, 40,* 160–164.

St. George, M., Kutas, M., Martinez, A., & Sereno, M. I. (1999). Semantic integration in reading: Engagement of the right hemisphere during discourse processing. *Brain, 122*(7), 1317–1325.

St. George, M., Mannes, S., & Hoffman, J. E. (1994). Global semantic expectancy and language comprehension. *Journal of Cognitive Neuroscience, 6*(1), 70–83.

St. George, M., Mannes, S., & Hoffman, J. E. (1997). Individual differences in inference generation: An ERP analysis. *Journal of Cognitive Neuroscience, 9*(6), 776–787.

Streb, J., Rösler, F., & Hennighausen, E. (1999). Event-related responses to pronoun and proper name anaphors in parallel and nonparallel discourse structures. *Brain and Language, 70*(2), 273–286.

Tallon-Baudry, C., & Bertrand, O. (1999). Oscillatory gamma activity in humans and its role in object representation. *Trends in Cognitive Sciences, 3*(4), 151–162.

Tooby, J., & Cosmides, L. (2000). Toward mapping the evolved functional organization of mind and brain. In M. S. Gazzaniga (Ed.), *The new cognitive neurosciences* (pp. 1167–1177). Cambridge, MA: MIT Press.

Tracy, J., Flanders, A., Madi, S., Natale, P., Delvecchio, N., Pyrros, A., & Laskas, J. (2003). The brain's response to incidental intruded words during focal text processing. *NeuroImage,* 18, 117–126.

Urbach, T. P., & Kutas, M. (2002). The intractability of scaling scalp distributions to infer neuroelectric sources. *Psychophysiology, 39*(6), 791–808.

Van Berkum, J. J. A., Brown, C. M., & Hagoort, P. (1999a). Early referential context effects in sentence processing: Evidence from event-related brain potentials. *Journal of Memory and Language, 41,* 147–182.

Van Berkum, J. J. A., Brown, C. M., & Hagoort, P. (1999b). When does gender constrain parsing? Evidence from ERPs. *Journal of Psycholinguistic Research, 28*(5), 555–571.

Van Berkum, J. J. A., Brown, C. M., Hagoort, P., & Zwitserlood, P. (2003A). Event-related brain potentials reflect discourse-referential ambiguity in spoken-language comprehension. *Psychophysiology, 40,* 235–248.

Van Berkum, J. J. A., Brown, C. M., Hagoort, P., Zwitserlood, P., & Kooijman, V., (2002a). Can people use discourse-level information to predict upcoming words in an unfolding sentence? Evidence from ERPs and self-paced reading. *Proceedings of the 8th International Conference of Cognitive Neuroscience (ICON-8),* Porquerolles, France, September 9-15, p. 105.

Van Berkum, J. J. A., Brown, C. M., Hagoort, P., Zwitserlood, P., & Kooijman, V., (2004). Can people use discourse-level information to predict upcoming words in an unfolding sentence? Evidence from ERPs and self-paced reading. Manuscript submitted for publication.

Van Berkum, J. J. A., Hagoort, P., & Brown, C. M. (1999c). Semantic integration in sentences and discourse: Evidence from the N400. *Journal of Cognitive Neuroscience, 11*(6), 657–671.

Van Berkum, J. J. A., Hagoort, P., & Brown, C. M. (2000). The use of referential context and grammatical gender in parsing: A reply to Brysbaert and Mitchell. *Journal of Psycholinguistic Research, 29*(5), 467–481.

Van Berkum, J. J. A., Kooijman, V., Brown, C. M., Zwitserlood, P., & Hagoort, P. (2002b). Do listeners use discourse-level information to predict upcoming words in an unfolding sentence? An ERP study. *Proceedings of the Cognitive Neuroscience Society 9th annual meeting (CNS-2002),* San Francisco, April 14–16. Supplement to *Journal of Cognitive Neuroscience* (p. 82).

Van Berkum, J. J. A., Zwitserlood, P., Brown, C. M., & Hagoort, P. (2003b). When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. *Cognitive Brain Research,* 17, 701–718.

Van den Brink, D., Brown, C. M., & Hagoort, P. (2001). Electrophysiological evidence for early contextual influences during spoken-word recognition: N200 versus N400 Effects. *Journal of Cognitive Neuroscience, 13*(7), 967–985.

Van Petten, C., Kutas, M., Kluender, R., Mitchiner, M., & Mclsaac (1991). Fractionating the word repetition effect with event-related potentials. *Journal of Cognitive Neuroscience, 3*(2), 131–150.

Weiss, S., & Mueller, H. M. (2003). The contribution of EEG coherence to the investigation of language. *Brain and Language, 85,* 325–343.

West, W. C., & Holcomb, P. J. (2002). Event-related potentials during discourse-level semantic integration of complex pictures. *Cognitive Brain Research, 13*(3), 363–375.

Notes

1. In engineering, the temporally extended response of a system to (something approximating) an infinitely short impulse at its input is termed an *impulse response*, and is considered to be one of the keys to characterizing and analyzing basic system behavior.

2. The hypothetical four-stage compartmentalization of comprehension displayed is chosen to make a methodological point, not to advance a particular model of language processing.
3. In spoken language, sentence-level prosodic cues such as those signaling a pause, a question, or a bit of irony, also drive the system. To the extent that these cues (some of which have a written-language counterpart in punctuation) cannot be tied to specific words, the image of sentence comprehension as a sequence of solely *word-driven* incremental impulse responses is incomplete.
4. Although we know that discourse-dependent ERP effects can have a surprisingly early onset (see the section on timing inferences), the effect displayed in Figure 13.2 begins to emerge somewhere in the first 50 ms from estimated onset of the inflection. We suspect that this rather disconcerting timing may have to do with how we estimated the acoustic onset of the inflection. To determine the latter, we displayed the speech waves of the two alternative adjectives (e.g., *groot* and *grote*) simultaneously on screen, and used this as well as their sound to determine at what point in time the two acoustic signals started to diverge in sound (which, for this example, would be by the end of the "t"). However, this procedure did not take into account such subtle cues as durational changes in the stem vowel (which are known to systematically correlate with the presence of an inflection; e.g., Jongman, 1998), for the simple reason that such cues are very difficult to capture in a practical measurement procedure. Because recent evidence suggests that listeners *can* make use of such early cues to an upcoming inflection in the stem of the word (Salverda, Dahan, & McQueen, 2003; Kemps, Ernestus, Schreuder, & Baayen, 2003; see also Gaskell & Marslen-Wilson, 2001), the estimated inflection onsets that we used for ERP averaging may well have been "too late" to an unknown variable extent (see Van Berkum et al., 2003, for further discussion).
5. In fact, when we presented a subset of the same materials in a noncumulative, moving-window, self-paced reading paradigm (Van Berkum et al., 2002a), readers did *not* slow down at the critical adjective if it had an expectation-incongruent gender inflection. They did slow down several words downstream at a second—and equally incongruently inflected—adjective. Because the latter effect still occurred *before* readers saw the noun, it supports our claim that discourse-level information is routinely used to predict specific upcoming words. At the same time, however, it illustrates that not everything seen in ERPs can also easily be picked up in self-paced reading (cf. Mitchell, chapter 2, this volume).
6. The study just discussed actually provides an interesting illustration of this problem. Because we were looking for an ERP effect tied to adjective *inflections*, the most appropriate ERP analysis was one in which we computed ERPs relative to the acoustic onset of the inflection within the adjective. However, for the same data we also computed ERPs relative to the acoustic onset of the adjective itself. Our adjectives were of variable length, and the critical inflectional suffix was therefore at a variable distance from adjective onset. The ERPs time-locked to adjective onset did reveal some traces of perturbation by incorrectly inflected adjectives. However, consistent with the fact that the (variably later) *inflection* provided the truly critical information, the ERP effect relative to adjective onset was broader, much less articulate, and not statistically significant.
7. The inflection-tied lexical prediction effect discussed in the previous section (Figure 13.2) illustrates the time-locking difficulties posed by spoken-language materials very well, both with respect to latency jitter (see note 6) and with respect to potential bias in one's estimate (see note 4). However, the problem is not specifically tied to ERPs,

and holds for any measure (e.g., eyetracking and probe response times) that must be related to an unfolding acoustic signal.

8. Note that for this logic to work, it is *not* important to know whether the P600 specifically indexes the initial detection of a syntactic problem, its further diagnosis, or subsequent repair. All one needs to assume is that, within the domain of language comprehension, a syntactic dead end reliably elicits a P600 effect (see Van Berkum et al., 2000).
9. To argue that the observed difference in N400 effect size does *not* reflect some aspect of how words like *pines* and *tulips* related to the dynamic discourse context, Kutas and Federmeier need to assume that the equal plausibility and predictability ratings obtained for these words rule out any other N400-relevant difference in how these words relate to the discourse. This is a very strong assumption, which need not be correct. In addition, there seems to be something odd about separating the knowledge encoded in one's representation of a discourse context from the structure of one's semantic memory. The former necessarily always draws upon, and thus incorporates, part of the latter. Moreover, although semantic memory can be said to have a rich *permanent* structure, what is the *relevant* structure may well be *dynamically* co-determined by the discourse context. That is, although a species-based classification will be relevant in many situations, one can easily imagine a context in which *palms* cluster with *tulips* instead of with *pines* (e.g., both *palms* and *tulips* will not be found up high in the mountains).
10. Although the problem is formulated for ERP effect size issues, it also occurs with fine-grained parametric studies of, say, the onset or peak latency of an ERP effect.
11. Because ERPs can be recorded in the presence or absence of an artificial secondary response task, they provide a unique opportunity to study the impact of such tasks. As might be expected, a task in which the subject is asked to evaluate linguistic stimuli on a critical feature (e.g., grammaticality judgments in an experiment looking for P600 effects to grammatical anomalies) can change the outcome of the experiment (see Osterhout & Mobley, 1995, for just such an example). Such task-dependence is not necessarily problematic or uninteresting. However, it does reveal that secondary response tasks can affect the processes involved in a language comprehension experiment (cf. Mitchell, chapter 2, this volume).
12. One might be tempted to argue that the self-paced reading task does not do so well in this respect because the reading time effects in this task frequently show up one or two words downstream from the critical word. However, although this spillover phenomenon is most salient in the self-paced reading task, the downstream displacement of effects is a general problem that affects *all* on-line measures. If the processing consequences to word X_w extend into the temporal domain of subsequent word X_{w+1} or beyond (i.e., the incremental impulse responses overlap), then both the ERP signals timelocked to, and the eye fixation on, the latter word may be contaminated by overlapping effects to the former.
13. In my experience, most subjects do not have a hard time listening to stories over four to five blocks of 15 min each, provided that the stories are sufficiently interesting, and provided that a few other conditions are met. One is that subjects should be given substantial breaks between recording blocks with each break offering some distraction (conversation, coffee, a check of electrode impedances, etc.). Another is to avoid a sleep-inducing environment, such as by *not* dimming the lights to dusk level, by *not* trying to block out all extraneous environmental sounds, and by mixing input modalities (e.g., announcing each visually displayed new discourse by means of a beep). Under these conditions, a 75-min ERP recording session is by no means necessarily

more problematic or arduous than a 15-min self-paced reading experiment (cf. Mitchell, chapter 2, this volume). Related to this, and probably partly due to social-psychological factors, subjects participating in an EEG experiment are very often more committed than their counterparts taking part in a run-of-the-mill 15-min RT task.

14. In discussing the relevance of neuroimaging data to psycholinguistics, Mitchell (chapter 2, this volume) could be taken to suggest that such data provide spatial XYZ-coordinates only and, as such, do not bear on traditional (spatially agnostic) theories of language processing. However, knowing which particular area of the brain is engaged by a particular task or condition (usually more than one) is not merely a matter of spatial localization. We often already know about *other* tasks or conditions that engage the exact same area and about how cognition breaks down if the area at hand is damaged. In addition, we sometimes also know from which other brain systems the area at hand receives input and to what other systems it projects its output. All of this can provide strong additional constraints on traditional theories of language comprehension, even if the latter do not speak about the brain at all.