Cognition 215 (2021) 104814

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

# Unexpected words or unexpected languages? Two ERP effects of code-switching in naturalistic discourse

Anthony Yacovone *, Emily Moya, Jesse Snedeker

*Harvard University, USA*

ARTICLE INFO

*Keywords:*
Bilingualism
EEG/ERP
Form-based prediction
N400
LPC
Discourse

ABSTRACT

Bilingual speakers often switch between languages in conversation without any advance notice. Psycholinguistic research has found that these language shifts (or *code-switches*) can be costly for comprehenders in certain situations. The present study explores the nature of these costs by comparing code-switches to other types of unexpected linguistic material. To do this, we used a novel EEG paradigm, the *Storytime* task, in which we record readings of natural texts, and then experimentally manipulate their properties by splicing in words. In this study, we manipulated the language of our target words (English, Spanish) and their fit with the preceding context (strong-fit, weak-fit). If code-switching incurs a unique cost beyond that incurred by an unexpected word, then we should see an additive pattern in our ERP indices. If an effect is driven by lexical expectation alone, then there should be a non-additive interaction such that all unexpected forms incur a similar cost. We found three effects: a general prediction effect (a non-additive N400), a post-lexical recognition of the switch in languages (an LPC for code-switched words), and a prolonged integration difficulty associated with weak-fitting words regardless of language (a sustained negativity). We interpret these findings as suggesting that the processing difficulties experienced by bilinguals can largely be understood within more general frameworks for understanding language comprehension. Our findings are consistent with the broader literature demonstrating that bilinguals do not have two wholly separate language systems but rather a single language system capable of using two coding systems.

## 1. Introduction

When bilinguals speak to one another, they often shift between their languages, producing utterances like “Can you get me *un café con leche y azúcar* [a coffee with milk and sugar]?” This flexible use of languages, or *code-switching*, occurs frequently in natural discourse (Poplack, 1980; Sebba, Mahootian, & Jonsson, 2012) and serves a variety of functions (Auer, 1988; Gumperz, 1982; Heller, 2007). Bilingual speakers are known to code-switch individual words, entire sentences, and even large portions of their conversations (e.g. Grosjean, 2001; Heredia & Altarriba, 2001; Milroy & Gordon, 2008). This process of integrating multiple languages on-the-fly can appear seamless, and often results in very few errors or overt breakdowns in communication (Poplack, 1980). But these switches may not be truly effortless: many behavioral and neurocognitive studies find that comprehenders incur costs associated with switching languages such as longer reaction times in lexical decision tasks and increases in the neural response to code-switched words relative to non-switched words (see Van Hell, Litcofsky, & Ting, 2015).

It is tempting to interpret these effects as evidence that comprehenders must actively switch from one language to another and that this process takes additional time and effort. But similar data patterns emerge in studies that do not involve code-switching. For example, hearing low frequency words or improbable sentence continuations will elicit similar effects in monolingual contexts (e.g. Forster & Chambers, 1973; Kutas & Federmeier, 2011). This raises an intriguing alternative hypothesis: perhaps the costs found in code-switching studies are caused by encountering unexpected input rather than a discrete process triggered by switching languages. The present study uses a novel paradigm combining electroencephalography (EEG) and naturalistic listening to explore this question.

In the remainder of this Introduction, we evaluate the evidence showing the costs of code-switching and their variability (Section 1.1) and describe two alternative theories about those costs (Section 1.2). We then outline the predictions that these theories make regarding two ERP responses: the N400 and the *Late Positive Complex* or LPC (Section 1.3). Finally, we end by describing the goals of the present study and the

* Corresponding author at: Department of Psychology, Harvard University, William James Hall, 33 Kirkland St., Cambridge, MA 02138, USA.
*E-mail address:* anthony_yacovone@g.harvard.edu (A. Yacovone).

https://doi.org/10.1016/j.cognition.2021.104814
Received 29 September 2019; Received in revised form 10 June 2021; Accepted 13 June 2021
Available online 21 July 2021

benefits of our novel paradigm, the *Storytime* task (Section 1.4).

### 1.1. The cost of code-switching in bilingual comprehension: evidence from ERPs

Over the past few decades, many researchers have studied the effects of code-switching on comprehension using EEG and *event-related potentials* (ERPs). ERPs are averaged electrical responses collected at the scalp and time-locked to the onset of a stimulus. ERPs vary systematically in their amplitudes, latencies, and/or scalp distributions, making them useful for characterizing when and to what degree different variables affect cognitive processes like language comprehension (see Kappenman & Luck, 2012; Luck, 2005, 2014). The ERP literature on code-switching largely reports a biphasic response to code-switched words in sentence contexts: an early negativity (e.g. an N400) followed by a *Late Positive Complex* or LPC (see Van Hell et al., 2015; Van Hell, Fernandez, Kootstra, Litcofsky, & Ting, 2018; Fernandez, Litcofsky, & Van Hell, 2019 for discussion). However, there are a few studies that do not find this biphasic pattern—for example, studies in which participants have lower levels of proficiency or experience with the matrix language often find no early negativities (e.g. Ruigendijk, Hentschel, & Zeller, 2016; see Zeller, 2020 for discussion). We return to these studies in the General Discussion. But for now, we will focus on the studies that *do* show the biphasic pattern during sentence comprehension and briefly describe the variability of these ERP effects in the code-switching literature.

#### 1.1.1. Switch-related negativities and their variability in the literature

Most ERP studies on code-switching find some kind of early negativity followed by a late positivity in response to code-switched words. We will refer to these effects as being *switch-related* merely because they occur in response to code-switched words. In using this term, we do not mean to imply that these switch-related effects reflect a process of language selection or language switching—in fact, we will be arguing that one of these effects (i.e. the earlier negativity) reflects the effects of lexical prediction rather than language switching per se.

In the code-switching literature, the most common switch-related negativity is the N400. The N400 response is a negative-going deflection in an ERP waveform that typically peaks around 400 ms post-stimulus onset. This response is argued to reflect how easily a word is accessed and/or integrated into its context—the larger the N400 amplitude, the greater the processing difficulty (Kutas & Federmeier, 2009, 2011; Van Petten, 1993). Switch-related N400 effects have been taken as evidence that switching between languages is costly and disrupts lexical processing (Alvarez, Holcomb, & Grainger, 2003; Grainger & Holcomb, 2009; Van Hell et al., 2018, 2015).

Some of the earliest studies to find N400 effects used the aptly-named *switch-task* paradigm in which a series of individual words are presented back-to-back and categorized by bilingual participants. In these tasks, "code-switching" occurs when a participant sees a word in one language and then a word in another language on the next trial. These studies find that switching languages between trials elicits increased N400 responses relative to non-switched trials (e.g. Alvarez et al., 2003; Midgley, Holcomb, & Grainger, 2009). In recent EEG studies, researchers have relied on more naturalistic paradigms to study code-switching. For example, some studies use single-sentence contexts or written discourses with intra-sentential code-switching, which is when one or more words are switched within a single utterance. The findings from these sentence comprehension studies largely support those from the earlier switch-tasks: there is an increased N400 response to code-switched words relative to the non-switched words (see Fernandez et al., 2019; Van Hell et al., 2018).

One tricky aspect of the ERP literature on code-switching is that not *every* study finds an early negativity with the latency and scalp distribution of a canonical N400. For example, in a foundational study by Moreno, Federmeier, and Kutas (2002), English-Spanish bilinguals read a mix of regular and idiomatic sentences (i.e. well-known proverbs). Their critical manipulation always occurred on the sentence-final word, which was either the expected, within-language word, its translation equivalent, or an unexpected, within-language word (see examples below).

(1) a. **Idiomatic sentences:** "Out of sight, out of…mind/brain/*mente* (mind)."
b. **Regular sentences:** "Each night the campers build a…fire/blaze/*fuego* (fire)."

In idiomatic sentences (1a), they found an LPC but no switch-related negativity—possibly because of the predictability of their idioms. We will return to this finding in the General Discussion, but for now, we focus on the results for the regular sentences (1b). In these more standard sentences, the authors found a left-lateralized negativity between 250 and 450 ms and an LPC between 450 and 850 ms in response to code-switched words (e.g. *fuego*). This early negativity was equivalent in magnitude to the canonical N400 elicited by their unexpected, within-language words (e.g. blaze). However, the left-frontal skew of this effect led the authors to interpret it as a *Left Anterior Negativity* (LAN) instead. Historically, LANs have been associated with increased demands on working memory (King & Kutas, 1995; Kluender & Kutas, 1993) and/or difficulties with morphosyntactic processing (e.g. Friederici, 2002; Gunter, Friederici, & Schriefers, 2000; Neville, Nicol, Barss, Forster, & Garrett, 1991). More recently, Ng, Gonzalez, and Wicha (2014) found a similar biphasic LAN-LPC pattern in response to code-switched words. In this study, Spanish-English bilinguals read short stories in English. Throughout the stories, some nouns and verbs were occasionally code-switched into Spanish (e.g. "The wind and the *sol* (sun) were disputing which was the stronger. Suddenly they *miraron* (saw) a traveler coming down the street…."). They report a LAN (350-450 ms) and an LPC (500-900 ms) to both code-switched nouns and verbs.

Taken together, one interpretation of these two studies is that the switch-related LAN and the switch-related N400 are functionally distinct, and thus the cognitive processes invoked in the studies that find LANs and in the studies that find N400s are systematically different (see Van Hell et al., 2018 for discussion). The burden of such an account would be to explain why the processes invoked by code-switches vary across studies and to identify replicable means of producing each distinct data pattern. In the code-switching ERP literature, there is little systematicity in the types of stimuli that elicit LAN vs. N400 effects. For example, studies using sentence-final code-switches have found LANs, N400s, and sometimes both effects overlapping with one another (for LANs, see Moreno et al., 2002; for N400 effects, see Proverbio, Leoni, & Zani, 2004; Van Der Meij, Cuetos, Carreiras, & Barber, 2011 with low proficiency bilinguals; FitzPatrick & Indefrey, 2014; Zeller, Hentschel, & Ruigendijk, 2016; Ruigendijk et al., 2016; for both, see Van Der Meij et al., 2011 with high proficiency bilinguals). In fact, when we look beyond code-switching to the broader psycholinguistic ERP literature, we see similar variation in the LAN and N400 effects elicited by unexpected and/or ungrammatical lexical items (Bornkessel-Schlesewsky & Schlesewsky, 2019; Caffarra, Mendoza, & Davidson, 2019; Fromont, Steinhauer, & Royle, 2020; Molinaro, Barber, & Carreiras, 2011; Molinaro, Barber, Caffarra, & Carreiras, 2015; Royle, Drury, & Steinhauer, 2013; Steinhauer & Drury, 2012; Tanner, 2015 for discussions of this debate). However, we will postpone discussion of this debate to the General Discussion.

We have observed three treatments of switch-related negativities in the ERP literature. Some authors treat LAN and N400 effects (and other negativities) as categorically distinct, invoking a difference in their function when interpreting their findings (e.g. Moreno et al., 2002; Ng et al., 2014; see Van Hell et al., 2018 for discussion). Some authors do not make strong predictions about which negativity they will find, conducting analyses consistent with both the N400 and the LAN (e.g.

Kaan, Kheder, Kreidler, Tomić, & Valdés Kroff, 2020). Finally, some authors note when negativities do not have the canonical distribution of N400s but nonetheless interpret these effects as reflecting the processes underlying the N400 (e.g. Van Hell & Witteman, 2009; Van Hell et al., 2015).

Adding to the complexity of the prior literature, we note that the LAN and the N400 are not the only early negativities observed in code-switching experiments. Other studies have interpreted their switch-related negativities as being *Phonological Mismatch Negativities* (PMNs, see Liao & Chan, 2016), *N1 effects* (e.g. Proverbio et al., 2004; Proverbio, Čok, & Zani, 2002), *N200 effects* (Khamis-Dakwar & Froud, 2007), *left-occipital N250 effects* (e.g. Van Der Meij et al., 2011), *fronto-central negativities* (Hut & Leminen, 2017), *"broad" negativities* (Zeller, 2020), and finally *anterior negativities* (ANs, Litcofsky & Van Hell, 2017 for second code-switched word; Zeller, 2020). All of these effects appear in addition to or in place of canonical N400 effects, varying in their precise timings (starting as early as 130 ms and lasting as late as 900 ms post-stimulus onset) and in their scalp distributions (ranging from left anterior, bilateral, fronto-central, to widespread). However, these effects also show clear commonalities: most of them take place, at least in part, during the typical N400 time window, and most show a scalp distribution that at least overlaps with the canonical N400 distribution. Furthermore, some of the variation in these effects could potentially be explained by differences in the presentation modality and the speed of language processing in a given population or during a particular task.

In the present paper, we have adopted the working hypothesis that the various switch-related negativities reflect a common underlying process (or set of processes)—and that this process is the same one that underlies the classic N400 effects. We do this both for ease of explanation and because we believe that it is the most parsimonious explanation given the existing data. On this hypothesis, the variation in latency and distribution would be attributed to the following: differences in processing speed due to features of the stimuli or the participants; differences in modality; differences in predictability; and differences in the other processes that are occurring within the same time window (see Moreno et al., 2002; Moreno, Rodríguez-Fornells, & Laine, 2008; Van Der Meij et al., 2011; Ng et al., 2014; Zeller, 2020 for similar interpretations). The challenge for such a hypothesis is to account for this variability and to make testable predictions. We return to this challenge in the General Discussion. Critically, our findings (and the validity of this experiment) do not depend on whether this working hypothesis is true. While our primary analysis will focus on the canonical N400 time window and electrode sites, we will also conduct exploratory analyses that investigate the precise distribution and timing of all of our effects.

#### 1.1.2. Switch-related LPCs and their variability in the literature

The second type of switch-related ERP components is the LPC, which typically peaks around 600 ms post-stimulus onset (over posterior electrode sites) and is argued to occur after initial lexical processing, i.e., after the N400/LAN (e.g. Fernandez et al., 2019; Litcofsky & Van Hell, 2017; Moreno et al., 2002, 2008; Ng et al., 2014; Proverbio et al., 2004; Van Hell et al., 2018, 2015; Van Hell & Witteman, 2009). These late-emerging, long-lasting positivities often occur in response to code-switched words in sentence contexts, but they can also be found in a variety of other linguistic (and non-linguistic) tasks (see Kuperberg, Brothers, & Wlotko, 2019; Van Petten & Luka, 2012). The precise interpretation of LPCs is still debated, but there seems to be agreement that they reflect the recognition of a high-level discrepancy (e.g. a language shift, a syntactic error, an unexpected event) and the reevaluation of the input to make sense of this unexpected event (see Coulson, King, & Kutas, 1998; Friederici, 2005; Hagoort, 1993; Hahne & Friederici, 1999; Kaan, Harris, Gibson, & Holcomb, 2000; Kolk & Chwilla, 2007; Kuperberg, 2007; Kuperberg et al., 2019; Litcofsky & Van Hell, 2017; Osterhout & Holcomb, 1992; Tanner, Grey, & Van Hell, 2017). On this interpretation, switch-related LPC effects would reflect the recognition of the language switch (Moreno et al., 2002) and the costs associated with integrating the new language into the discourse (Van Hell et al., 2018).

The switch-related LPC has been observed many times alongside switch-related negativities. There is, however, variation in the size of these effects (and when they occur) that seems to be related to factors like the predictability of the switch (FitzPatrick & Indefrey, 2014; Moreno, Rodríguez-Fornells, & Laine, 2008; Van Hell et al., 2018), the switching direction (Fernandez et al., 2019; Liao & Chan, 2016; Litcofsky & Van Hell, 2017), the participants' language proficiency (Alvarez et al., 2003; Moreno et al., 2002; Ruigendijk et al., 2016; Van Der Meij et al., 2011), and their experience with code-switching (e.g. Proverbio et al., 2004). We return to this variability in the General Discussion.

In sum, the ERP literature on code-switching provides strong, converging evidence that comprehenders are sensitive to an unforeseen shift in the language being used and experience some processing difficulties. The question addressed in the present study is whether these difficulties are specific to switching languages or whether they are simply an indirect consequence of processing an unexpected word.

### 1.2. Two theories about the costs of code-switching

Broadly speaking, there are two ways in which we could imagine bilinguals tackling the task of understanding multilingual utterances. First, language identification could precede lexical processing: a bilingual could initially determine which language is being spoken (perhaps on the basis of phonetic features, see Caramazza & Brones, 1979; Dijkstra, 2005; Dijkstra & Van Heuven, 2002; Grainger & Beauvillain, 1987; Grainger & Dijkstra, 1992; Macnamara & Kushnir, 1971; Soares & Grosjean, 1984; Scarborough, Gerard, & Cortese, 1984; Van Hell & Tanner, 2012). Then, they could switch into that language, and find the relevant word. On this account, code-switching costs would arise from the need to switch languages prior to accessing the code-switched word (Alvarez et al., 2003; Bultena, Dijkstra, & Van Hell, 2015; Grainger & Holcomb, 2009; Green, 1998). Second, lexical access could occur prior to (or independent of) recognizing the language of the word being processed: a bilingual could simultaneously map the sounds they hear (or the letters/signs they see) onto lexical forms in both of the languages that they know (e.g. Dijkstra, Grainger, & Van Heuven, 1999; Duyck, Van Assche, Drieghe, & Hartsuiker, 2007; Van Hell & De Groot, 1998). On this second account, language identification might only be achieved after the word is accessed and recognized as belonging to a particular lexicon (Dijkstra & Van Heuven, 2002; Moreno et al., 2002)—in fact, one could imagine a comprehension system in which the listener never *actively* recognized which language the word was in.

The evidence to date favors this second theory of bilingual comprehension. Many studies show that bilinguals simultaneously activate words from two languages as a spoken word unfolds. For example, Russian-English bilinguals hearing the sounds "*shar…*" will activate both the English word *shark* and the Russian word *sharik* (balloon), as both words match the initial phonemes that they heard (Marian & Spivey, 2003). The fact that bilinguals initially entertain both words and then arrive at the correct word after phonological disambiguation suggests that it is not necessary to distinguish between different lexicons during comprehension (e.g. Grainger & Beauvillain, 1987; Hartsuiker, Pickering, & Veltkamp, 2004; Kroll, Dussias, Bogulski, & Kroff, 2012; Li, 1996; Loebell & Bock, 2003; Marian & Spivey, 2003; Spivey & Marian, 1999). At first glance, this model of bilingualism is hard to reconcile with the studies that find costs to code-switching. If you do not need to switch from one lexicon to another before accessing a word, why would code-switched words be processed more slowly or effortfully? We see two alternative explanations for these code-switching effects, both of which come from an expectation-based framework for language comprehension (see Hale, 2001; Levy, 2008; Pickering & Gambi, 2018; Pickering & Garrod, 2013; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Venhuizen, Crocker, & Brouwer, 2019 for expectation-based frameworks).

One type of code-switching effect could result from later, post-lexical processes that occur after the listener realizes that the speaker has switched from one language to another. In EEG, we might expect these post-lexical effects to be indexed by the LPCs because they arise later than components linked to the processing of lexical forms and meanings (e.g. Brothers, Swaab, & Traxler, 2015; Grainger, Kiyonaga, & Holcomb, 2006; Holcomb & Grainger, 2006; Lau, Holcomb, & Kuperberg, 2013; see Nieuwland, 2019 for review of form-based components).

A second type of code-switching effect might occur—not because the listener switches from one lexicon to another—but rather because they have made a specific prediction about the form (i.e. the language) of the word they are about to hear, and that prediction is violated when they hear a code-switched word instead. On this account, code-switched words are no different than any other unexpected word (cf. Moreno et al., 2002). This hypothesis is consistent with our current understanding of the N400. The N400 was first discovered in contexts where a highly predictable word is replaced with an unexpected word, e.g., “He spread the warm bread with…*socks*” (Kutas & Hillyard, 1980). Subsequent studies have demonstrated that the magnitude of the N400 varies continuously with the predictability of a word (Borovsky, Elman, & Kutas, 2012; Brown & Hagoort, 1993; Federmeier & Kutas, 1999; Fernandez et al., 2019; Kutas & Federmeier, 2000; Kutas & Hillyard, 1984; Lau et al., 2013; Lau, Almeida, Hines, & Poeppel, 2009; Van Berkum, Hagoort, & Brown, 1999). N400s also decrease for a given word as the cumulative contextual constraints make that word increasingly likely (Van Petten, 1993). This pattern is compatible with a framework in which top-down processes generate predictions about upcoming words (Altmann & Kamide, 1999; Kamide, Altmann, & Haywood, 2003), making it easier to access the meanings of words that are consistent with those predictions (Federmeier, 2007; Hale, 2001; Kuperberg, 2016; Kuperberg et al., 2019; Levy, 2008; Schwanenflugel & LaCount, 1988). Some studies reveal that the N400 is sensitive to expectations that are linked to meaning (or semantic features) of the word (e.g. Federmeier & Kutas, 1999; Federmeier, McLennan, De Ochoa, & Kutas, 2002; Kuperberg, 2007; Kuperberg et al., 2019). Other studies have also found evidence that these expectations can lead to the pre-activation of syntactic or phonological features of the word (DeLong, Urbach, & Kutas, 2005, 2017; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; Wicha, Bates, Moreno, & Kutas, 2003; Wicha, Moreno, & Kutas, 2003, 2004; cf. Ito, Martin, & Nieuwland, 2017; Nieuwland, 2019). If top-down processing generates an expectation for a particular lexical item within a particular language (rather than just the concept encoded in that word), then we would expect to get N400 effects to code-switched words purely as a side effect of these lexical predictions. In the next section, we consider both hypotheses (i.e. discrete switch costs vs. general expectation-based costs), and then discuss an experimental manipulation designed to tease them apart.

### *1.3. Testing the one-cost and two-cost hypotheses*

The present study tests two alternative theories for the N400 effects in code-switching studies. The first theory claims that language identification or recognition (e.g. English *or* Spanish) precedes lexical processing. On this hypothesis, the N400 to code-switched words reflects the cost of switching from one lexicon to another, while the N400 to words that do not fit the context reflects a slowdown in lexical access in the absence of contextual clues. We will call this the *two-cost* hypothesis. The second theory claims that lexical processing occurs prior to and independent of recognizing the language of the word being processed. On this hypothesis, the N400 to code-switched words and the N400 to words that do not fit the context have the same root cause. In both cases, the listener hears a word that is inconsistent with their expectations, and thus that word is more difficult to access. We will call this the *one-cost* hypothesis.

These accounts make different predictions about what would happen if a bilingual encountered a code-switched word *that was also a poor fit for the context.* The two-cost account predicts additivity in the N400 response, such that the size of the N400 would be roughly equivalent to the sum of the N400 effects for the two separate violations. Critically, the one-cost account predicts that the N400 response to the double violation should be roughly the same as the cost for either the (correct) code-switched word or the poorly fitting word in the matrix language—as in all three cases, the expected word did not appear. The present study tests whether these two N400 effects are additive or whether all three violations have similar N400 effects.

There are two studies with data that bear on these hypotheses: Both Liao and Chan (2016) and FitzPatrick and Indefrey (2014) conducted experiments with 2 × 2 manipulations of the language of the target word (switched vs. non-switched) and how well the word fits into the preceding sentence context. These factorial designs provide critical data that could test the question of additivity—however, neither group considered their data in light of the two hypotheses above. Instead, they arrived at very different conclusions, perhaps because they set out to test different questions or used different interpretive frameworks to understand their data.

In the first relevant study, Liao and Chan (2016) had Mandarin-Taiwanese bilinguals listen to sentences that were played word-by-word with 200 ms pauses between each word. In addition to manipulating the presence/absence of a code-switch and the contextual fit of their target words, they also manipulated the direction of the language switching, i.e., switching from the participants' dominant language into their weaker one or vice versa. The authors concluded that the costs associated with code-switching are greater in cases of dominant-to-weaker language switching (which is also the switching direction used in the present study). When collapsing across switching direction, the authors find an interaction between contextual fit and code-switching in an early negative component—namely, the Phonological Mismatch Negativity (PMN), which emerged between 250 and 350 ms. This interaction is consistent with the one-cost hypothesis, as the negativities are similar across all three violation conditions and significantly different from the baseline condition (i.e. the expected word in the expected language).

Three features of this study, however, prevent us from drawing strong conclusions with respect to our current hypotheses: 1) The use of word-by-word auditory presentation may have resulted in processing strategies that are different from those in a more naturalistic listening task, perhaps because it leaves more time for prediction and subarticulation; 2) The interaction was found on a component that is not typically observed in code-switching studies, possibly due to the presentation method. Thus, it is unclear whether or not this finding would generalize to the N400 and LAN effects, which are the most prevalent initial effects of code-switching in the ERP literature; 3) Because the pattern of effects is radically different across the two switching directions, it is difficult to know how to interpret findings that appear when you collapse across them. Taken together, we cannot tell (from the data presented) whether the one-cost data pattern is reliably present in either switching direction.

In the second relevant study, FitzPatrick and Indefrey (2014) had Dutch-English bilinguals listen to sentences in either their native language (Dutch) or their non-native language (English). The authors' central goal was to explore bilinguals' comprehension of interlingual homophones (e.g. *pet* can mean *an animal companion* in English or *a cap-like hat* in Dutch). However, they also conducted two experiments (one in English, one in Dutch) with a 2 × 2 manipulation of code-switching and semantic congruence. In this study, the authors pursued a different analytic approach. Rather than directly comparing the three violation conditions to one another, they focused on the presence and the timing of the semantic congruity effects within each language separately. Their conclusions are broadly consistent with the two-cost hypothesis. Specifically, they propose that there are semantic incongruity effects for all incongruous words (regardless of language) and that these effects emerge earlier for non-switched words because they can be accessed more easily. They also propose that there is an early transient negativity for congruous code-switched words due to the priority given to the matrix language during lexical access. But curiously, their data patterns also seem consistent with the one-cost hypothesis: there is an interaction in the early N400 time window due to the fact that the effects in the three violation conditions appear to be similar in magnitude, timing, and scalp distribution. However, comparisons across these conditions are hindered by the fact that the sentence frames are different for each condition.

In sum, the studies to date do not provide conclusive evidence for either the one-cost or the two-cost hypothesis, largely because these studies did not originally set out to address these particular hypotheses. The two studies above reached divergent conclusions, despite having broadly similar data patterns—an issue that we will return to in the General Discussion.

### 1.4. The present study

The present study asks whether the neural markers associated with code-switching costs are best understood as difficulties specific to switching languages or as indirect consequences of processing unexpected words or encountering unexpected events. To test this question, we systematically compared bilinguals' ERP responses to three types of words: code-switched words, unexpected (within-language) words, and double violations (code-switched words that weakly fit the context). If there are unique costs to code-switching, we should expect an additive N400 response to the double violation condition. This study differs from most of the prior ERP studies on code-switching in one critical way. Many of the prior studies have used artificial tasks (e.g. lexical decisions, naming tasks) with unnatural code-switches (i.e. switches in formal, written texts or predictable sentences with the last word switched). Recently, many observers have noted that these artificial contexts are not the most accurate way to study how bilinguals understand code-switching in the wild (Blanco-Elorrieta & Pylkkänen, 2016, 2017, 2018; Fernandez et al., 2019; Van Hell et al., 2018, 2015). The present study takes a critical step toward a more naturalistic approach with our Storytime task. In our paradigm, participants listen to real, unscripted, spoken narratives. We took these narratives and spliced in a carefully counterbalanced experimental manipulation. This allowed us to precisely study intra-sentential code-switching in a rich, variable, naturalistic context. To preview our findings, we were able to successfully replicate the N400 and LPC effects found in the prior literature using our design. And critically, this design allowed us to test whether the processing costs of language and contextual fit were additive.

## 2. Method

### 2.1. Participants

Thirty-four Spanish-English bilinguals from Harvard University participated in this experiment. We excluded two participants due to experimenter error, resulting in 32 participants in the final sample. We did not perform an a priori power analysis—rather we based our final sample size on the prior literature. During data collection, however, we implemented a stopping rule: participants' EEG data were cleaned incrementally and replaced if more than 25% of all trials were rejected (see rejection criteria in Section 2.4.1). This procedure continued until we had usable data from 32 participants.

We recruited participants from student-run organizations and from Harvard's study pool. Participants were compensated $10/h or received two study credits for participating. We screened participants for eligibility by asking them six questions about their proficiency in Spanish and their overall exposure to code-switching in their community. This language screener is accessible on the Open Science Framework (OSF; see https://osf.io/jwqpr/). Participants self-reported that they were highly-proficient in Spanish (*intermediate-level* = 1, *advanced-level* = 3, *native-level* = 28) and had considerable exposure to code-switching (*never heard code-switching* = 1, *sometimes* = 9, *often* = 22). All participants reported learning Spanish before age eight with the average age of acquisition being 0;11 (*SD* = 1;11).

**Table 1**
Participants' responses from the LEAP Questionnaire.

| *Assessment Type* | Language | | | |
|---|---|---|---|---|
| ***Language Proficiency Assessment*** | **Spanish** | | **English** | |
| Average age of acquisition | 0;8 | (0;4) | 3;9 | (0;8) |
| Average age of fluent speaking | 6;0 | (1;0) | 6;10 | (0;11) |
| Average age of fluent reading | 8;5 | (1;1) | 7;7 | (0;10) |
| Percentage of language exposure | 30.5 | (3.4) | 69.3 | (3.3) |
| Language exposure composite score[1] | 4.3 | (0.3) | 7.1 | (0.3) |
| Self-rated proficiency (speaking) | 8.6 | (0.3) | 9.8 | (0.1) |
| Self-rated proficiency (listening) | 9.1 | (0.2) | 9.8 | (0.1) |
| Self-rated proficiency (reading) | 8.0 | (0.1) | 9.8 | (0.1) |
| Percentage of time speaking | 32.6 | (4.5) | 67.3 | (4.5) |
| Language proficiency composite score[2] | 8.6 | (0.2) | 9.8 | (0.1) |
| ***Language Dominance Assessment*** | **Spanish** | | **English** | |
| Number of participants listing this language as dominant (out of 22) | 2 | | 20 | |
| Number of participants listing this language as first acquired (out of 22) | 19 | | 3 | |

*Notes:*
Means and standard errors are reported. All ages are reported as years followed by months. [1] The composite scores for *language exposure* (highest score = 10) were created by averaging self-reported ratings (out of 10) of how often participants are currently exposed to Spanish/English when 1) interacting with friends, 2) interacting with family, 3) reading, 4) language apps or websites, 5) watching TV, and 6) listening to radio/music. [2] The composite scores for *language proficiency* (highest score = 10) were created by averaging the self-reported scores (out of 10) for speaking, understanding, and reading in each language.

We did not intend for this initial screener to be a robust language history survey—thus, after the experiment, we recontacted all of our original participants to have them complete the Language Experience and Proficiency (LEAP) Questionnaire (Marian, Blumenfeld, & Kaushanskaya, 2007; Kaushanskaya, Blumenfeld, & Marian, 2020). Roughly two-thirds of our study population agreed to participate and completed the LEAP Questionnaire (22 out of 32 participants). After receiving the additional information from the LEAP Questionnaire, we determined that our population was dominant in English and considered Spanish to be their first language. Although participants considered themselves to be largely dominant in English, the levels of proficiency across both languages were comparable with slightly more variability in the proficiency for Spanish (see Table 1). The results from the LEAP Questionnaire are largely consistent with our findings from the initial screener—more information can be found on OSF (https://osf.io/jwqpr/).

### 2.2. Stimuli

To preview our stimuli, we manipulated 120 target words in sentences within two oral stories, which were largely in English. There were two factors in our design: 1) how well does a word fit into its preceding context putting aside its language (strong-fit, weak-fit) and 2) what language is the word in (English, Spanish). Below is an example of a target sentence in all four conditions:

(2) a. And the wig itself is so hot and heavy on my **head**. (**Strong-fit** English)
b. And the wig itself is so hot and heavy on my **cabeza**. (**Strong-fit** Spanish)
c. And the wig itself is so hot and heavy on my *cranium*. (*Weak-fit* English)
d. And the wig itself is so hot and heavy on my *cráneo*. (*Weak-fit* Spanish)

Thus, the Spanish conditions involved code-switching while the English conditions did not. Note, all Spanish target words were translations of either the English strong or weak-fit conditions (i.e. **head-cabeza***, cranium-cráneo*). In the sections below, we explain how these stimuli were created.

### *2.2.1. Oral story selection*

We selected two stories from a collection of unscripted, oral performances known as 'Moth' stories. The Moth is a non-profit organization dedicated to "the art and craft of storytelling" (see https://themoth.org/). These Moth stories contain properties of naturalistic speech (e.g. disfluencies, redundancies, colloquialisms) that are typically absent in more formally-scripted performances. The stories that we selected were originally performed in English (roughly 20 min each) and had a combined total of 343 sentences. From these stories, we selected 120 target sentences (described below), resulting in an approximate 2:1 filler-to-target sentence ratio.

### *2.2.2. Target sentence and English noun selection*

Words that are preceded by a supportive context are processed more easily, as indexed by faster behavioral responses and reduced N400 responses (Ehrlich & Rayner, 1981; Fischler & Bloom, 1979; Jordan & Thomas, 2002; Kutas & Federmeier, 2011). Thus, for our experiment, we wanted the original target words to be as predictable (and as easy to process) as possible. One limitation to using naturally produced narratives is that we were not able to create highly predictable sentence contexts—so, we settled on selecting the most predictable nouns available in the Moth stories. To characterize the predictability of the nouns in these stories, we conducted two cloze tasks: one using written versions of our stories and the other using our final auditory stimuli (see Taylor, 1953 for information on cloze tasks). Both of these cloze tasks were created on the IbexFarm experimental software (http://spellout.net/ibexfarm/) and made available to participants on Amazon's Mechanical Turk (https://www.mturk.com).

The first cloze task was designed to collect the cloze probabilities for *all* of the nouns in both stories. To do this, 72 participants read one of the two stories from beginning to end. Participants saw short, fragmented sentences that ended right before a noun. They were then asked to guess the next word. After guessing, they were shown the actual noun, and this procedure continued until participants had guessed every noun in the story. We then calculated each noun's *cloze probability,* i.e., the proportion of times that participants provided the target noun given its context. Cloze probability is argued to be a good measure of how easily a word can be predicted during language comprehension (e.g. Federmeier & Kutas, 1999; Staub, Grant, Astheimer, & Cohen, 2015), and it is inversely correlated with a word's N400 amplitude (Kutas & Hillyard, 1984). Based on these data, we identified the most predictable nouns by sorting the cloze probabilities and taking the top 120 targets. Occasionally, one noun (e.g. head) had high cloze probabilities in multiple sentences (i.e. one noun *type* had multiple high-cloze *tokens*); however, we never used the same target noun more than three times (and never more than twice in a single story). These 120 nouns became the strong-fit English target words, and they had an average cloze probability of 61% with a range of 13–98%.

Given this range of cloze values, we designed a second cloze task to characterize how predictable our target words were in the final recordings that we used in our EEG study. Note, in Section 2.2.7 below, we describe how these final recordings were created. In this audio cloze task, 45 participants heard both stories in their entirety, and we counterbalanced which story was played first. The recordings would pause right before each target word, and participants would then guess the next word. After guessing, the recording would rewind to the start of that target sentence, so that the participants could hear the actual story continuation. At the end of the task, participants were asked a series of questions to determine their level of engagement and overall comprehension during the task. Results indicated that participants understood the stories, as their comprehension accuracy was 91.7% (*SE* = 2.1%). Similar to the written cloze task, the target words had an average cloze probability of 61.3% (*SE* = 2.3%); however, this time the range was slightly wider with values ranging from 2.8–94.3%. In the General Discussion, we will address the implications of having a wide range of cloze values for our targets. The full list of target nouns and their cloze values can be found in Appendix A, as well as on OSF (see https://osf.io/jwqpr/).

### *2.2.3. Weak-fit English noun selection*

Next, we selected the weak-fit English target nouns. We wanted the strong and weak-fitting pairs to be semantically related to one another. To do this, we took the same sentences that we identified above and replaced the high cloze noun with a noun that still made sense in that context—but had never been used by our MTurk norming sample in the written cloze task (e.g. And the wig itself is so hot and heavy on my *cranium*). Thus, these weaker alternatives had a cloze probability value of roughly zero given our target contexts—although, this does not imply that these words could *never* be used in our sentences. Critically, the weak-fit nouns expressed events that were plausible and did not disrupt the overall storyline (e.g. It had a *mind/brain* of its own; I put it on one of my dresser *drawers/shelves*; My hair does fall out, first in these strands in my brush, and then in clumps in my shower *drain/hole*).

After creating the strong and weak-fit pairs, we quantified the semantic relatedness between them by calculating their cosine similarities. We used the *LSAfun* package (Günther, Dudschig, & Kaup, 2015) in the R statistical computing environment (R Core Team, 2020). To do this, we first selected a semantic space in which each target word is represented as a single vector—for our analyses, we used the semantic space from Baroni, Dinu, & Kruszewski (2014). Then, we measured the cosine of the angle between the vectors for each strong and weak-fit word pair. Cosine similarity values can range between −1 (highly dissimilar) and 1 (highly similar). A cosine value of zero indicates that the two words are orthogonal to one another (for more information, see Günther et al., 2015). Across all pairs, the average cosine similarity was 0.26, which is equivalent to the similarity between the words *dog* and *mouse* in this semantic space. The range of similarities was 0.02–0.69, which is similar to the comparison of *dog* and *osprey* and then *dog* and *puppy* respectively.

Next, we decided to assess the fit of our words within our target sentences. To do this, we conducted a naturalness rating task on IbexFarm and Amazon's Mechanical Turk. In this task, participants read all of the target sentences from one of the two stories. The sentences were presented in their entirety with either the strong-fit or weak-fit target. All target words were marked with asterisks (e.g. *head*). Participants were then instructed to rate the naturalness of the target word in the sentence using a 7-point Likert scale (7 = *Very Natural*, 1 = *Very Unnatural*). An unnatural word was described as a word that a person might have a hard time imagining someone saying in this context (and vice versa for a natural word). To determine any differences across conditions, we used a two-tailed paired-samples *t*-test. We found that our strong-fit targets were rated significantly higher ($M = 6.36$, $SE = 0.06$) than our weak-fit targets ($M = 3.28$, $SE = 0.08$), $t(129) = 32.19$, $p < .001$.

### *2.2.4. Spanish noun selection*

In our design, English was the matrix language, which meant that the Spanish words served as the code-switched items, and the English words served as controls. We assumed that English would be the dominant language for the majority of our study population, as they all attended a university where the language of instruction is English. Results from the LEAP Questionnaire confirmed that this assumption was correct, as most participants said English was their dominant language. Prior studies have shown that bilinguals are better able to predict upcoming words when comprehending sentences in their dominant language (see Ito, 2016; Ito et al., 2017; Ito & Pickering, 2021; Ito, Pickering, & Corley, 2018; Kotz & Elston-Güttler, 2004; Liao & Chan, 2016). In the present study, the critical words in the Spanish conditions were translation equivalents of the strong-fit and weak-fit English targets. The translations were provided by the second author (EM), who is a native Spanish speaker. In some cases, the direct translation of an English target was a Spanish cognate (e.g. *ceremony* and *ceremonia*). The use of cognates was judged to be unavoidable, so we equated the number of cognates in the strong-fit and weak-fit conditions: 36 cognates per condition, 72 cognates in total out of 240 Spanish words. After testing, we realized that two of these cognates are considered to be variants (*subjeto* for *sujeto*) or are not formally-accepted (*disturbia* for *disturbio*) according to the *Diccionario de la lengua española* (Dictionary of the Spanish language). Given this finding and the number of cognates in our study, we conducted our primary analyses with and without these cognate items. However, removing the cognates did not change the overall pattern of findings. The results from the analyses without cognates are available in our annotated analyses on OSF (see https://osf.io/jwqpr/). Again, all target trials are listed in Appendix A and on OSF.

### 2.2.5. *Assessing our critical manipulations within our study population*

As we mentioned above, we recontacted all of our original study participants and asked them to complete a language survey and a ratings task. In the ratings task, we asked participants to re-read the original stories and provide naturalness ratings for our strong and weak-fitting English words. To do this, participants read both stories in their entirety, chunk-by-chunk. Each chunk contained one English target word (either the strong or weak-fitting version). Participants then rated the naturalness of the word given the story context on a sliding scale from 0 (*Very Unnatural*) to 100 (*Very Natural*). After rating this target word, participants were presented with a potential Spanish translation of that word—half of these translations were the ones used in the EEG study while the other half were foils. The foils were simply the Spanish translations of other words from the trials that the participant did not see. Finally, participants rated these translations as being acceptable or unacceptable (or they indicated that they did not know the Spanish word, the English word, or both of the words presented). This procedure continued until participants had read both stories and rated all of the target words that they had encountered in the original EEG study.

We had 20 out of the original 32 participants complete this ratings task. Results indicated that participants consistently rated the strong-fit English words as being more natural ($M$ = 94.3, $SE$ = 0.9) than their weak-fit English alternatives ($M$ = 35.9, $SE$ = 2.0). Participants also strongly accepted the translations that we used in the original study ($M$ = 90% acceptable, $SE$ = 1.0%) and strongly rejected our foil translations ($M$ = 2.8% acceptable, $SE$ = 1.0%). Looking at the code-switched conditions individually, we found that strong-fit and weak-fit Spanish translations were accepted 95.8% ($SE$ = 1.2%) and 83.5% ($SE$ = 2.2%) of the time respectively. Finally, participants knew nearly all of our target words: strong-fit English words ($M$ = 99.8% known, $SE$ = 0.1%); strong-fit Spanish words ($M$ = 97.5%, $SE$ = 0.7%); weak-fit English words ($M$ = 98.4%, $SE$ = 0.4%); and weak-fit Spanish words ($M$ = 93.5%, $SE$ = 1.2%).

### 2.2.6. *Other properties of the stimuli: word frequency, length, and sentence position*

There are a few other stimulus properties that we did not consider when initially selecting our target words; however, they are still important to characterize. These properties are word frequency, word length, and the word's position in our target sentences—and we describe them in more detail below.

*Word frequency.* For our target words, we used the standardized word frequencies (per million words) from the SUBTLEX$_{US}$ (Brysbaert & New, 2009) and the SUBTLEX$_{ESP}$ (Cuetos, Glez-Nosti, Barbon, & Brysbaert, 2011) subtitle corpora. The SUBTLEX$_{US}$ corpus has roughly 51 million words from American English subtitles (1990–2007). The SUBTLEX$_{ESP}$ corpus has roughly 41 million words from Spanish subtitles (1990–2009) that contain both Iberic and Latin American language variants. These Spanish subtitles came from a range of Spanish-speaking countries such as Argentina, Chile, Colombia, Mexico, Peru, and Spain, as well as from the United States. To evaluate differences in word frequencies across conditions, we used a series of two-tailed, paired-samples $t$-tests (Bonferroni-corrected $\alpha = 0.01$). The two irregular cognates (mentioned above) did not have any frequency values, so we included the frequency values for their accepted forms: *sujeto* and *disturbio*. When comparing the frequency of all English words ($M$ = 176.54, $SE$ = 22.5) to all Spanish words ($M$ = 166.98, $SE$ = 18.7), there was no significant difference in frequency, $t(239) = 0.33$, $p = .74$. However, there were pairwise differences between conditions, such that the strong-fit English words ($M$ = 328.72, $SE$ = 40.30) were more frequent than the weak-fit English words ($M$ = 23.11, $SE$ = 4.32), $t(119) = -7.70$, $p < .001$, and the strong-fit Spanish words ($M$ = 287.81, $SE$ = 32.65) were more frequent than the weak-fit Spanish words ($M$ = 43.67, $SE$ = 9.48), $t(119) = -7.28$, $p < .001$. There were no significant differences between the two strong-fit conditions, $t(119) = -1.47$, $p = .14$, nor the two weak-fit conditions, after correcting for multiple comparisons, $t(119) = 2.48$, $p = .015$.

*Word length.* Next, we compared the length of our target words. To do this, we calculated two measures of word length: the raw number of syllables and the duration (ms) of the words from our recordings. To evaluate any differences, we again used a series of two-tailed, paired-samples $t$-tests (Bonferroni-corrected $\alpha = 0.0125$). Unsurprisingly, there were significant differences in the number of syllables between all four conditions, reflecting the tendencies for both Spanish words and less frequent words to have more syllables in general: the strong-fit English words ($M$ = 1.40, $SE$ = 0.06) were shorter than the strong-fit Spanish words ($M$ = 2.63, $SE$ = 0.09), $t(119) = 14.49$, $p < .001$; the weak-fit English words ($M$ = 1.79, $SE$ = 0.07) were shorter than the weak-fit Spanish words ($M$ = 2.99, $SE$ = 0.09), $t(119) = 13.60$, $p < .001$; the strong-fit English words were shorter than the weak-fit English words, $t(119) = 4.97$, $p < .001$; and the strong-fit Spanish words were shorter than the weak-fit Spanish words, $t(119) = 2.94$, $p < .01$. For target word duration in milliseconds, the strong-fit English words ($M$ = 538.75, $SE$ = 23.48) were significantly shorter than both the weak-fit English words ($M$ = 617.24, $SE$ = 24.15), $t(119) = -4.67$, $p < .001$ and the strong-fit Spanish words ($M$ = 647.17, $SE$ = 25.51), $t(119) = -6.25$, $p < .001$. However, there were no significant differences between the durations of the weak-fit Spanish words ($M$ = 700.99, $SE$ = 27.10) and the strong-fit Spanish words, $t(119) = -1.77$, $p = .078$, nor the two weak-fit conditions, $t(119) = -2.23$, $p = .027$ (after correcting for multiple comparisons).

*Sentence position.* Finally, we evaluated where our critical words appeared in each target sentence. Roughly 30% of all target words appeared at the end of the sentence—the rest of the targets appeared somewhere in the middle. To quantify the position of our target words, we calculated how many words preceded the targets and how much of the total sentence had been heard prior to the target. On average, there were 13 words prior to our critical words (range: 3–35 words) and roughly 70% of the entire sentence had been heard by the onset of our targets (range: 12.5–100%). Below, we have summarized all of the relevant properties of our naturalistic stimuli (see Table 2).

### 2.2.7. *Audio stimulus creation*

After selecting our target nouns and sentences, we created the materials for our Storytime paradigm. First, we recorded the two stories in their entirety, making an effort to preserve the disfluencies and redundancies from the original Moth performances. We then recorded each of the target sentences individually in each of the four conditions. Next, we spliced all target words from these individual recordings into the larger story recordings, which allowed us to keep the audio before and after the targets identical across conditions. To avoid splicing artifacts, we respected co-articulation by splicing in, at most, the word before and after the target. Finally, we found all of the target onset times manually using the phonetic software, PRAAT (Boersma & Weenink,

**Table 2**
Critical properties of our experimental stimuli.

| Condition | Frequency | Syllables | Duration (ms) | Word Known[1] | Translation Acceptability | Naturalness (0–100) |
|---|---|---|---|---|---|---|
| *Strong-fit English* | 328.7 (40.3) | 1.40 (0.1) | 538.8 (23.5) | 99.8% (0.1) | – | 94.3 (0.9) |
| *Weak-fit English* | 23.1 (4.3) | 1.79 (0.1) | 617.2 (24.2) | 98.4% (0.4) | – | 35.9 (0.2) |
| *Strong-fit Spanish* | 287.8 (32.6) | 2.63 (0.1) | 647.2 (25.5) | 97.5% (0.7) | 95.8% (1.2) | – |
| *Weak-fit Spanish* | 43.7 (9.5) | 2.99 (0.1) | 701.0 (27.1) | 93.5% (1.2) | 83.5% (2.2) | – |
| **Average cloze probabilities** | | *Written Cloze (N = 36)*<br>*Auditory Cloze (N = 45)* | | $M = 61.0\%$ (13.0–98.0%)<br>$M = 61.3\%$ (2.8–94.3%) | | |
| **Average sentence positions** | | *Amount of prior context*<br>*Number of prior words* | | $M = 70.6\%$ (12.5–100%)<br>$M = 13$ words (3–35 words) | | |

*Notes:*
Means are reported alongside the range or their standard errors. The top half of the table summarizes the critical properties of our manipulations across the four conditions. [1]Word Known variable represents the percentage of words recognized by bilinguals in each condition. The bottom two rows reflect the average cloze probabilities and sentence positions for our target baseline words (i.e. the strong-fit English conditions).

2001), which we later used to time-lock our ERP responses.

Since we used these onset times in our final analyses, we wanted to ensure their accuracy. First, all onset times were determined by the second author (EM), who is a native speaker of Spanish. Second, these onset times were confirmed by the first author (AY), who is a native speaker of English. Third, we transcribed each target word into IPA and considered how the phonetic differences across the two languages could causes differences in onset times (e.g. the lack of aspiration for word-initial plosives in Spanish). Finally, we relied on visual cues like formant transitions, changes in pitch, frequency, and/or intensity, as well as our own intuitions when marking word boundaries. An example of a specific time-locked stimulus, and all of our phonetic transcriptions can be found on OSF (see https://osf.io/jwqpr/).

We used a Latin Square design with four lists. Targets were assigned to item groups in such a way that in any given list no adjacent items were ever in the same condition. In our Storytime paradigm, there is no traditional trial structure, meaning that all intervening sentences serve as fillers. Some of our target sentences occurred back-to-back, whereas others occurred with 16 sentences in between. On average, the number of intervening sentences was 1.73 ($SE = 0.23$). This number may make it seem like our target words were presented in rapid succession; however, our sentences varied widely in their total durations. Thus, a more accurate characterization of the timing between target trials is the *inter-stimulus interval* (ISI), which represents the time between the offset of one target word and the onset of the next. The average ISI in our recordings was 17.8 s ($SE = 1.35$) with a range of 1.8 to 74.3 s.

Finally, we wanted to characterize the speech rate and the average *stimulus onset asynchrony* (SOA) of our recordings. These two properties are often tightly controlled in traditional psycholinguistic experiments—thus, we calculated these metrics for ease of comparison. To calculate the speech rate, we used a PRAAT script written by De Jong and Wempe (2009), which finds the nucleus of each syllable in a recording and uses that information to determine metrics like speech rate, phonation time, and average syllable duration automatically. The average speech rate across recordings was 2.84 ($SE = 0.02$) syllables per second. Next, we calculated the average SOA (i.e. the time between the onset of a target word and the onset of the next target word). To do this, we obtained the onset times for each word in our recordings using the Gentle forced aligner from Ochshorn & Hawkins 2016 (see https://lowerquality.com/gentle/). Then, we calculated the SOA values for each word, removing the values for all of the target words (as they differed across story versions) and all of the function words (as they are extremely short and would skew the overall average). Results indicated that, on average, the SOA in our stories was 521.9 ms ($SE = 8.0$).

## 2.3. *Procedure*

In the present study, participants passively listened to two short stories during a single EEG recording session. This naturalistic listening technique circumvents typical difficulties associated with traditional experimental designs, i.e., the need for many disjointed, out-of-the-blue sentences, and an extensive use of filler sentences. This technique also allows participants to hear rich discourses, promotes attention, and is arguably more engaging. Each story lasted about 20 min, and there was a short break in between them. The order of story presentations was counterbalanced across participants. We intended to test the same number of participants in each list, but one participant was run in the wrong list, resulting in a slightly uneven distribution (i.e. 7, 9, 8, and 8 per list). All participants sat approximately 40 in. away from a TV monitor, which displayed an unrelated video of a beach sunset (available on OSF, https://osf.io/jwqpr/). In this video, the sun was slowly moving along a vertical axis in the center of the display. This video served as a focal point for participants and helped minimize sharp horizontal and vertical eye movements throughout the study. We encouraged participants to blink as little as possible and to reduce facial tension (i.e. keep their forehead and jaw relaxed). At the end of the study, participants were fully debriefed and given the opportunity to ask any questions. All of our experimental procedures were approved by the Harvard Committee on the Use of Human Subjects (CUHS).

## 2.4. *EEG recording*

We recorded the electroencephalogram using Brainvision's actiChamp System. Online signals were recorded from 31 active Ag/AgCl electrodes embedded in an elastic cap (EASYCAP GmbH). The ground and reference electrodes were the pre-frontal electrodes FPz and FP1 respectively. A pair of passive EOG electrodes connected to the BIP2AUX adapter was attached above and below the left eye to monitor for vertical eye movements. We continuously recorded at a sampling rate of 500 Hz and kept electrode impedances below 20 kΩ.

### 2.4.1. *EEG pre-processing*

We pre-processed and analyzed our EEG data using both EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014) toolboxes. First, we downsampled the data to 200 Hz and re-referenced offline to the average of the left and right mastoids. The EEG signals were then filtered using an IIR filter with a bandwidth of 0.01-30 Hz. We then identified and corrected eye blink artifacts using an Independent Component Analysis (ICA). Next, we created epochs that extended from 200 ms before stimulus onset to 2000 ms post-stimulus onset. We then performed a two-step artifact rejection process: First, we subjected all epochs to an automatic rejection procedure that removed trials with voltages exceeding −90 or 90 μV. This procedure rejected 15.2% of all 3840 trials (4 conditions × 30 trials per condition × 32 electrodes = 3840 observations). No condition had more than 17% of their trials rejected: strong-fit English = 16.25% (*SE* = 0.02%); strong-fit Spanish = 16.67% (*SE* = 0.02%); weak-fit English = 14.06% (*SE* = 0.01%); weak-fit Spanish = 13.75% (*SE* = 0.02%). Using a generalized logistic mixed model, we confirmed that there were no statistical differences in rejection rates between the two strong-fit conditions ($\widehat{\beta} = 0.03$, $SE = 0.13$, $z = 0.27$, $p = .79$), the two English conditions ($\widehat{\beta} = -0.17$, $SE = 0.13$, $z = -1.33$, $p = .18$), nor the strong-fit English and the weak-fit Spanish conditions ($\widehat{\beta} = -0.21$, $SE = 0.13$, $z = -1.59$, $p = .11$). Second, we visually inspected each electrode to check the quality of the EEG recording. Specifically, we looked for eye motion artifacts (e.g. horizontal movements), electrocardiographic (ECG or EKG) and other muscular artifacts, and instances of power line noise, channel noise, and channel pop-off effects. If a single electrode had multiple artifacts that were not removed or corrected using the prior methods, we interpolated the entire electrode channel. However, we never interpolated the three main electrodes along the midline (Fz, Cz, and Pz). On average, we interpolated roughly 3 out of 32 electrodes across all participants.

### 2.5. Statistical analyses

#### 2.5.1. Pre-registered mean amplitude analyses (300-500 ms)

We first averaged the ERP amplitudes from our pre-registered N400 time window of 300-500 ms post-stimulus onset for each trial (and channel location). This process created 3840 mean amplitude values (120 trials × 32 electrodes) for each participant prior to exclusions. For our mean amplitude analyses, however, we only used the averages from three midline electrodes (Fz, Cz, and Pz). We then modeled these averages with a linear mixed effects model using the *lme4* package in the R statistical computing environment (Bates, Mächler, Bolker, & Walker, 2014; R Core Team, 2020). Our model had dummy-coded, fixed effects of *contextual fit* (strong-fit = 0, weak-fit = 1) and *language* (English = 0, Spanish = 1) as well as their interaction.[1] The model had a maximal random effects structure: there were random intercepts and random slopes for *language*, *contextual fit*, and their interaction for both participant and item grouping factors.

To evaluate significance, we adopt the convention of having an absolute value of *t* greater than 2 (Gelman & Hill, 2007). This is due to the on-going debate about how to best calculate the appropriate degrees of freedom for the test statistics in linear mixed effects models (see Baayen, 2008). However, we also report the *p*-values as calculated by the *lmerTest* package, as both methods of evaluation arrived at the same conclusions. The code for our statistical analyses and model comparisons can be found on OSF (see https://osf.io/jwqpr/).

#### 2.5.2. Exploratory permutation-based cluster mass analyses (0-2000 ms)

In an exploratory analysis, we used a permutation-based cluster mass technique (Fields & Kuperberg, 2020; Groppe, Urbach, & Kutas, 2011a; Maris & Oostenveld, 2007) to investigate the full range of effects during comprehension. This approach should both confirm any effects observed between 300 and 500 ms and detect other effects not captured by our pre-registered mean amplitude analyses. Mean amplitude analyses have been shown to have limited power for detecting small, long-lasting effects, as they often involve averaging across many electrodes and time points with small or absent effects.

Permutation-based cluster mass analyses do not have this limitation, instead they preserve power for effects that emerge slowly over time and broadly across the scalp (Fields, 2019; Groppe, Urbach, & Kutas, 2011b; for simulations, see Fields & Kuperberg, 2020). Permutation-based cluster mass analyses employ the following procedure: First, an ANOVA is performed at each electrode site for each time-point in the target window. The results from each of these spatially and temporally-distinct ANOVAs are compared to a threshold for cluster inclusion. We used a *p*-value of 0.01, as recommended for exploratory analyses looking at long time-windows (see Fields, 2019). All spatially and temporally-adjacent points (i.e. neighboring electrodes at similar times) with *p*-values exceeding this threshold are grouped into a single cluster. Then, for each cluster, we calculate a cluster mass statistic by summing all of the cluster's *F*-values. Finally, we evaluate a cluster's significance using permutation-based corrections for multiple comparisons. To do these corrections, we first create a distribution of possible cluster statistics computed from randomly-permuted data (with null effects). We then compare our observed cluster statistics to the null distribution to determine significance at a predetermined alpha level. For example, if we set α = 0.05, a significant cluster statistic would need to fall outside of the 95 percentile (i.e. 1- α) of the null distribution.

Prior to our cluster analyses, we downsampled the data to 100 Hz using the *boxcar* filter, which averages adjacent time-points together, reducing the data to the desired sampling rate. This procedure left us with 200 samples between -5 ms to 1985 ms post-stimulus onset. Note, this unusual time-window is a product of downsampling and re-baselining from −200 to 0 ms. We implemented our analyses using the Factorial Mass Univariate Toolbox extension (FMUT; Fields, 2017; Fields & Kuperberg, 2020) for the Mass Univariate Toolbox (MUT; Groppe et al., 2011a). We used the recommended number of 100,000 permutations and α = 0.05 (Fields, 2019) for our main ANOVA and our four pairwise comparisons, which further addressed effects of contextual fit and language. The pairwise tests were corrected for multiple comparisons by applying a new Bonferroni-corrected alpha level (α = 0.0125). In the sections below, we first present the results from the mean amplitude analyses, and then those from the cluster mass analyses.

## 3. Results and discussion

### 3.1. Averaged waveforms and topographic voltage maps

Fig. 1 shows the averaged waveforms for the three midline electrodes (Fz, Cz, and Pz), as well as the combined averages for left anterior, right anterior, left posterior, and right posterior electrodes. When interpreting these waveforms, it is important to take into account the differences that are often found between auditory and visual ERPs. Typically, auditory ERP components have earlier onsets, later offsets, and wider/broader distributions across the scalp relative to visual ERP components (e.g. Fernandez et al., 2019; Grey, Schubel, McQueen, & Van Hell, 2018; Grey & van Hell, 2017; Holcomb & Neville, 1991; Kutas & Federmeier, 2011; Liao & Chan, 2016; Ruigendijk et al., 2016). These differences are presumably due to the fact that visual words are presented all at once while auditory words unfold over hundreds of milliseconds (see Connolly, Phillips, & Forbes, 1995; Van Petten, Coulson, Rubin, Plante, & Parks, 1999).

### 3.2. Mean amplitude analysis at Fz, Cz, and Pz (300-500 ms): the N400 (pre-registered)

In our pre-registered time window of 300-500 ms, there were main effects of *contextual fit* ($\widehat{\beta} = -3.51$, $SE = 0.72$, $t = -4.86$, $p < .001$) and *language* ($\widehat{\beta} = -2.62$, $SE = 0.66$, $t = -3.99$, $p < .001$) at midline electrodes Fz, Cz, and Pz. The weak-fitting words and Spanish code-switches elicited greater N400 amplitudes relative to strong-fitting words and English non-switches respectively. These main effects, however, were superseded by an interaction between *contextual* fit and *language* ($\widehat{\beta} =$

[1] We ran additional models that included *midline electrode site* (Fz, Cz, or Pz) as either a random effect or a control variable. The models with *midline electrode site* as a random effect resulted in singular fits. The model with *midline electrode site* as a control variable did not improve model fit, $\chi^2(27, N = 32) = 0.65$, $p = .72$; thus, we collapsed across these midline electrode sites in our final analysis.

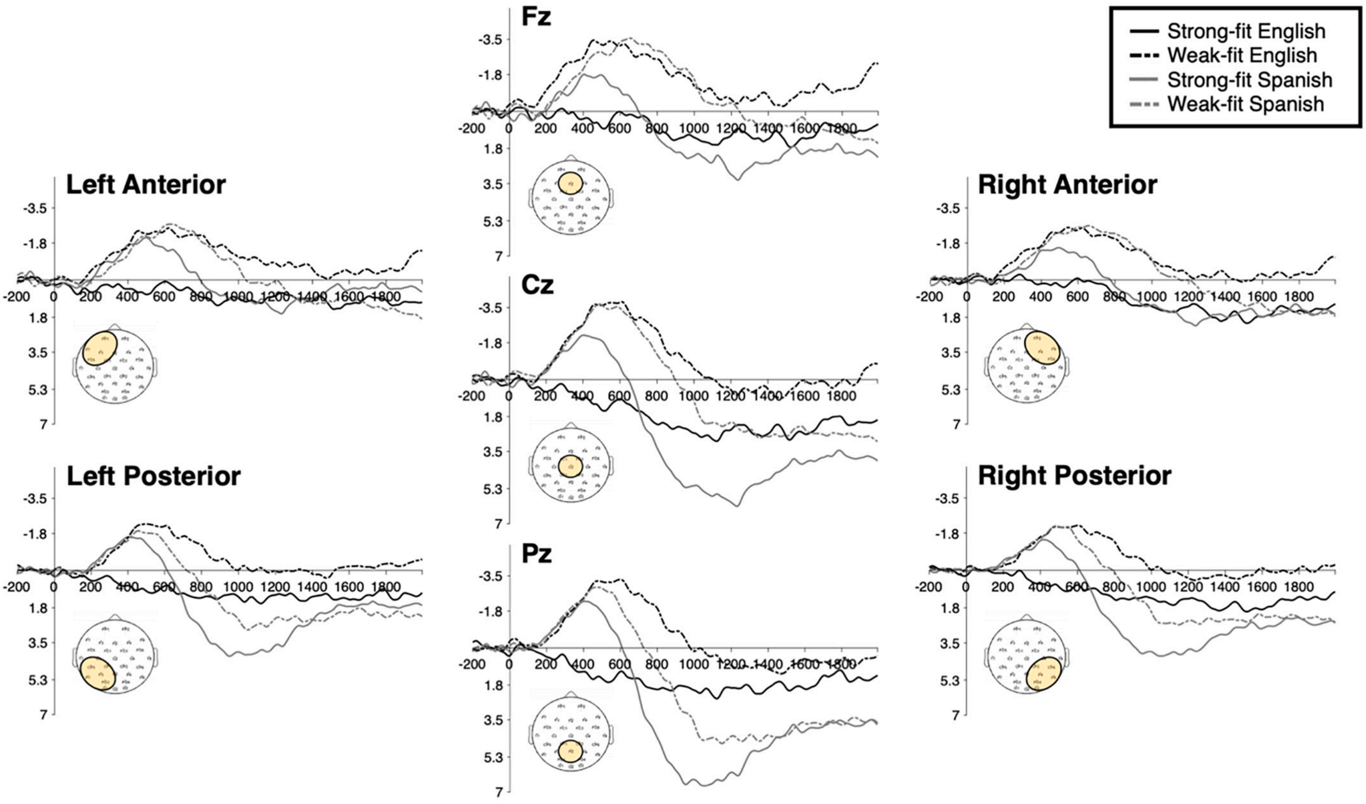

**Fig. 1.** *Grand waveforms for all conditions.* The averages (μV) for the midline electrodes Fz, Cz, and Pz are presented in the center panel. The combined averages for left anterior, right anterior, left posterior, and right posterior electrodes are positioned in their respective quadrants. The two dark lines at each site indicate the English conditions, while the lighter lines indicate the Spanish code-switched conditions. The solid lines indicate strong-fitting conditions, while the dotted lines indicate weak-fitting conditions. The canonical N400 effect is seen for all violation conditions (300-500 ms) and the LPC effect is seen for all code-switched words (700-1200 ms). Both effects are most prominent over parietal site Pz. The sustained negativity for weak-fitting words (700-1200 ms) is most prominent over frontal site (Fz). All waveforms were subjected to an additional low-pass filter (10 Hz) for plotting purposes.

2.87, $SE = 1.01$, $t = 2.86$, $p < .01$).[2]

To unpack this interaction, we performed planned pairwise comparisons using the *emmeans* package in R (Lenth, Singmann, Love, Buerkner, & Herve, 2020). The reported *p*-values were first obtained by comparing the pairwise estimates against a standard normal distribution (rather than the *t* distribution) and then adjusted for multiplicity using the Tukey method. The pairwise comparisons revealed a main effect of *contextual* fit between English conditions ($\widehat{\beta} = 3.52$, $SE = 0.72$, $z = 4.86$, $p < .0001$) such that weak-fit English words were more difficult to process, eliciting greater N400 responses relative to strong-fit English words. There were no differences between the N400 amplitudes elicited by the two Spanish code-switch conditions, suggesting that they were similarly difficult to process for listeners ($\widehat{\beta} = 0.64$, $SE = 0.68$, $z = 0.95$, $p = .78$). There was also a main effect of *language* between strong-fit words such that the strong-fit Spanish code-switches were more difficult to process and elicited greater N400 responses relative to strong-fit English words ($\widehat{\beta} = 2.62$, $SE = 0.66$, $z = 3.99$, $p < .001$). Finally, there were no differences between the two weak-fit conditions ($\widehat{\beta} = -0.25$, $SE = 0.66$, $z = -0.38$, $p = .98$) nor the strong-fit Spanish and the weak-fit English conditions ($\widehat{\beta} = 0.89$, $SE = 0.58$, $z = 1.53$, $p = .42$). Taken together, these comparisons show that all unexpected conditions (i.e. the weak-fit English and both Spanish conditions) were more challenging for bilingual listeners than the expected strong-fit English condition—and moreover, that all forms of unexpected words elicited the same magnitude of comprehension difficulty (as indexed by their equally-sized N400 responses between 300 and 500 ms).

For ease of comparison to prior work, we also conducted a set of exploratory analyses to see if there were any distributional differences across our N400 effects. To do this, we ran three separate mixed effects models: The first model compared strong-fit English words to both Spanish conditions (collapsing across contextual fit). The second model compared strong-fit English words to the weak-fit English words. The last model compared the three violation conditions to each other. All three models included distributional factors of *hemisphere* (left vs. right), *laterality* (lateral vs. medial), and *anteriority* (pre-frontal, frontal, centro-temporal, and occipital).[3] Results indicated that, when comparing strong-fit English conditions to both Spanish conditions (collapsing across contextual fit), there was a main effect of *language* ($\widehat{\beta} = -1.46$, $SE = 0.34$, $t = -4.24$, $p < .001$), confirming our prior findings. The only distributional factor that interacted with *language* in this model was *laterality* such that the N400 effect for Spanish words was more negative at medial electrode sites than at lateral ones ($\widehat{\beta} = -1.55$, $SE = 0.49$, $t = -3.18$, $p < .01$). There were no other significant two-way, three-way, or

[2] We also implemented these models with covariates for participants' proficiency levels in English and in Spanish. The pattern of significance did not change when controlling for proficiency differences in the subset of 22 participants that completed the LEAP Questionnaire. More information about these analyses can be found in our annotated analysis script on OSF (https://osf.io/jwqpr/).

[3] These distributional factors are originally from Moreno et al. (2002) and Ng et al. (2014). Both studies used these factors to argue for left-lateralization of their switch-related negativities (i.e. LAN effects). Specific details about which electrodes were used in each group can be found in our annotated analysis script on OSF (https://osf.io/jwqpr/).

four-way interactions. In the second model, when comparing strong-fit English to weak-fit English conditions, there was a main effect of *contextual fit* ($\widehat{\beta} = -1.41$, $SE = 0.39$, $t = -3.66$, $p < .001$), again confirming our prior results. Similar to the Spanish conditions, the only distributional factor that interacted with *contextual fit* was *laterality*, as the N400 effect was more negative at medial electrode sites than at lateral ones ($\widehat{\beta} = -1.92$, $SE = 0.54$, $t = -3.53$, $p < .001$). Again, there were no other significant two-way, three-way or four-way interactions. Finally, the direct comparison of all three unexpected conditions did not yield any significant differences, reaffirming that all three N400 effects had similar magnitudes and scalp distributions.

### 3.3. *Permutation-based cluster mass analysis across all electrodes: (exploratory)*

We conducted an exploratory permutation-based cluster mass analysis using all electrodes and all 200 time points between -5 ms to 1985 ms post-stimulus onset. We first report the results for the interaction in the main ANOVA and then the results from four pairwise comparisons. Extensive information about all of the results, the statistical procedure, and the raw output can be found on OSF (see https://osf.io/jwqpr/). First, our analysis revealed a significant cluster for the interaction that lasted between 355 and 535 ms (Summed *F*-statistic = 2125.00, $p < .05$). This cluster was broadly-distributed across centro-parietal electrode sites, and the effect was greatest at 385 ms over electrode FC1, which neighbors electrode Cz (see Fig. 2 for raster plots, waveforms, and scalp topographies). This significant interaction confirms the findings from our mean amplitude analyses, which revealed an interaction in the pre-registered 300-500 ms window such that the Spanish code-switched words and weak-fitting English words elicited similar N400 responses.

Given this pattern of effects, one might wonder whether there are any effects of code-switching that are independent of predicting a specific word. To explore this, we conducted a cluster analysis comparing the weak-fit English and the weak-fit Spanish conditions. Both are unpredicted words, but the latter involves a language shift. The analysis revealed a late positivity restricted to parietal electrodes between 885 and 1985 ms (Summed *F*-statistic = 9310.81, $p < .01$). This LPC effect peaked at 1025 ms over parietal electrode P4 (see Fig. 3). We take this LPC effect as evidence that bilinguals recognized that the speaker switched languages, i.e., a high-level (unexpected) discrepancy that needed to be re-evaluated (Friederici, 2005; Kaan et al., 2000; Kolk & Chwilla, 2007; Kuperberg et al., 2019; Litcofsky & Van Hell, 2017; Van Petten & Luka, 2012).

Similarly, one might ask whether there are any effects of contextual fit that are independent of predicting a specific word. We explored this by comparing the strong-fit Spanish and the weak-fit Spanish conditions. Both are unpredicted word forms, but the latter condition also involves a concept that is a poor fit for the context. This analysis revealed a broadly-distributed negativity lasting between 545 and 1265 ms (Summed *F*-statistic = 11,242.94, $p < .0125$). This sustained negativity peaked at 955 ms over the midline electrode Cz (see Fig. 4); however, the effect became more frontally-distributed toward the end of the cluster. Sustained negativities, especially those that are frontally-distributed, have been associated with increased working memory demands (e.g. Coulson & Kutas, 2001; Kutas & King, 1996), continued activity associated with word identification (Liao & Chan, 2016), and/or cognitive control processes (Lee & Federmeier, 2006, 2009, 2012; Nieuwland, Otten, & Van Berkum, 2007; Nieuwland & Van Berkum, 2006). We interpret this sustained negativity as reflecting increased or persisting difficulties with integrating a weak-fitting word into the unfolding context.

The remaining two pairwise comparisons involve conditions that differ along one dimension in our original experimental design but, by hypothesis, differ in terms of two cognitive processes each. In comparing the strong-fit English and the strong-fit Spanish conditions, we are comparing a word that is both predictable and in the matrix language to a word whose word form is unpredicted and involves code-switching. Thus, we might expect to see two effects: an early effect (i.e. the N400) reflecting the unexpected word (as in Fig. 2) and a late positive component reflecting the language shift (as in Fig. 3). The analysis revealed two significant clusters. The first cluster was a negativity distributed along the midline, which lasted between 235 and 595 ms (Summed *F*-statistic = 9431.80, $p < .0125$) and peaked at 435 ms over parietal electrode Pz (see Fig. 5). This early negativity reflects the N400 response captured in our previous analyses. The second cluster was a late positivity distributed across the parietal electrodes between 755 and 1305 ms (Summed *F*-statistic = 7472.399, $p < .0125$), which peaked at 965 ms over parietal electrode P3 (see Fig. 5). Again, the presence of the LPC seems to index recognition of the language switch, as these effects appeared in both Spanish code-switched conditions regardless of contextual fit.

In the last comparison, between the strong-fit English and the weak-fit English conditions, we are comparing a word that is predictable and easily integrated into the context with one that is unpredicted and hard to integrate. Thus, we might expect to first see an initial N400 effect reflecting the processing of the unexpected word (as in Figs. 2 and 5) followed by a later long-lasting negativity that begins toward the end of this time window (as in Fig. 4). This long-lasting negativity may reflect the continued difficulty of integrating weak-fitting words (regardless of language) into a broader discourse (see Liao & Chan, 2016 for similar effects). Because these two effects are adjacent in time, overlapping in space, and in the same voltage direction, they should be continuous in the cluster analysis, resulting in a single long-lasting cluster. Indeed, the analysis revealed a long-lasting negativity between 265 and 1195 ms (Summed *F*-statistic = 35,173.77, $p < .001$) that peaked at 515 ms over electrode CP1, which neighbors electrode Cz (see Fig. 6); however, the effect became more frontally-distributed toward the end of the cluster (similar to the sustained negativity in Fig. 4). We believe this long-lasting cluster reflects the summation of the (centro-parietal) N400 prediction effect and the sustained negativity observed in the other weak-fit condition.

## 4. General discussion

In this study, we tested whether switch-related ERP effects are better understood as direct costs associated with switching languages or as indirect consequences of processing unexpected words. To explore this, we factorially manipulated contextual fit and the presence of code-switching to explore bilinguals' responses to weak-fitting, within-language words and to strong and weak-fitting code-switched words. Given prior findings, we expected to find increased switch-related negativities for all three violation types and LPC effects for code-switches regardless of contextual fit (e.g. FitzPatrick & Indefrey, 2014; Liao & Chan, 2016; Van Hell et al., 2018, 2015). Using our novel Storytime paradigm, we successfully replicated these prior findings, as well as a lesser-known sustained negativity effect for weak-fitting words (e.g. Liao & Chan, 2016). Critically, we found that the N400 effect for double violations (i.e. weak-fitting, code-switched words) was equivalent to the effects for words that were either weak-fitting or code-switched. This pattern suggests that the N400 effects for code-switching are simply a specific case of lexico-semantic predictions being violated. In these rich contexts, listeners predict a particular word in a particular language. When that prediction is violated, lexical access and/or integration is more difficult, resulting in an increased N400 response. This cost is the same regardless of whether the prediction is violated due to the language, the meaning, or both. This pattern contrasted with the two later effects that we found: the LPC and a sustained negativity. The LPC was unique to the code-switching conditions and occurred regardless of contextual fit, and the sustained negativity was unique to words whose meaning did not quite match the context regardless of whether they were in English or Spanish.

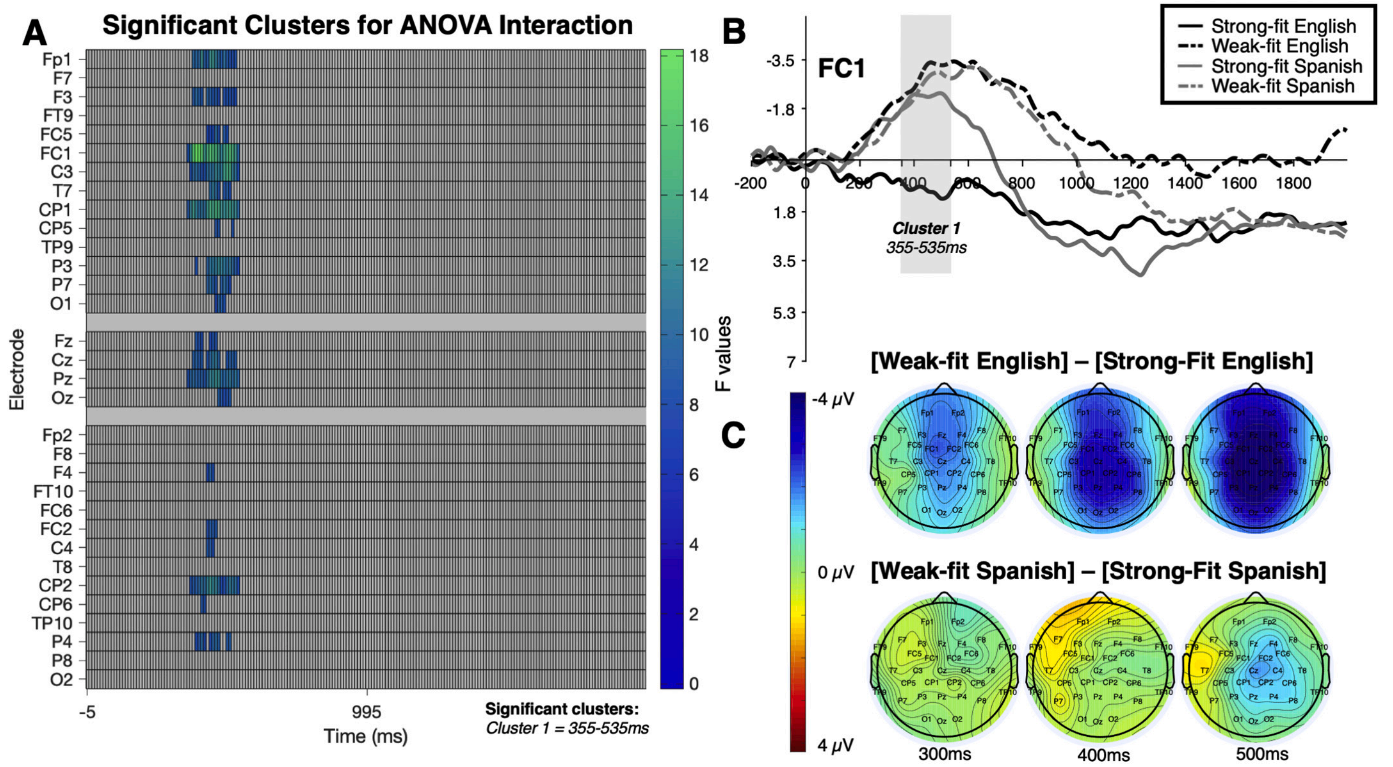

**Fig. 2.** *Cluster mass results from main ANOVA interaction.* This graphic indicates A) when and where the main interaction was significant in the trial (i.e. broadly-distributed effect between 355 and 535 ms); B) the waveform at electrode FC1, where the interaction effect was maximal; and C) the topographic maps of the difference waves for contextual fit between language conditions. All values in (B) and (C) are μVs. These plots show a significant interaction represented as typical N400 effects for all three violation types.

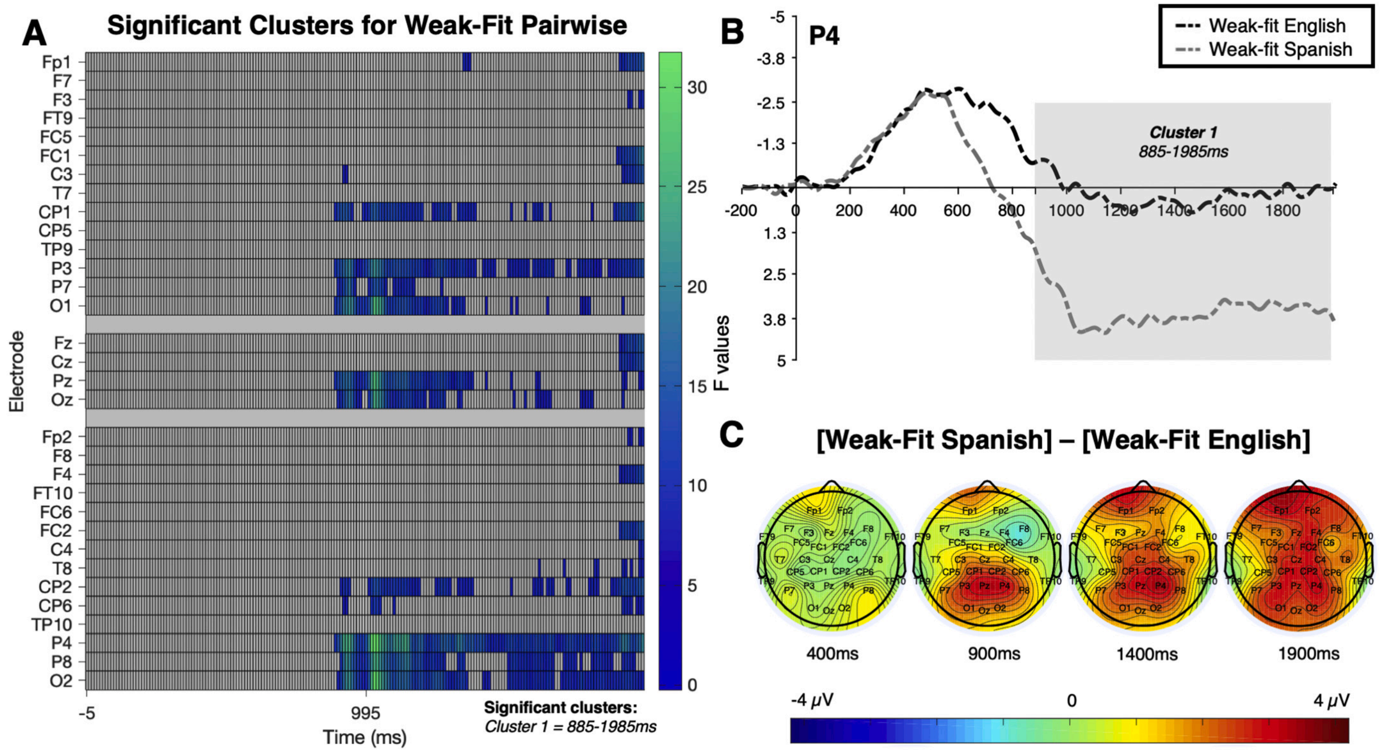

**Fig. 3.** *Cluster mass results from weak-fit comparison.* This graphic indicates A) when and where the weak-fit conditions differed significantly (i.e. posterior LPC effect between 885 and 1985 ms); B) the waveform at electrode P4, where the effect was maximal; and C) the topographic maps of the difference wave for language within weak-fitting conditions. All values in (B) and (C) are μVs. These plots show a late-emerging posterior positivity in response to code-switching for the weak-fitting conditions.

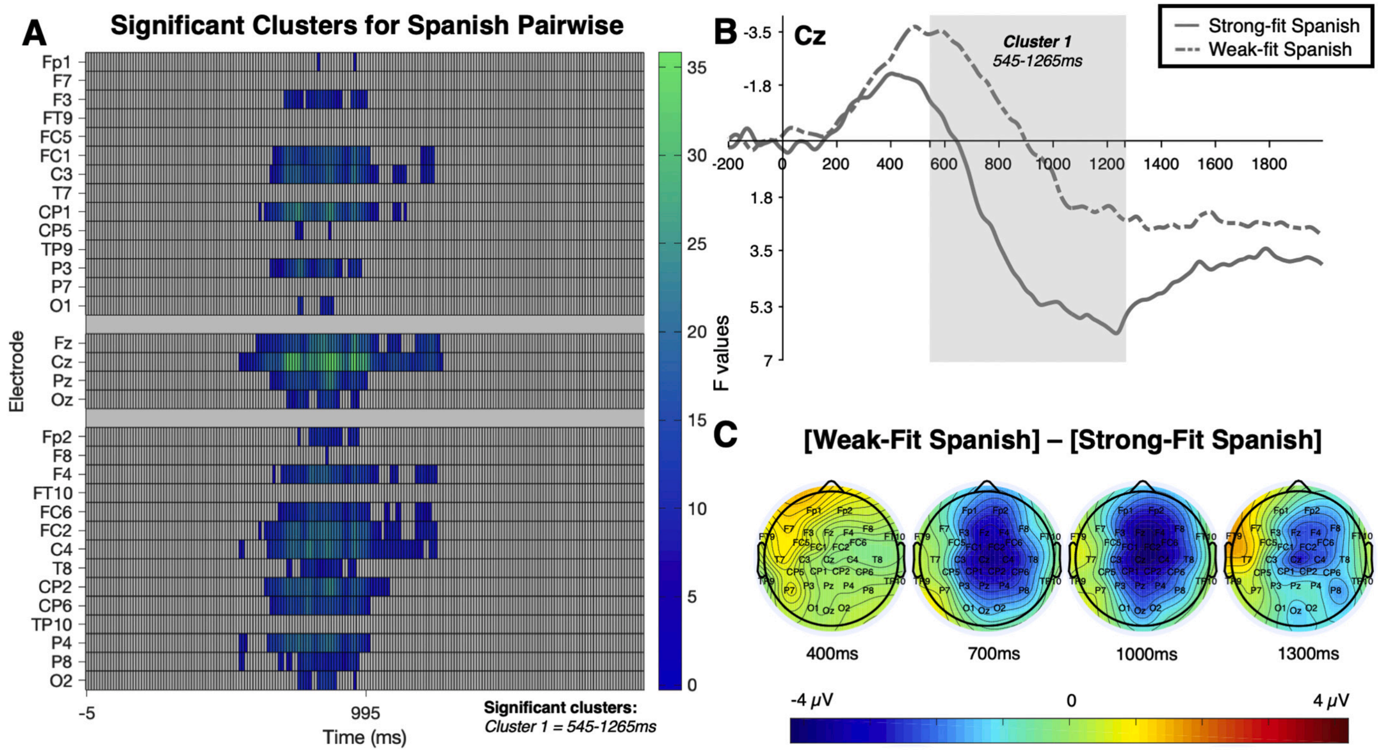

**Fig. 4.** *Cluster mass results from Spanish comparison.* This graphic indicates A) when and where the Spanish code-switched conditions differed significantly (i.e. sustained negativity between 545 and 1265 ms); B) the waveform at electrode Cz, where the effect was maximal; and C) the topographic maps of the difference wave for contextual fit within Spanish conditions. All values in (B) and (C) are μVs. These plots show a sustained negativity in response to weak-fitting words within Spanish conditions.

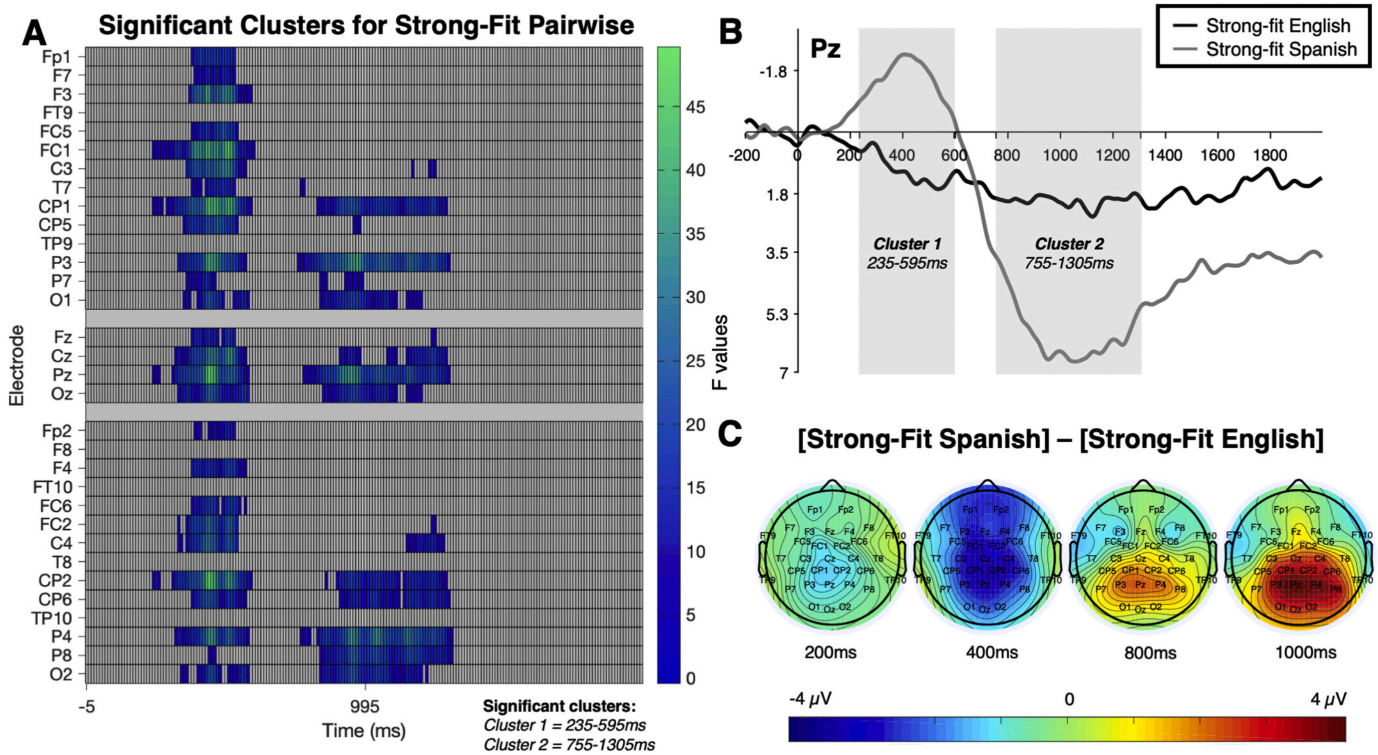

**Fig. 5.** *Cluster mass results from strong-fit comparison.* This graphic indicates A) when and where the strong-fit conditions differed significantly (i.e. early N400 effect between 235 and 595 ms and posterior LPC effect between 755 and 1305 ms); B) the waveform at electrode Pz, where the effects were maximal; and C) the topographic maps of the difference wave for language within strong-fit conditions. All values in (B) and (C) are μVs. These plots show an early N400 effect for the unexpected strong-fit Spanish word and a late-emerging posterior positivity reflecting the code-switch.

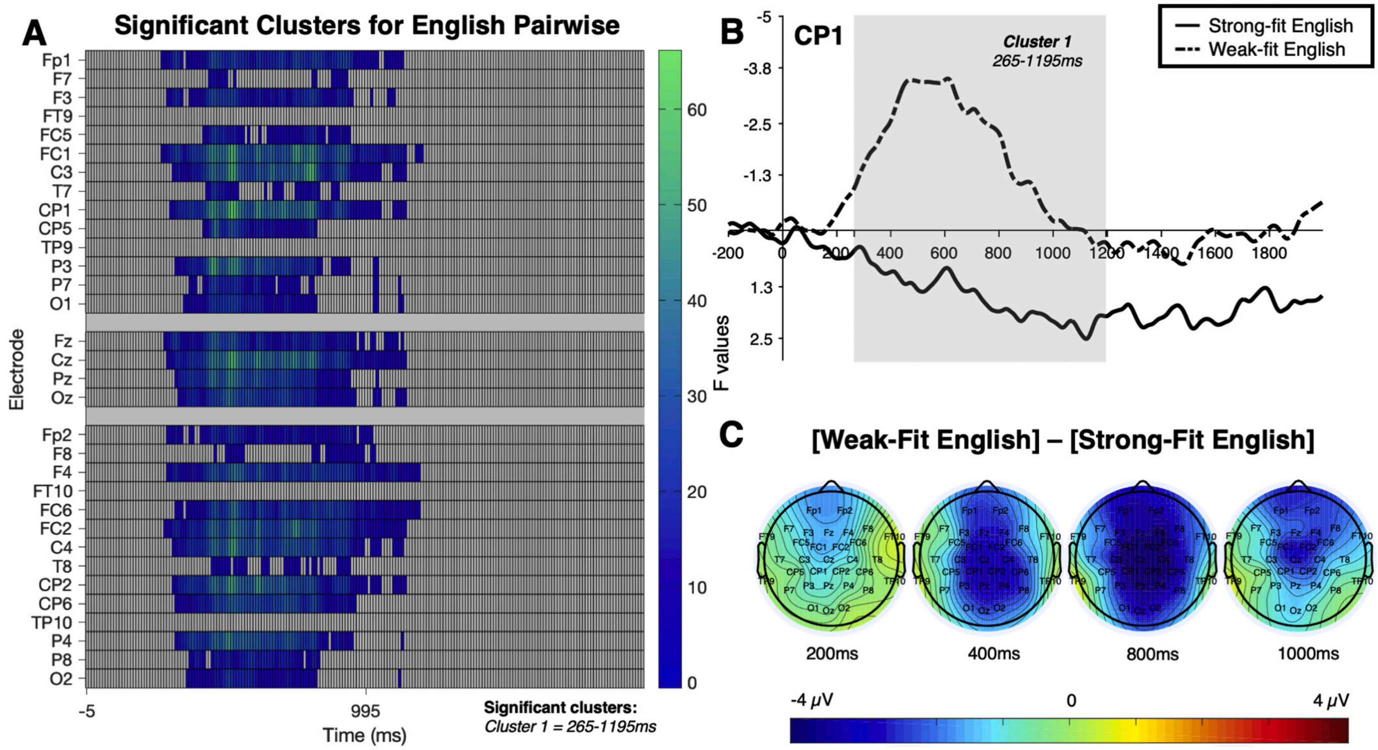

**Fig. 6.** *Cluster mass results from the English comparison.* This graphic indicates A) when and where the English conditions differed significantly (i.e. one long-lasting negativity between 265 and 1195 ms); B) the waveform at electrode CP1, where the effect was maximal; and C) the topographic maps of the difference wave for contextual fit within English conditions. All values in (B) and (C) are μVs. These plots show a long-lasting negativity, which we interpret as an overlapping N400 effect and a sustained negativity for the weak-fitting, unexpected English word.

We interpret the present findings as supporting the one-cost hypothesis presented in the Introduction. These findings demonstrate that early lexical processing can often occur prior to or independent of recognizing the language of the word being processed. Our results are consistent with a body of evidence showing that bilinguals can simultaneously map the sounds that they hear (or the letter/signs they see) onto lexical forms in both (or all) of the languages that they know (e.g. Dijkstra et al., 1999; Duyck et al., 2007; Van Hell & De Groot, 1998). Given this simultaneous mapping *and* a system that predicts lexical forms, we should expect that any unique cost of switching languages should occur after the initial costs of processing an unexpected word.

This data pattern is quite robust. As we mentioned in the Introduction, there are two other code-switching studies that used similar designs to our own and found similar data patterns (FitzPatrick & Indefrey, 2014; Liao & Chan, 2016). However, because these studies were designed with different questions in mind, they did not directly assess the predictions of the one-cost and two-cost hypotheses.

In the remainder of this General Discussion, we will address five issues: First, we will discuss the Liao and Chan (2016) and FitzPatrick and Indefrey (2014) studies in more detail, evaluating their interpretations of their findings in light of the present data (Section 4.1). Next, we discuss the prior code-switching studies that find LAN effects rather than canonical N400 effects, situating these data in the larger debate about the functional difference between them (Section 4.2). We then turn to prior studies that do not find *any* early negativities related to code-switching (Section 4.3). Given these findings, we then consider the implication of our results for theories about the functional significance of the N400 (Section 4.4). Finally, we integrate our findings into the prior literature on the LPC (Section 4.5), and end by describing the methodological contribution of this paper by examining the advantages and limitations of the Storytime paradigm (Section 4.6).

### *4.1. Previous factorial manipulations of contextual fit and language switching*

As we noted above, there are two other code-switching studies that used designs similar to ours and found very similar data patterns but interpreted them in different ways. Both Liao and Chan (2016) and FitzPatrick and Indefrey (2014) found the three basic ERP effects that were present in our study: an early interaction in a negative component (~250-450 ms), a longer-lasting negativity for all words that did not fit the context (collapsing across language), and an LPC effect for all code-switched words (collapsing across fit).[4] Despite the similarities across studies, the authors arrived at divergent conclusions about the costs of switching languages during comprehension. The main issues surrounding these alternative interpretations involve the conceptualization of the effects themselves and whether or not the authors posit a unique cost associated with code-switching during comprehension.

For example, Liao and Chan (2016) interpreted their early negativities as a variant of the PMN, which emerges after listeners hears a word with phonological features that mismatch the features of the word they expected to hear given the context (see Connolly et al., 1995; Connolly & Phillips, 1994). In their study, they argue that bilinguals were able to pre-activate information about the form of the upcoming words (and not just their semantic features). This interpretation is highly plausible given their design: participants listened to word-sized audio chunks presented one-by-one with 200 ms pauses in between them, and all of their target words were sentence-final. Thus, the slow and choppy presentation of their sentences, coupled with the fact that their targets always appeared

[4] Although, it is important to note that the LPC effects for the double violation conditions in Liao and Chan (2016) and FitzPatrick and Indefrey (2014) were heavily reduced (and sometimes non-significant) due to the overlapping sustained negativities that were also present in this particular condition.

in the same sentence position, could have allowed for robust prediction. Critically, under their interpretation, there is no unique cost to code-switching—rather, there is just one cost associated with perceiving an unexpected sound, and that cost is similar across all violation conditions. Thus, their theoretical conclusions are broadly compatible with ours; the outstanding issue is whether their PMN effects can be interpreted as reflecting an identical (or at least similar) set of underlying processes as our N400 effects. There are three reasons to think this might be the case: First, as the authors note, their PMN did not have the canonical frontal distribution of a typical PMN but instead had a distribution more similar to an N400. Second, their slower presentation method and the predictable position of their violations may have speeded up processing, shifting the N400 effect to a slightly earlier time window (see Kutas, Van Petten, & Kluender, 2006; Brothers et al., 2015 for related discussions). Lastly, as we will discuss below in Section 4.2.2, there are good reasons to believe that all language-related negativities come from the same functional family and vary continuously rather than categorically.

In contrast, FitzPatrick and Indefrey (2014) conceptualize their results in a very different way: They argue that the N400 effect for the strong-fitting code-switches reflects the initial unavailability of the meaning of the code-switched word. They argue that the cost of code-switching does not manifest itself as a greater amplitude on the N400 response but rather as a delay in lexical access that results in a transient negativity while lexical meaning is unavailable. On this account, weak-fitting code-switched words do not show greater N400 amplitudes than strong-fitting code-switched words because, in both cases, the meaning is not initially accessed.

To address these competing hypotheses, we can look at the predictions that each account makes for code-switched words in various sentence contexts. According to FitzPatrick and Indefrey (2014), the meaning of code-switched words should always be delayed. Thus, we should see the initial N400 effect (the transient negativity) for both code-switches regardless of whether the target word is predictable or not. According to our proposal (the one-cost account), the initial N400 effect reflects the violation of a lexical prediction, not the evaluation of meaning. Thus, we should only see this pattern when the word is predictable. In contrast, the later sustained negativity reflects the degree to which the meaning of the word matches or mismatches the prior context (regardless of how predictable the word was or whether it was code-switched or not).

In order to test these predictions, we would want to look at sentences with unpredictable target words, as this condition provides the most robust contrast between the hypotheses: On the delayed access account, strong-fitting code-switches should continue to show an initial negativity (relative to same language continuations) because their meaning is always initially unavailable. Moreover, the magnitude of this negativity should not be influenced by the constraint of the sentences. In contrast, our one-cost account predicts that strong-fitting code-switches should not show any early negativity, because the lexical form of the target word cannot be predicted in unpredictable contexts. This finding would indicate that, in the unpredictable contexts, the semantic features of the input (regardless of language) were interpreted in a bottom-up fashion with all strong-fitting words showing smaller N400 responses than all weak-fitting words. This account would also predict that the only index of recognizing the language switch would occur later as an LPC effect.

In our exploratory analyses, we compared the ERP effects from trials with very high cloze values and very low cloze values. Specifically, we analyzed the top 15% of trials ($M = 91\%$, Range = 88.5–94.3%) and the lowest 15% of trials ($M = 17\%$, Range = 2.9–34.3%). In the highest cloze group, we saw the same pattern that we found in the primary analysis: a very large early N400 effect that was similar in size across the three violation types, a later sustained negativity for weak-fitting words, and an LPC for code-switched words (see Fig. 7 below). In the lowest cloze group, there was little to no early N400 effect, but there was still a sustained negativity and an LPC (for similar findings using eye-tracking methods in low constraint sentences, see Altarriba, Kroll, Sholl, & Rayner, 1996; cf. Hoversten & Traxler, 2020).

We confirmed these findings with a post-hoc mixed effects model that looked at average N400 amplitudes from centroparietal electrodes between 300 and 600 ms. We included random intercepts for participants and items, and fixed effects of *language* (English = 0, Spanish = 1), *contextual fit* (strong-fit = 0, weak-fit = 1), *cloze* (highest cloze = 0, lowest cloze = 1), and all of their interactions. We found a significant three-way interaction ($\widehat{\beta} = -1.04$, $SE = 0.51$, $t = -2.04$, $p < .05$), which suggested that the lower two-way interaction of *contextual fit* and *language* differed across cloze groups between 300 and 600 ms. Pairwise comparisons revealed that this effect was due to the fact that the two code-switched conditions significantly differed in the lowest cloze items ($\widehat{\beta} = -1.12$, $SE = 0.25$, $z = 4.51$, $p < .001$) but not in the highest cloze items ($\widehat{\beta} = 0.46$, $SE = 0.26$, $z = 1.75$, $p = .08$). All other pairwise comparisons (within each cloze group) were significant (see our

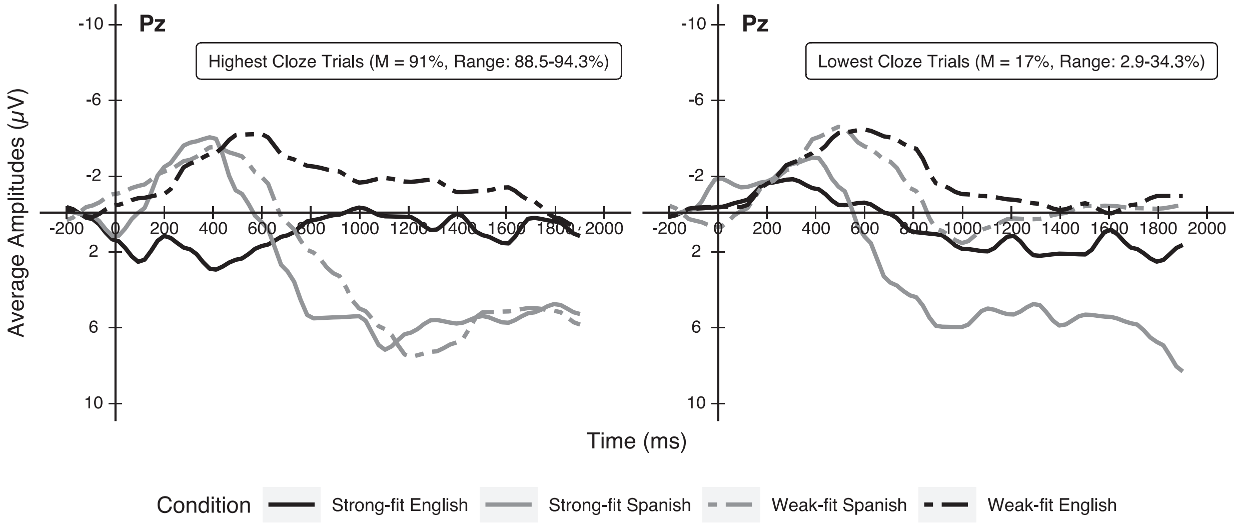

**Fig. 7.** *Average waveforms at Pz across highest and lowest cloze items.* These waveforms were recreated for plotting purposes by taking the average amplitude values every 100 ms from −200 to 2000 ms (e.g. 100-200 ms, 500-600 ms, 1200-1300 ms). The lines were then fit to these averages using local regression (loess) smoothing techniques in R (see our annotated code on OSF, https://osf.io/jwqpr/). The two dark and lighter lines represent English words and Spanish code-switches respectively. The solid and the dotted lines represent strong and weak-fitting conditions respectively.

analyses on OSF for the full model summaries). These findings suggest that the meanings of code-switched words are initially available, and that lexical access is not delayed, resulting in the early differentiation between the two code-switched conditions in the unpredictable contexts.

There are two other interesting findings to pull from these exploratory analyses: First, the LPC effects for the weak-fitting code-switches were heavily reduced in the lowest cloze trials—a similar pattern was reported in FitzPatrick and Indefrey (2014). Second, the N400 response for the baseline condition (i.e. the strong-fitting English words) increased in the lowest cloze group, confirming that the prediction of the target word was not as robust in these less predictable trials. In fact, we found a small correlation between the size of the N400 response for baseline conditions and the predictability of the target word (see Fig. 8). As the cloze probability of the target word increases, the N400 response for that word becomes less negative, $r(118) = 0.18$, $p < .05$. Taken together, these exploratory findings suggest that the meanings of code-switched words are accessible during the initial stages of lexical processing—and moreover, that the variability in the magnitude of this early negativity heavily depends on how much the listener expected to hear a particular word in a particular language.

### 4.2. Understanding variability in switch-related negativities

In the Introduction, we noted that most ERP studies on code-switching find a biphasic pattern consisting of an early negativity followed by later positivity in response to code-switched words in sentence contexts. However, we also noted that there is considerable variation in the timing and scalp distribution of these switch-related effects. In this section, we focus on understanding this variability. Researchers have used many different labels for the switch-related negativities in their studies—for example, switch-related PMNs, N1s, N200s, N250s, N400s, and LANs have all been reported (see Kutas, Moreno, & Wicha, 2009; Payne, Ng, Shantz, & Federmeier, 2020 for an overview). These effects vary in both their timing and their distribution, ranging from effects that are localized to left anterior or fronto-central electrode sites to effects that spread widely across the scalp (see Moreno et al., 2008; Litcofsky & Van Hell, 2017 for reviews). Nevertheless, there is a family resemblance between them—namely, they are all negativities that take place, at least in part, during the typical N400 time window (~200-600 ms) and overlap in distribution with a typical N400 (i.e. widespread with a centroparietal focus).

The present study found a switch-related negativity with a classic N400 distribution and timing—thus, we interpreted the effect in line with prior treatments of the N400 in the broader psycholinguistic literature. Specifically, we treated it as an index of the difficulty of lexico-semantic prediction (e.g. Federmeier, 2007; Kutas et al., 2006; Kuperberg et al., 2019; Kutas & Federmeier, 2011; Lau et al., 2013; Otten & Van Berkum, 2008; Van Berkum et al., 2005; Wlotko & Federmeier, 2015 for a review). While many other studies have found switch-related components that look like classic N400s, switch-related effects that look like LANs are also common. In the broader psycholinguistic literature, LAN effects are often associated with an increased demand on working memory (King & Kutas, 1995; Kluender & Kutas, 1993) and/or difficulties with morphosyntactic processing (e.g. Friederici, 2002; Gunter et al., 2000; Neville et al., 1991).

In reading the code-switching literature, we found several approaches for dealing with this variability. One approach is to treat these two effects as functionally distinct, with each component reflecting a different process. On this approach, switch-related LANs are argued to index difficulties associated with integrating two language systems with different morphosyntactic features (e.g. Moreno et al., 2002; Ng et al., 2014) whereas switch-related N400s reflect difficulty in accessing the meaning of the code-switch and integrating it into the prior context (e.g. Proverbio et al., 2004; Ruigendijk et al., 2016). A second approach is to avoid making explicit commitments about the nature of these components. In practice, this often means not making a strong prediction about which effect will occur in a given study and conducting an analysis that should capture either effect if it is present (e.g. Kaan et al., 2020). A final approach is to simply note that switch-related effects vary in their distribution and often do not have the canonical N400 morphology but then treat the effects as being functionally equivalent to the N400 (see Van Hell et al., 2015; Van Hell & Witteman, 2009).

The present study was framed with the working hypothesis that LANs and N400s reflect a common underlying process (or set of processes). We chose this framing for the sake of simplicity and clarity, but also because we believe that there is compelling evidence that all language-related negativities within this time window belong to the same functional family of *mismatch negativities* (see Bornkessel-Schlesewsky & Schlesewsky, 2019 for parallel discussion). Nothing in the design of our study depended on this working hypothesis; as we noted above, many researchers have used these measures without making this theoretical commitment. However, a full interpretation of our findings and of the prior literature requires that we revisit this hypothesis. We begin by laying out the case for functionally distinct LANs and N400s and then evaluating the argument in light of the data (Section 4.2.1). Next, we make the case for the unitary nature of switch-related negativities, examining how this theory would account for the observed differences across studies (Section 4.2.2). Finally, we address an alternative theory that LAN components are an epiphenomenon resulting from instances of component overlap (Section 4.2.3).

#### 4.2.1. The LAN and N400 as functionally distinct components

There is a long tradition in the psycholinguistic literature of treating LANs and N400s as functionally distinct components, with the LAN indexing morphosyntactic processes and the N400 indexing lexico-semantic processes (see Caffarra et al., 2019 for discussion). Thus, it is unsurprising that these two effects are often treated as distinct in the

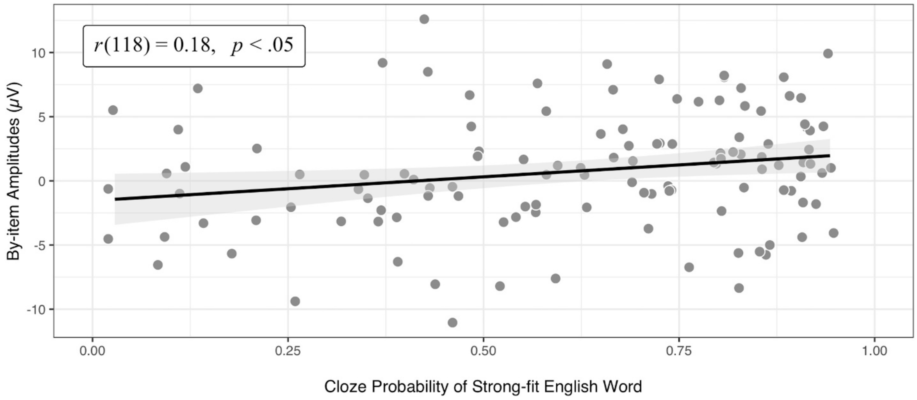

**Fig. 8.** *Average N400 response amplitudes across target predictability.* Each observation represents the by-item average amplitude between 300 and 600 ms from midline electrodes Fz, Cz, and Pz for strong-fitting English words. The x-axis is plotting the cloze probability values obtained from our audio cloze ratings task. As the predictability of the target word increases, the N400 response to that word is reduced (i.e. it becomes more positive). There is a small correlation between response amplitude and predictability, $r(118) = 0.18$, $p < .05$.

code-switching literature as well (see Litcofsky & Van Hell, 2017). If these two effects are functionally distinct, then we should expect to find them in different populations or under different conditions. There are two sets of findings in the code-switching literature that provide prima facie support for this claim, but this evidence is limited and open to alternative interpretations.

First, there are studies that find the switch-related LANs and N400s in different populations. For example, Van Der Meij et al. (2011) recruited native Spanish speakers with either high or low proficiency in English and asked them to read English sentences. Half of these sentences had a word code-switched into Spanish. As predicted, the authors found a biphasic pattern in response to code-switched words. However, there were distributional differences in the switch-related negativities based on speakers' proficiency in English: For low proficiency speakers, there was a canonical N400 effect, which the authors took as evidence that accessing words from the non-matrix language incurred additional processing costs. For high proficiency speakers, there was a widespread negativity that extended to left frontal electrode sites, which the authors suggested may be more akin to the LAN effects observed in balanced bilingual populations (e.g. Moreno et al., 2002; Ng et al., 2014). This distributional difference was taken as tentative evidence that highly proficient L2 speakers (and balanced bilinguals) are more influenced by the grammar of their second language than less proficient speakers, resulting in more difficulty integrating codeswitches with the matrix language.

This hypothesis, however, has not stood up well to further scrutiny. In a very similar study with Finnish-English bilinguals, Hut and Leminen (2017) found essentially the opposite pattern —a widespread negativity that extended to left anterior electrodes in their less proficient group and a canonical N400 effect in their more proficient group. Two other studies comparing bilinguals with varying levels of proficiency found canonical N400 effects with no topographical differences between the groups (e.g. Proverbio et al., 2004; Ruigendijk et al., 2016).

The second way to demonstrate a functional dissociation between switch-related LANs and N400s would be to identify stimulus factors that influence one component but not the other. Ng et al. (2014) report one analysis of this kind. They presented short stories (averaging four sentences in length) to Spanish-English bilinguals. These stories contained four instances in which target words were code-switched into Spanish. The authors report a left-lateralized negativity in response to these code-switched words, which they interpreted as a LAN effect. This left-lateralized effect was unexpected; the authors intended to explore how the size of the switch-related *N400 effect* was influenced by the position of the word in the story. N400 effects are typically smaller for words that appear later in a sentence, consistent with standard accounts linking N400 effects to lexical access and integration (see Kambe, Rayner, & Duffy, 2001; Federmeier, 2007; Van Petten & Kutas, 1990, 1991; cf. Van Petten, 1995 for connected discourse). They found that this switch-related negativity was not modulated by the position of the code-switched word in the story, as the amplitude of the effect was no different for the first two switches than the last two. However, when collapsing across switched and non-switched words, they found an N400 effect that was modulated by discourse position. The authors concluded that their switch-related negativity was a LAN and that it was functionally distinct from the N400 in their study.

Critically, Ng and colleagues' analysis rests on a null interaction in a study that may not have sufficient power to detect an effect of the relevant size. The evidence that the N400 is sensitive to their discourse manipulation is complex: when collapsing across switches and non-switches, there was a small *word position* by *word class* interaction due to the fact that nouns in either language showed smaller N400s later in the discourse, while verbs did not. Thus, the critical evidence for a functional dissociation would be a modulation of this effect—namely, a three-way interaction between *word class*, *word position*, and *switching* such that the difference between code-switched nouns and non-switched nouns would be greater at the beginning of the story than at the end. It is unclear how large we would expect such a modulation to be, but presumably it would be smaller than the two-way interaction itself (since no crossover is predicted). Given that their two-way interaction was just within the standard limits of significance ($p = .04$), it seems likely that a fairly large modulation could be missed with this design.

Future studies could address this issue with larger samples or more powerful manipulations of context. Word position effects on N400 magnitudes are thought to be a side effect of predictability, i.e., the more context that precedes a target word, the stronger the lexical prediction for that word may become (e.g. Payne, Lee, & Federmeier, 2015; Van Petten & Kutas, 1991; Van Petten & Luka, 2012). Thus, a more direct test of this functional distinction claim would be to test for differences in code-switching effects when the original target word is predictable or unpredictable. In Section 4.1, we presented evidence from an exploratory analysis of this kind using our own data: we found that the magnitude of the N400 response for strong-fitting code-switched words was reduced (relative to other violation conditions) when the target words were less predictable (see Fig. 7). Moreover, we would also predict that the size of the N400 effects from both violation types (i.e. code-switched words vs. weak-fit words) should vary continuously across cloze probability. To test this, we conducted another set of post-hoc mixed models: The first model directly compared strong-fit English words to both Spanish conditions. We found a significant interaction between *cloze probability* and *language* ($\widehat{\beta} = -2.86$, $SE = 1.12$, $t = -2.55$, $p < .05$), suggesting that the magnitude of the switch-related N400 effect (collapsing across contextual fit) increased as the predictability of the target word increased. In the second model, we directly compared the two English conditions and found a significant interaction between *cloze probability* and *contextual fit* ($\widehat{\beta} = -4.58$, $SE = 1.27$, $t = -3.60$, $p < .001$), suggesting that the N400 effects for weak-fitting, non-switched words also increased alongside predictability. Below, we summarize these findings by plotting the N400 effect sizes for each violation condition across cloze values (see Fig. 9).

In sum, the prior evidence for a functional distinction between switch-related LANs and switch-related N400s is quite weak. There is no known set of factors that will reliably produce switch-related LANs as opposed to N400s. This problem extends to the broader psycholinguistic literature. In studies that are intended to elicit LANs, the observed effect is often not left-lateralized (e.g. Foucart & Frenck-Mestre, 2011, 2012; Hasting & Kotz, 2008; Lau, Stroud, Plesch, & Phillips, 2006; Nevins, Dillon, Malhotra, & Phillips, 2007; Osterhout & Mobley, 1995; Tokowicz & MacWhinney, 2005) and sometimes has the morphology of a canonical N400 instead (Courteau, Martignetti, Royle, & Steinhauer, 2019; Fromont et al., 2020; Guajardo & Wicha, 2014; Nieuwland, Martin, & Carreiras, 2013; Severens, Jansma, & Hartsuiker, 2008; Wicha et al., 2004).

A further reason to reject this hypothesis is that it fails to account for other ways in which switch-related negativities vary in their timing and distribution (e.g. bilateral ANs, Litcofsky & Van Hell, 2017 on second code-switch; Zeller, 2020; N1 effects, Proverbio et al., 2002, 2004; N200 effects, Khamis-Dakwar & Froud, 2007; left-occipital N250s, Van Der Meij et al., 2011; fronto-central negativities, Hut & Leminen, 2017; broad negativities, Zeller, 2020; PMNs, Liao & Chan, 2016; see Kutas et al., 2009; Payne et al., 2020 for reviews). Thus, it seems unlikely that code-switches in sentence contexts produce 4 or 5 categorically distinct effects that are elicited by differences across studies that we do not yet understand.

### 4.2.2. The LAN and the N400 as a unitary phenomenon

The second hypothesis regarding LANs and N400s is that both negativities reflect the same set of underlying processes—and thus belong to the same functional family, despite their differences in timing and scalp distribution. This functional family is often referred to as *mismatch negativities* (MMNs) because they are thought to reflect the degree to which top-down expectations mismatch the incoming sensory input. In a

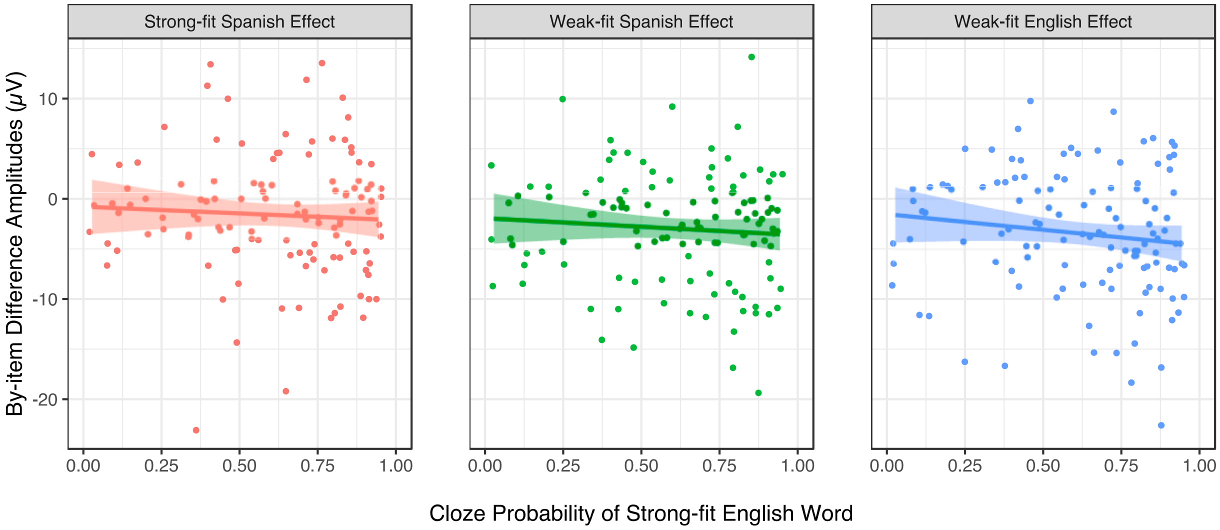

**Fig. 9.** *Average N400 effect amplitudes across target predictability.* Each observation represents the average N400 effect amplitude (violation – baseline) between 300 and 600 ms from midline electrodes Fz, Cz, and Pz. The x-axis is plotting the cloze probability values obtained from our audio cloze task. For each violation type, the N400 effect size becomes more negative as the predictability of the expected word (i.e. the strong-fitting English word) increases.

recent proposal, Bornkessel-Schlesewsky and Schlesewsky (2019) argue that all language-related negativities can be understood as MMNs. They argue that differences in the timing and distribution of these negativities arise from differences in the specific stimuli that cause the mismatch—namely, differences in stimulus complexity, the window of temporal integration needed to detect the mismatch, and the nature of the representation that fails to match some top-down expectation. For example, the authors propose that ERP effects that are more LAN-like may arise when the top-down prediction involves more morphemic (rather than lexical) representations. But critically, all of these negativities (e.g. N400, LANs, ELANs, PMNs) reflect the same basic construct: *precision weighted prediction error* (i.e. an error signal that is inversely related to the uncertainty before the critical word, see Molinaro et al., 2011, for a related proposal, and Moreno et al., 2008 and Zeller, 2020 for discussions of how these same principles may apply to code-switching).

This proposal from Bornkessel-Schlesewsky and Schlesewsky (2019) is flexible enough to explain all of the existing data: it can account for both the LAN-like and N400-like effects; it can explain why these effects arise in response to code-switching without needing to posit two mutually-exclusive error detection processes; it can allow for intermediate data patterns; and it can encompass the host of other negativities that have been observed in code-switching paradigms (see list above). However, the flexibility of this proposal is also its greatest weakness—without a host of auxiliary hypotheses, it fails to predict when each pattern should occur. This weakness, however, may simply reflect the limits of our current knowledge. As we noted above, we do not yet have a set of factors that reliably give rise to these subtle differences in switch-related negativities.

#### 4.2.3. *The LAN as an epiphenomenon resulting from component overlap*

There is a final hypothesis about the relationship between LANs and N400s. Some researchers argue that the LANs in most biphasic patterns may be epiphenomenal, arising in circumstances in which negative and positive components (typically N400s and LPCs) are both present and cancel out one another due to their temporal and spatial co-occurrence (see Osterhout, 1997; Osterhout, McLaughlin, Kim, Greenwald, & Inoue, 2004; Guajardo & Wicha, 2014; Tanner & Van Hell, 2014; Tanner, 2015 for discussion; but also see Caffarra et al., 2019). In spoken language studies, the N400 is typically broadly distributed, not strongly lateralized, and emerges at centro-parietal electrode sites from 200 to 600 ms (Kutas & Federmeier, 2011). The LPC is often posteriorly distributed, can be right-skewed, and emerges at parietal electrode sites between 500 and 800 ms (Kuperberg et al., 2019; Tanner & Van Hell, 2014; Van Petten & Luka, 2012). Thus, the N400 response often begins prior to the LPC and can overlap in time and space with these late positivities.

In the ERP literature, component overlap may occur for various reasons: First, a single individual may generate both an N400 and an LPC in response to the same stimulus (e.g. a code-switched word). When both components are laid on top of each other, what remains is the left-most portion of the N400 in the earlier time windows (i.e. the LAN) and a positivity in the later windows at posterior electrode sites (see Osterhout & Mobley, 1995; Tanner, 2015). Second, component overlap may be artificially created when researchers average the ERP data across individual participants—a procedure that is standard in most ERP studies (Tanner & Van Hell, 2014). At the individual-level, ERPs rarely show a true biphasic pattern with equally robust negativities and positivities (Tanner & Van Hell, 2014; Tanner, 2015; although, see Caffarra et al., 2019). Individual participants' data are often more negative (i.e. large early negativities with small late positivities) or more positive (i.e. small early negativities with large late positivities; see Osterhout et al., 2004; Tanner & Van Hell, 2014). Thus, when the study population contains both types of individuals, the averaging procedure can create a strong biphasic pattern.

This epiphenomenal hypothesis makes a few predictions about switch-related negativities: First, we expect that no study should find switch-related LAN effects without LPCs. To the best of our knowledge, this prediction holds true, as no code-switching study to date reports pure LAN effects and no late positivities. In the broader psycholinguistic literature, there are only a handful of studies that show pure LAN effects without late positivities (e.g. Coulson et al., 1998; Kim & Sikos, 2011; O'Rourke & Van Petten, 2011)—however, this pattern remains relatively rare. Second, in studies without LPC effects, we expect to find canonical N400 effects. In two prominent code-switching studies without LPCs, for example, the authors indeed report switch-related negativities with the canonical latency and distribution of N400 effects (see Fernandez et al., 2019; Proverbio et al., 2004).

Finally, we expect that the studies, experimental conditions, and individual participants that show larger LPCs should be more likely to have LANs and not N400s. There are two studies that find left-lateralized negativities in their most proficient L2 speakers but classic N400s in less proficient speakers. In both cases, the more proficient speakers also showed larger LPC responses to code-switched words, suggesting that the size of the LPC may contribute to the degree of lateralization

(Ruigendijk et al., 2016; Van Der Meij et al., 2011). Tellingly, in Hut and Leminen (2017), both aspects of this pattern are reversed: less proficient speakers show greater left lateralization and a trend toward a larger LPC (though this interaction is not significant). This pattern suggests that the two components move in unison, just as this last hypothesis would predict.

Future studies could more directly assess how the presence (and strength) of late-arriving positivities influence the scalp distribution of earlier negativities. Many psycholinguistic studies have successfully suppressed late positivities over the course of an experiment by increasing the amount of surprising material—essentially, reducing the novelty of the violations (e.g. Hahne & Friederici, 1999). Thus, if one were to increase the number (or predictability) of the code-switches in a study, the LPC response should weaken over the course of the experiment. One could then observe how the distribution of switch-related negativities is affected by the strength of the LPC effects. Moreover, this study has the added benefit of evaluating switch-related LANs and N400s within the same individual, controlling for proficiency and the type of information that they may prioritize when processing the code-switches.

In sum, while there is considerable research to be done, we feel that the evidence to date does not provide strong support for the hypothesis that the LAN and the N400 are functionally distinct components. Instead, the data suggest that they reflect similar underlying processes with the differences in distribution reflecting either differences in the stimuli that give rise to the mismatch (Section 4.2.2) or the effect of overlapping components on the observed morphology (Section 4.2.3). Thus, we will continue to discuss switch-related negativities as instances of N400s and to interpret them in light of the extensive work on the functional nature of this component.

### *4.3. When should we expect reduced or absent switch-related N400 effects?*

On our hypothesis, the one-cost account, the larger N400s to code-switched words occur because comprehenders have predicted that they will encounter a particular word (or one of a small set number of candidate words), and that expectation is violated after encountering the translation equivalent instead. On this theory, the N400 effect is not driven by an increased N400 response to violations but rather by a decreased N400 to words that are expected due to the pre-activation of lexical and semantic information and the subsequent reduction in the processing load (e.g. Federmeier, 2007; Kuperberg et al., 2019; Kutas et al., 2006; Kutas & Federmeier, 2011; Lau et al., 2013; Van Berkum et al., 2005). A challenge for our hypothesis is explaining why some code-switching studies fail to find an N400 effect or any switch-related negativity (e.g. Moreno et al., 2002 with idioms; Ruigendijk et al., 2016 with intermediate-level speakers; Litcofsky & Van Hell, 2017). On the one-cost account, there are two obvious explanations for the lack of N400 effects for code-switching: 1) the effect could be absent because the comprehender is failing to predict the upcoming word (or its features) in the matrix language; or 2) the effect could be absent because the comprehender is predicting both the expected word *and* the code-switched word to a similar degree. These two explanations seem to account for most (if not all) of the missing N400 effects.

The most common type of missing or reduced N400 effects occurs in studies where the matrix language is less familiar to the participants than the code-switched language. For example, Ruigendijk et al. (2016) investigated the comprehension of sentence-final German-to-Russian code-switches in spoken sentences. They recruited German monolinguals (no knowledge of Russian) and Russian speakers with either high or intermediate German proficiency. Both groups with high proficiency in German (i.e. the matrix language) showed N400 effects for code-switching into Russian. The intermediate group showed no difference in their N400 responses to the German and Russian targets—in fact, the authors argue that the intermediate group showed large N400 responses to *both* the expected and code-switched conditions, suggesting a lack of pre-activation for any of the target words (see similar findings in our exploratory analyses above). This is consistent with our first explanation for missing N400 effects. Similarly, Liao and Chan (2016) only find switch-related N400 effects when switching from participants' dominant language into their weaker language. Finally, Van Der Meij et al. (2011) report delayed, less robust N400 effects in less proficient speakers. In semantic violation paradigms, participants also typically show weaker N400 effects in their second language than in their first language (see Ito, 2016; Ito et al., 2017; Ito et al., 2018; Martin et al., 2013; Van Hell & Tanner, 2012; Weber-Fox & Neville, 1996). Thus, the most parsimonious explanation of these data patterns is that the ability to predict words on-the-fly largely depends on fluency in the matrix language (e.g. Ito, 2016; Ito et al., 2017; Ito et al., 2018; Kotz & Elston-Güttler, 2004).

Similar to being unable to use an unfamiliar language to make predictions, we can also have a challenging paradigm that makes it difficult to predict upcoming words. For example, in a study by Litcofsky and Van Hell (2017), highly-proficient Spanish-English bilinguals read sentences (word-by-word) that switched mid-sentence from one language to another. In this study, some sentences began in English, some began in Spanish, and half of the sentences contained code-switches. The authors did not find any differences between the N400 responses to switched and non-switched words in either switching direction. We believe that the lack of an N400 effect (in either switching direction) reflects the fact that predicting *any* word in this study was difficult for participants. Support for this claim comes from the fact that both code-switched words *and* non-switched words elicited robust N400 responses (on the order of 2 μV). This data pattern resembles the one from Ruigendijk et al. (2016) referenced above—i.e., when their intermediate L2 speakers could not predict in the matrix languages, there were robust N400 responses for their code-switched words and their baseline controls. The present study also finds increased N400 responses to our baseline controls when the target words are not easily predicted (see Figs. 7 and 8).

The challenge, however, for this theory is that Litcofsky and Van Hell (2017) had highly proficient speakers and used sentences that do not seem highly unpredictable (although, they did not assess the predictability of their materials). For this reason, we believe that prediction has broken down because of the paradigm itself, perhaps due to their use of Rapid Serial Visual Presentation (RSVP) and/or their fast presentation rate (300 ms per word, 500 ms SOA). If the paradigm has made prediction more difficult, then we would expect that using the same stimuli with a more naturalistic presentation (i.e. auditory presentation where each word is presented at a speed that is correlated to the word's length) would allow prediction to occur more easily. Evidence in favor of this prediction comes from a study by Fernandez et al. (2019) who adapted the study by Litcofsky and Van Hell (2017) using naturalistic auditory presentation and the same set of sentences. In contrast to the prior study, the authors find N400 effects for code-switched words in *both switching directions*, suggesting that prediction was enhanced in this more naturalistic presentation method. Moreover, the N400 response amplitudes for their baseline conditions appear to be reduced; however, this point remains speculative, as we should not readily compare N400 response amplitudes across written and spoken modalities.

Our second explanation for missing N400 effects is that, under some circumstances, bilinguals may predict the code-switched form in addition to the matrix form. Logically, this should happen most often when the location of the switch is highly predictable. The clearest example of such an effect is a study by Moreno et al. (2002), which presented a mix of regular and idiomatic sentences to English-Spanish bilinguals. In each sentence, the last word was manipulated to be the expected, within-language word, its translation equivalent, or an unexpected within-language word (e.g. “Out of sight, out of mind/brain/*mente* [mind].”). In regular sentences, the authors observed the typical N400 effect for the unexpected, within-language word and a biphasic pattern for the code-switched word consisting of a left-lateralized negativity (250-450 ms) and an LPC (450-850 ms). For the idiomatic sentences, however, they authors only report an N400 effect for the unexpected, within-language words (brain), but no negativity for the code-switched words (*mente*). We suspect that this reflects participants' ability to predict the final word of the sentence well ahead of the time (since they know the idiom) and retrieve the relevant item in both languages (since they know that it is equally likely to end in either word). In this study, the average amplitude from 250 to 450 ms for the expected words in both regular and idiomatic sentences was about 5 μV. For code-switched words, the average amplitude in this time window for regular sentences was more negative (about 3 μV), reflecting the switch-related negativity reported in the study. However, in idiomatic sentences, the average amplitude for code-switched words was nearly identical to that of the expected word (about 5 μV). This evidence supports the idea above that participants were able to predict *both* the English and Spanish words in the highly predictable idioms but not in the regular sentences.

This raises an interesting question of whether code-switching in the wild ever becomes predictable enough to facilitate lexical processing in this way. Code-switching is not random, instead it is argued to serve a range of discourse functions that might allow a listener to predict a switch (Auer, 1988; Gumperz, 1982; Heller, 2007; Poplack, 1980; Sebba et al., 2012). For example, there may be some words that are always code-switched, making the within-language word heavily dispreferred and arguably unexpected. In Spanish, the use of the English word *email* is preferred over the Spanish term *correo electrónico*. Similarly, many Spanish speakers will use *bar* instead of *la cantina, cervecería,* or *coctelería* when discussing where to meet up for drinks. In this scenario, the 'expected' code-switch (*email, bar*) should elicit smaller N400 effects relative to the more unusual, within-language word (*correo electrónico, cantina*). Another interesting question would be whether or not the bilinguals consider these borrowed words to be "code-switches" at all—as they are probably widely accepted in their language. This theory would make the following prediction: bilinguals that consider the words to be "code-switches" should show LPC effects, whereas those that do not consider the word as a language switch would not show LPC effects. To the best of our knowledge, there is no study that investigates these hypotheses, but clearly under our account, N400 effects for code-switches are predicted to disappear when code-switching is expected, and LPC effects should emerge whenever the word is interpreted as an unexpected switch in the language.

Taken together, the variability in the literature on switch-related N400 responses can be accounted for by a simple prediction account. In the next section, we address the implications of these findings for our theories of the N400 and its sensitivity to form-based predictions.

### 4.4. *What does this tell us about the functional significance of the N400?*

For decades, the N400 response has been used in psycholinguistic studies to determine the degree to which a particular context leads comprehenders to make lexico-semantic predictions, easing the processing of a word once it is encountered (Federmeier, 2007; Kutas et al., 2006; Lau et al., 2013; Otten & Van Berkum, 2008; Van Berkum et al., 2005; Wlotko & Federmeier, 2015). There is an overwhelming amount of evidence showing that N400 responses are reduced when comprehenders are able to predict or pre-activate *semantic* features associated with upcoming words (e.g. Federmeier & Kutas, 1999; Federmeier et al., 2002; see Federmeier, 2007; Kuperberg, 2007; Kutas & Federmeier, 2011; Kuperberg et al., 2019 for reviews). In contrast, there is less evidence that the N400 response is sensitive to the pre-activation of features associated with a word's grammatical, phonological, or orthographic form (see Nieuwland, 2019). Nevertheless, there are a handful of studies that demonstrate that under certain circumstances comprehenders *can* anticipate these form-based features, resulting in reduced N400 responses (e.g. Brothers et al., 2015; DeLong et al., 2005; Ito et al., 2017; Ito, Corley, Pickering, Martin, & Nieuwland, 2016; Van Berkum et al., 2005; Wicha et al., 2004; Wicha, Bates, et al., 2003; Wicha, Moreno, & Kutas, 2003). We argue that the present study provides more evidence for this hypothesis: In our rich contexts, bilinguals seem to predict a particular word in a particular language. This preactivation of a language-specific form leads to the reduction of the N400 response for that particular word and not its translation equivalent, which matches in semantic features. Our hypothesis makes the interesting prediction that, if the comprehender predicts form-based features for the expected word, any exact or near-cognate of that word would result in a reduced N400 effect. This is consistent with the literature on cognates showing facilitated lexical processing and N400 reductions as a function of form overlap (Christoffels, Firk, & Schiller, 2007; De Groot, 1993; De Groot & Nas, 1991; Dijkstra, Van Hell, & Brenders, 2015; Gollan & Acenas, 2004). Similar N400 reductions have been observed in monolingual populations when words are slightly misspelled (e.g. cake vs. *ceke*, see Kim & Lai, 2012). We take this as evidence that the N400 response is sensitive to some degree of form-based prediction, at least at the level of a particular lexical item, specified for its language.

However, we fully acknowledge that the evidence for prediction of pure word form features is limited and highly controversial at the moment (see Nieuwland et al., 2018; Nieuwland, 2019 for further discussion). Thus, an alternative explanation for our data could be that language is represented (and predicted) in a similar way to other form features like grammatical gender or number. There is ample evidence to suggest that comprehenders are capable of pre-activating features like grammatical gender (e.g. Wicha et al., 2004; Wicha, Bates, et al., 2003; Wicha, Moreno, & Kutas, 2003). Under this account, bilinguals would need to predict both the semantic features *and* the language feature of an upcoming word in order to explain our pattern of results. If this proves to be the case, it might provide support for theories where lexical representations are divided into two levels: the lemma, which links conceptual and syntactic features, and the lexeme, which contains links to the phonological features (see Roelofs, Meyer, & Levelt, 1998; cf. Caramazza, 1998). We would, however, need to specify that lemmas are sensitive to a language feature in the same way that they are sensitive to other grammatical and syntactic features like gender, number, and person (for similar proposals, see Poulisse & Bongaerts, 1994; Bullock & Toribio, 2019). In order to test this hypothesis, we would need to have a manipulation in which two words share semantic *and* language features, but differ in word form features. Future work could investigate words that have acceptable alternative spellings (e.g. ax/axe, donut/doughnut, dialog/dialogue) in order to assess the degree to which semantic, language, and form-based features are predicted independently. To the best of our knowledge, there is no study that uses a manipulation of this kind. Thus, for the purpose of the present discussion, we will say that the most parsimonious way to interpret our data is to assume that word form prediction can occur and that the limited evidence reflects the limits of experimental power, variability in the strength of predictive cues, the time available to make predictions, and/or the motivation or speed of processing in the research participants across studies.

### 4.5. *What does the present study tell us about LPC effects?*

The present study was specifically designed to explore the N400 rather than the LPC. As a result, our main hypothesis (the one-cost account) makes no differential predictions regarding the LPC effects for strong and weak-fitting code-switches. The code-switching literature suggests that LPC effects to code-switching are influenced by two main factors: the expectedness of the code-switching event and participants' language proficiency (e.g. Moreno et al., 2002, 2008; Proverbio et al., 2004; Ruigendijk et al., 2016; Van Der Meij et al., 2011; Van Hell et al., 2018, 2015; Van Hell & Witteman, 2009). LPC effects are often reduced (or missing) in experiments where the code-switching manipulation is highly predictable (see Proverbio et al., 2004). In the present study, our code-switching manipulation occurred at seemingly random intervals throughout the stories. Thus, participants were unable to guess when a code-switched word might appear—unlike in some prior studies that always manipulate the last word in the target sentence. We return to this point below in our discussion of the benefits and limitations of the Storytime paradigm. LPC effects can also be smaller and earlier when participants are more proficient in the code-switched language (Moreno et al., 2002; Ruigendijk et al., 2016; Litcofsky & Van Hell, 2017; cf. Van Der Meij et al., 2011). In the present study, we did not manipulate proficiency levels in either English or Spanish. Our study population was largely dominant in English (the matrix language) but still reported high proficiency in Spanish (see Table 1). Nonetheless, there was still some variability in the levels of proficiencies across participants—thus, we conducted a set of exploratory analyses to see if proficiency influenced the size of our LPC effects. However, these analyses did not reveal any significant effects of Spanish or English proficiency levels, perhaps due to the homogeneity of our study population. More information about these analyses can be found in our annotated analysis script on OSF (https://osf.io/jwqpr/).

Taken together, the present study provides critical information for understanding the functional significance of the LPC effects found in both monolingual and bilingual contexts. Most researchers agree that switch-related LPC effects index two aspects of comprehending a code-switch: First, the initial recognition of the switch, and then the subsequent reanalysis of the input and the prior context (Litcofsky & Van Hell, 2017; Moreno et al., 2002; Van Hell et al., 2018). This reanalysis process is argued to involve sentence or discourse-level restructuring, which is why LPC effects are seldom found in studies using single, isolated words (cf. Alvarez et al., 2003; Chauncey, Grainger, & Holcomb, 2008; Midgley et al., 2009; see Van Hell et al., 2018, 2015 for discussion). LPC effects, more broadly, have been argued to reflect a failure to update a participant's *mental model*, which is her, "high-level representation of meaning that is established and built during comprehension, based on the preceding linguistic and non-linguistic context, the comprehender's real-word knowledge, and her beliefs about the communicator and the broader communicative environment" (Kuperberg et al., 2019). This contrasts with the N400, which is seen as an index of lexico-semantic activation and integration. Thus, we expect that the LPC will not reflect the fit of the word within its context (weak vs. strong) but instead will reflect the degree to which the speech act fits into the broader communicative environment. This interpretation makes the prediction that using a design that supports code-switching events may reduce (or even eliminate) these posterior effects (see Moreno et al., 2002, 2008; Blanco-Elorrieta & Pylkkänen, 2016, 2017, 2018 for more discussion).

### 4.6. Methodological advantages and limitations to the Storytime paradigm

One goal of this study was to explore code-switching in a more natural context than is typically used in experimental studies. For this reason, we used auditory materials rather than written text (cf. Moreno et al., 2002; Ng et al., 2014) since code-switching is more common in spoken language (Fernandez et al., 2019; Litcofsky & Van Hell, 2017; Van Hell et al., 2018). To make the task more realistic and engaging, we used complete natural discourses rather than isolated words or sentences (cf. Alvarez et al., 2003; Liao & Chan, 2016; Moreno et al., 2002; Ruigendijk et al., 2016). Finally, we gave our participants no task beyond enjoying the story (cf. Ng et al., 2014), in an effort to remove potential strategic considerations.

One core characteristic of our Storytime paradigm is that our discourses are naturally produced, meaning that the speaking rate is not tightly controlled and arguably faster than those in traditional psycholinguistic experiments. One benefit of producing words naturally, however, is that each word is presented at a speed that is correlated to the word's length. The alternative is presenting all words at the same stimulus onset asynchrony (SOA), which may benefit shorter words but make longer words more challenging to process. Previous studies looking at form-based prediction in sentences found that at faster SOAs (e.g. 500 ms), the N400 effects are reactive (bottom-up) rather than predictive (Ito et al., 2016; Ito et al., 2017). In fact, the average SOA in our recordings is 521.9 ms ($SE = 8.0$) with a range of 20 ms to 2570 ms. This finding would suggest that the faster speaking rates in our paradigm would make it harder for listeners to make form-based predictions—-which seems counter to our argument that naturalistic listening enhances prediction.

There are, however, two other temporal variables that are relevant to prediction, both of which we suspect will favor natural discourse. First, there is the time that passes between the material that generates the expectation and the input in which the expectation is realized. For example, an expectation for an upcoming word could be formed many words (or even many sentences) before the word is actually produced. Take, for example, this discourse: "Tim kept talking about wanting a cat for his birthday. So, when his birthday finally arrived, I went to the pet shop and bought him an adorable black and white...**cat**." In this case, a listener might have the time to make robust predictions even if the speech was quite rapid. This seems more likely to happen in a connected discourse in which ideas build upon each other over many sentences. Second, lexical prediction will depend on the amount of processing time needed to make the conceptual prediction and retrieve the relevant form. If a particular concept or lexical item is already active (because it is central to the discourse, appeared earlier, or is in the same semantic neighborhood as words that appeared earlier), then it may take less time to access that word for the purposes of form-based prediction—just like it would take less time to produce it. These factors would appear to favor prediction in rich connected discourse relative to isolated sentences. Nevertheless, we suspect that there might be even more prediction in a given discourse if the speech rate is slower. Future studies should address the interaction between these temporal variables to better understand the conditions under which we make form-based predictions.

Next, there are clearly ways in which our paradigm did not fully capture the rich context that code-switching typically occurs in. First, our design required that participants hear unexpected words in addition to code-switched words. This may have led participants to process unexpected items (e.g. code-switches) in a different way than they otherwise would have (see Van Berkum et al., 2005 for reasons to avoid implausible discourses). Second, our study involved single-word insertions rather than sentence alternations (i.e. where the sentence continues in the other language following the code-switch). Single-word insertions are less frequent than sentence alternations (e.g. Litcofsky & Van Hell, 2017; Van Hell et al., 2018, 2015; cf. Poplack, 2018), especially in Spanish-English bilingual communities (Deuchar, Davies, Herring, Parafita Couto, & Carter, 2014; Fernandez et al., 2019; Milroy & Muysken, 1995). We did this to minimize the differences across conditions such that effects of one trial would be unlikely to bleed into the next. Because we were primarily interested in the processes occurring immediately at the time of the switch—rather than downstream effects on subsequent words—this choice seemed optimal. But consistent use of single-word insertions may have made it more difficult for our bilinguals to adapt or may have introduced additional difficulties when switching back into the matrix language. Third, we selected the target words based solely on their cloze probabilities and then randomly assigned them to conditions. This meant that our code-switching events were not clearly motivated by the discourse, cultural practices, and/or language accessibility (see the *email* and *bar* examples in Section 4.3).

However, we still believe that these findings provide useful (and generalizable) information about how code-switching is processed in natural conversation. In most contexts, with most words, the discourse pressures and internal forces that lead one speaker to switch languages are unlikely to be completely transparent to the listener. Thus, much of the time, the within-language word will probably be expected by the listener, rather than its translation equivalent. When this is the case, we should expect the pattern of effects found here. Some initial support for this hypothesis comes from an MEG study by Blanco-Elorrieta and Pylkkänen (2017). They used naturally occurring dialogues between bilingual speakers and found an increased activation to code-switched words in auditory cortex, suggesting that listeners may predict the within-language target and have to put in extra effort to overcome this even in these natural dialogues (see Blanco-Elorrieta & Pylkkänen, 2016, 2017, 2018). There is also a possibility that natural speech contains acoustic properties that could better signal code-switching—for example, the rate of speech may be faster for the matrix language relative to code-switched material, or there could be natural pauses prior to code-switching when produced naturally, or even subtle changes in articulators may provide natural cues rather than the unnatural transitions generated from splicing. Future work could address these issues with respect to the N400 by using dialogues between bilingual speakers as the base for stimulus creation, by implementing sentence alternations rather than single-word insertions, and by eliminating the weak-fit conditions.

## 5. Conclusion

In bilingual communities, speakers often switch between languages, and their listeners seem to readily follow them. Psycholinguistic research has suggested that these code-switches may be costly for listeners in some situations. The present study explored those costs by comparing them to the difficulties associated with hearing unexpected words within a single language context. Using our novel Storytime paradigm, we found three effects: an initial prediction effect (the N400), a post-lexical recognition of the switch in languages (the LPC), and a prolonged integration difficulty associated with weak-fitting words regardless of language (the sustained negativity). Together, these findings suggest that the difficulties that bilinguals encounter in understanding code-switched words can largely be understood within more general frameworks for understanding language comprehension. This work is consistent with other findings suggesting that a bilingual is not someone with two separate and competing languages living in their mind (e.g. Dijkstra & Van Heuven, 2002). Rather, bilinguals are individuals with a language system optimized to handle two coding systems, where a single lexical concept can be readily mapped onto two distinct forms (e.g. Emmorey, Borinstein, Thompson, & Gollan, 2008; Van Hell & De Groot, 2008). As a result, the phenomenon of comprehending code-switched words in conversation can be understood as comprehending an unexpected word that just happens to be in another language.

## Funding

This work was supported in part by Harvard's Barbara Ditmars Fund awarded to Anthony Yacovone and Harvard College Research Program funds awarded to Emily Moya.

## Acknowledgements

We would like to thank Maribelle Dickins, Karen Andres, Paulina Piwowarczyk, Moshe Poliak, and Harita Koya for their assistance; Tanya Levari for invaluable training, support, and inspiration; Alfonso Caramazza and other members of the Harvard Psychology community for their feedback on earlier drafts; Sarah Brown-Schmidt and three anonymous reviewers for their insightful questions and recommendations, all of which strengthened this work; and finally, all of the individuals who participated in our studies. Various aspects of this research project were presented at the 32nd and 34th Annual CUNY Conference on Human Sentence Processing (March 2018 in Davis, CA; March 2021 virtually hosted in Philadelphia, PA).

## Appendix A. Critical Trials

This appendix contextualizes all 120 target words and their violation conditions. In each story snippet, there are four alternative words: (strong-fit English | strong-fit Spanish | weak-fit English | weak-fit Spanish). Each target context is listed with the cloze probability of the strong-fit English word for that sentence.

All of the story scripts and the audio recordings from the experiment can be found on OSF (https://osf.io/jwqpr/). Trials 1–62 are from *"Hair Today, Gone Tomorrow"* by Jenny Allen. Trials 63–120 are from *"The Scammer Who Loved Me (Not)"* by Sofija Stefanovic. The original performances can be found on the Moth story webpage (see https://themoth.org/).

| Trial | Cloze | Sentence Snippet |
|---|---|---|
| 1 | 91.43% | ...Now, when you have the kind of chemotherapy I had, you lose your (hair \| pelo \| fur \| pelaje)... |
| 2 | 91.43% | ...My hair, my life, my hair, my (life \| vida \| spirit \| ánimo), I don't know... |
| 3 | 8.57% | ...I have this kind of unruly hair, and it has a (mind \| mente \| brain \| cerebro) of its own... |
| 4 | 80.00% | ...I've had it cut by a mental (patient \| paciente \| subject \| subjeto)... |
| 5 | 94.29% | ...Would I try to hide it, or would I announce it to the whole (world \| mundo \| sphere \| esfera)?... |
| 6 | 48.57% | ...In other (words \| palabras \| names \| apodos), was I gonna be a scarf person or a wig person?... |
| 7 | 80.00% | ...I showed up places suddenly wearing a scarf all the (time \| veces \| months \| meses)… |
| 8 | 42.86% | ...So I, uh, uh, had this dilemma, uh, whether to wear a (scarf \| bufanda \| tarp \| lona) or a wig… |
| 9 | 2.86% | ...you know cuz I, uh, have a lot of (self \| persona \| soul \| alma) righteous integrity... |
| 10 | 57.14% | ...So, one night, uh, I ran into my (friend \| amiga \| comrade \| camarada) Ruth... |
| 11 | 45.71% | ...our (teeth \| dientes \| molars \| muelas) are only supposed to last us about forty five years... |
| 12 | 82.86% | ...And they even went to the same (school \| escuela \| dungeon \| mazmorra) cuz she'd recommended it... |
| 13 | 45.71% | ...So, I'd seen her a few weeks earlier just wearing a (scarf \| bufanda \| tarp \| lona)... |
| 14 | 40.00% | ..."I never thought I'd wear a (wig \| peluca \| pelt \| pellejo)!" ... |
| 15 | 17.14% | ...And, uh, I love free (things \| cosas \| objects \| objetos) you know... |
| 16 | 20.00% | ...So, a couple of (weeks \| semanas \| cases \| asuntos) go by, and my hair does fall out... |
| 17 | 31.43% | ...It sort of gradually gave up the (ghost \| fantasma \| shadow \| sombra)... |
| 18 | 14.29% | ...You know, first in these strands in my (brush \| cepillo \| bristles \| cerdas)... |
| 19 | 71.43% | ...And then in clumps in my shower (drain \| desagüe \| hole \| hueco)... |
| 20 | 85.71% | …every time I looked in the (mirror \| espejo \| glass \| vidrio), my baldness told me how sick I was... |
| 21 | 25.71% | ...And so I thought, "Well maybe I'll go to the (store \| tienda \| depot \| almacén) and get a wig"... |
| 22 | 71.43% | ...it was called "bits and (pieces \| trozos \| dots \| puntos)"... |
| 23 | 40.00% | ..."I don't know, how many (kinds \| tipos \| genres \| géneros) do you have?"... |
| 24 | 85.71% | ...they had wigs made from the hair of Caucasian European (women \| mujeres \| dames \| damas)... |

(continued on next page)

(*continued*)

| Trial | Cloze | Sentence Snippet |
|---|---|---|
| 25 | 42.86% | ...the Indian hair wigs were in the middle (range \| intervalo \| span \| lapso)... |
| 26 | 94.29% | ...it's just obscene to spend this kind of (money \| dinero \| loot \| plata) on a wig... |
| 27 | 88.57% | ...four times as expensive as the hair of Indian (women \| mujeres \| dames \| damas)?... |
| 28 | 60.00% | ...And, uh, I took the bag (home \| hogar \| abode \| domicilio)... |
| 29 | 48.57% | ...I put it in a corner of my (bedroom \| dormitorio \| gym \| gimnasio)… |
| 30 | 82.86% | ...old ladies, who's pushing her grocery (cart \| carro \| truck \| camión) down Broadway... |
| 31 | 42.86% | ...It seems like a lot of (trouble \| molestia \| tumult \| disturbia)… |
| 32 | 91.43% | ...And for some (reason \| razón \| sense \| sentido), this is completely unexpected by me... |
| 33 | 54.29% | ...Having no (eyebrows \| cejas \| sideburns \| patillas) makes me feel very naked… |
| 34 | 82.86% | ...My hair was part of my head, but my eyebrows were part of my (face \| cara \| mask \| mascara)... |
| 35 | 94.29% | ...And I decide, you know, this might be the perfect occasion to wear my (wig \| peluca \| pelt \| pellejo)... |
| 36 | 74.29% | ...cuz every time I look in a store (window \| ventana \| door \| puerta), I recognize myself… |
| 37 | 82.86% | ...I feel like I've done something really bad like robbed a (bank \| banco \| fund \| fondo)... |
| 38 | 20.00% | ...And now I'm trying to just kind of lose myself in the (crowd \| multitud \| troupe \| compañía)… |
| 39 | 57.14% | ...And I'm uncomfortable having it in my personal (space \| espacio \| quarters \| cuarteles)... |
| 40 | 42.86% | ...we're gonna be spending about four hours in the scorching Chicago (sun \| sol \| light \| luz)… |
| 41 | 74.29% | ...So, I put the (hat \| sombrero \| crown \| corona) on over my wig... |
| 42 | 85.71% | ...And I take my place in a folding chair among this ocean of folding (chairs \| sillas \| beds \| camas)… |
| 43 | 68.57% | ...by the second (hour \| hora \| chapter \| capitulo) or so, uh, my head is just baking... |
| 44 | 48.57% | ...And, um, these little rivulets of (sweat \| sudor \| saline \| salina) are coming down… |
| 45 | 88.57% | ...And the wig itself is so hot and heavy on my (head \| cabeza \| cranium \| cráneo) ... |
| 46 | 91.43% | ...It feels like I'm wearing my cat, um, on my (head \| cabeza \| cranium \| cráneo)... |
| 47 | 74.29% | ...you know, in one of those nearly free fall slow (motion \| movimiento \| travel \| viaje) moments… |
| 48 | 91.43% | ...My wig comes off with my (hat \| sombrero \| crown \| corona)... |
| 49 | 37.14% | ...I've gone through all this (trouble \| molestia \| tumult \| disturbia) only to end up like this... |
| 50 | 80.00% | ...I feel so bad for the (people \| gente \| citizens \| ciudadanos) behind me I can't even look at them... |
| 51 | 51.43% | ...And I, uh, sit through the rest of the (ceremony \| ceremonia \| ritual \| rito), and then I go back… |
| 52 | 80.00% | ...I put it way in the back of one of my dresser (drawers \| cajones \| shelves \| estantes)... |
| 53 | 74.29% | ...But, uh, for the first (time \| veces \| months \| meses), I'm kinda glad to see it... |
| 54 | 42.86% | ...And for two (years \| años \| units \| unidades), we have just a lot of fun… |
| 55 | 65.71% | ...And eating the delicious (food \| comida \| morsel \| bocado) that she cooks at her house... |
| 56 | 71.43% | ...It's when people come up to you and tell you an inspiring (story \| historia \| fable \| fabula) life... |
| 57 | 71.43% | ...at the (end \| fin \| closure \| cierre) of these stories, you always ask the people “how is she doing?”... |
| 58 | 45.71% | ...We crack each other up with our (stories \| historias \| fables \| fabulas)... |
| 59 | 71.43% | ...I meet in various doctors' waiting (rooms \| cuartos \| cubicles \| cubículos) die... |
| 60 | 82.86% | ...ask yourself if you are ever, ever, gonna wear that bridesmaid's (dress \| vestido \| shirt \| camisa)... |
| 61 | 85.71% | ...And I might lose my (hair \| pelo \| fur \| pelaje) again... |
| 62 | 68.57% | ...Better to let (people \| gente \| citizens \| ciudadanos) ask me questions... |
| 63 | 62.86% | ...I found myself in another (relationship \| relación \| linkage \| enlace) with a woman called Cindy... |
| 64 | 34.29% | ...And it was based on deception and guilt, and it left me feeling like (crap \| mierda \| feces \| heces)... |
| 65 | 37.14% | ...I had written for a TV (show \| programa \| exhibit \| obra) a couple years back… |
| 66 | 11.43% | ...they start a (relationship \| relación \| linkage \| enlace) with them… |
| 67 | 65.71% | ...I celebrated Bill's (birthday \| cumpleaños \| appointment \| cita) with him... |
| 68 | 25.71% | ...he sort of knew in the back of his (head \| cabeza \| cranium \| cráneo) that he was being scammed... |
| 69 | 91.43% | ...scam victims were pretty gullible (people \| gente \| citizens \| ciudadanos)… |
| 70 | 54.29% | ...He was a really smart, worldly (man \| hombre \| specimen \| especie)... |
| 71 | 60.00% | ...And so last (year \| año \| session \| sesión), my boyfriend Michael and I make the big move... |
| 72 | 74.29% | ...And every time, Michael goes to (work \| trabajo \| task \| tarea) and makes new friends... |
| 73 | 80.00% | ...And I stay at (home \| hogar \| abode \| domicilio) researching scams... |
| 74 | 85.71% | ...And I tell her what the (weather \| clima \| biome \| bioma) is like in Mumbai... |
| 75 | 82.86% | ...we're all lying to each other at this (point \| punto \| spot \| lugar)... |
| 76 | 94.29% | ...she thinks I'm a middle-aged (man \| hombre \| specimen \| especie) in Mumbai... |
| 77 | 80.00% | ...And I also use like a fake unisex (name \| nombre \| label \| etiqueta)... |
| 78 | 34.29% | ...I'm a(n) (fan \| aficionada \| preacher \| predicador)... |
| 79 | 11.43% | ...Uh, because (soccer \| fútbol \| skiing \| esquiar) is big in Senegal... |
| 80 | 62.86% | ...my Senegalese (girlfriend \| novia \| affiliate \| afiliada) and I chat online... |
| 81 | 80.00% | ...And when my (boyfriend \| novio \| suitor \| pretendiente) is not at work, I tactfully close my laptop... |
| 82 | 68.57% | ...I get so many (pings \| sonidos \| pulses \| pulsos) from Cindy... |
| 83 | 88.57% | ...the most attentive person I've ever semi-dated in my (life \| vida \| spirit \| ánimo)... |
| 84 | 82.86% | ...Which is, in some (ways \| maneras \| paths \| caminos), really nice… |
| 85 | 40.00% | ...technically that's her (job \| trabajo \| post \| puesto), right, as a scammer... |
| 86 | 11.43% | ...But no (money \| dinero \| loot \| plata) requests... |
| 87 | 91.43% | ...this is taking up quite a lot of my (time \| tiempo \| months \| meses)... |
| 88 | 48.57% | ...So I type, “Hey Cindy I have a (confession \| confesión \| disclosure \| revelación) to make”... |
| 89 | 80.00% | ...And by this (point \| punto \| spot \| lugar) I thought that I would be Cindy free... |
| 90 | 82.86% | ...she's asking me for an (picture \| imagen \| likeness \| cuadro)... |
| 91 | 91.43% | ...I don't have any pictures on my hard (drive \| disco \| gadget \| dispositivo)... |
| 92 | 62.86% | ...hey, listen you have been lying to me for several (weeks \| semanas \| cases \| casos) now... |
| 93 | 57.14% | ...And in this (moment \| momento \| juncture \| coyuntura) because I'm kind of prone to feeling anxious... |
| 94 | 77.14% | ...trace back to the real me, and then send me a dead rat in the (mail \| correo \| bundle \| bulto)... |
| 95 | 8.57% | ...So, I end up dealing with all this (stuff \| cosas \| substance \| substancia) ... |
| 96 | 65.71% | ...I can see by the little (dots \| puntos \| blobs \| manchas), and she sends it... |
| 97 | 88.57% | ...But, I have fallen in (love \| amor \| rapture \| rapto) with you... |
| 98 | 77.14% | ...even though you're a (woman \| mujer \| maiden \| doncella)... |
| 99 | 91.43% | ...And she picks up the (phone \| teléfono \| appliance \| aparato) and says “Hello” and I say “Hello”... |
| 100 | 8.57% | ...And suddenly my scammer not only has a (face \| cara \| mask \| mascara)… |

(*continued on next page*)

(*continued*)

| Trial | Cloze | Sentence Snippet |
|---|---|---|
| 101 | 37.14% | ...And I say, “Oh, do you have (kids \| niños \| youths \| juventudes)”... |
| 102 | 85.71% | ...And I think maybe she's not telling me the (truth \| verdad \| reality \| realidad), that she's a parent... |
| 103 | 57.14% | ...my (friend \| amiga \| comrade \| camarada) said that maybe Cindy's a scammer... |
| 104 | 71.43% | ...Hell, what if I've got this whole (thing \| cosa \| object \| objeto) wrong... |
| 105 | 68.57% | ...I can't help but feel guilty cuz I think of her tired (voice \| voz \| tone \| tono)... |
| 106 | 57.14% | ...I think about that (baby \| bebé \| suckling \| mamón) crying... |
| 107 | 94.29% | ...And I think that 140 dollars isn't really that much (money \| dinero \| loot \| plata)... |
| 108 | 91.43% | ...I see that 50% of its population lives in (poverty \| pobreza \| beggary \| mendigos)... |
| 109 | 65.71% | ...And she reminds me of her hard (life \| vida \| spirit \| ánimo)... |
| 110 | 2.86% | ...And she says, “I'm trying to be a good (girl \| niña \| damsel \| damisela)”... |
| 111 | 2.86% | ...And I feel like a (jerk \| imbécil \| rascal \| pícaro) for stringing her along... |
| 112 | 34.29% | ...But in any (case \| caso \| sample \| muestra), while I'm typing, I find myself crying... |
| 113 | 40.00% | ...tell me about her real self and about being a (scammer \| estafador \| rogue \| pillo)... |
| 114 | 54.29% | ...And the next (day \| día \| dawn \| amanecer), Cindy writes back and ignores most of my email... |
| 115 | 88.57% | ...I'm not going to do it unless she admits to being a (scammer \| estafador \| rogue \| pillo)... |
| 116 | 85.71% | ...And we go back and forth like this for about a(n) (week \| semana \| case \| asunto)... |
| 117 | 14.29% | ...But still I told myself maybe, you know, she's a single (parent \| madre \| creator \| creadora)... |
| 118 | 51.43% | ...you overlook so much bad (stuff \| cosas \| substance \| substancia)... |
| 119 | 25.71% | ...I wonder if her (baby \| bebé \| suckling \| mamón) still cries while she's scamming people... |
| 120 | 91.43% | ...she'll one (day \| día \| dawn \| amanecer) take her revenge... |

## Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2021.104814 or on OSF (https://osf.io/jwqpr/).

## References

Altarriba, J., Kroll, J. F., Sholl, A., & Rayner, K. (1996). The influence of lexical and conceptual constraints on reading mixed-language sentences: Evidence from eye fixations and naming times. *Memory & Cognition, 24*(4), 477–492.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*(3), 247–264.

Alvarez, R. P., Holcomb, P. J., & Grainger, J. (2003). Accessing word meaning in two languages: An event-related brain potential study of beginning bilinguals. *Brain and Language, 87*(2), 290–304.

Auer, P. (1988). A conversation analytic approach to code-switching and transfer. *Codeswitching: Anthropological and Sociolinguistic Perspectives, 48*, 187–213.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting linear mixed-effects models using lme4*. arXiv preprint arXiv:1406.5823.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don’t count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 238–247).

Blanco-Elorrieta, E., & Pylkkänen, L. (2016). Bilingual language control in perception versus action: MEG reveals comprehension control mechanisms in anterior cingulate cortex and domain-general control of production in dorsolateral prefrontal cortex. *Journal of Neuroscience, 36*(2), 290–301.

Blanco-Elorrieta, E., & Pylkkänen, L. (2017). Bilingual language switching in the laboratory versus in the wild: The spatiotemporal dynamics of adaptive language control. *Journal of Neuroscience, 37*(37), 9022–9036.

Blanco-Elorrieta, E., & Pylkkänen, L. (2018). Ecological validity in bilingualism research and the bilingual advantage. *Trends in Cognitive Sciences, 22*(12), 1117–1126.

Boersma, P., & Weenink, D. (2001). *Praat speech processing software*. Institute of Phonetics Sciences of the University of Amsterdam. http://www. praat. org.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2019). Toward a neurobiologically plausible model of language-related, negative event-related potentials. *Frontiers in Psychology, 10*, 298.

Borovsky, A., Elman, J. L., & Kutas, M. (2012). Once is enough: N400 indexes semantic integration of novel word meanings from a single exposure in context. *Language Learning and Development, 8*(3), 278–302.

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: Prediction takes precedence. *Cognition, 136*, 135–149.

Brown, C., & Hagoort, P. (1993). The processing nature of the N400: Evidence from masked priming. *Journal of Cognitive Neuroscience, 5*(1), 34–44.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.

Bullock, B. E., & Toribio, A. J. (2019). Conceptual and empirical arguments for a language feature: Evidence from language mixing. In *Contributions of romance languages to current linguistic theory* (pp. 93–113). Cham: Springer.

Bultena, S., Dijkstra, T., & Van Hell, J. G. (2015). Language switch costs in sentence comprehension depend on language dominance: Evidence from self-paced reading. *Bilingualism: Language and Cognition, 18*(3), 453–469.

Caffarra, S., Mendoza, M., & Davidson, D. (2019). Is the LAN effect in morphosyntactic processing an ERP artifact? *Brain and Language, 191*, 9–16.

Caramazza, A. (1998). The interpretation of semantic category-specific deficits: What do they reveal about the organization of conceptual knowledge in the brain? *Neurocase, 4*(4–5), 265–272.

Caramazza, A., & Brones, I. (1979). Lexical access in bilinguals. *Bulletin of the Psychonomic Society, 13*(4), 212–214.

Chauncey, K., Grainger, J., & Holcomb, P. J. (2008). Code-switching effects in bilingual word recognition: A masked priming study with event-related potentials. *Brain and Language, 105*(3), 161–174.

Christoffels, I. K., Firk, C., & Schiller, N. O. (2007). Bilingual language control: An event-related brain potential study. *Brain Research, 1147*, 192–208.

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*(3), 256–266.

Connolly, J. F., Phillips, N. A., & Forbes, K. A. (1995). The effects of phonological and semantic features of sentence-ending words on visual event-related brain potentials. *Electroencephalography and Clinical Neurophysiology, 94*(4), 276–287.

Coulson, S., King, J. W., & Kutas, M. (1998). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language & Cognitive Processes, 13*(1), 21–58.

Coulson, S., & Kutas, M. (2001). Getting it: Human event-related brain response to jokes in good and poor comprehenders. *Neuroscience Letters, 316*(2), 71–74.

Courteau, É., Martignetti, L., Royle, P., & Steinhauer, K. (2019). Eliciting ERP components for morphosyntactic agreement mismatches in perfectly grammatical sentences. *Frontiers in Psychology, 10*, 1152.

Cuetos, F., Glez-Nosti, M., Barbon, A., & Brysbaert, M. (2011). SUBTLEX-ESP: frecuencias de las palabras espanolas basadas en los subtitulos de las peliculas. *Psicológica, 32*(2), 133–144.

De Groot, A. M. (1993). *Word-type effects in bilingual processing tasks. The bilingual lexicon* (pp. 27–51). Amsterdam: John Benjamins.

De Groot, A. M., & Nas, G. L. (1991). Lexical representation of cognates and noncognates in compound bilinguals. *Journal of Memory and Language, 30*(1), 90–123.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods, 41*(2), 385–390.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience, 8*(8), 1117.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2017). Is there a replication crisis? Perhaps. Is this an example? No: A commentary on Ito, Martin, and Nieuwland (2016). *Language, Cognition and Neuroscience, 32*(8), 966–973.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134*(1), 9–21.

Deuchar, M., Davies, P., Herring, J., Parafita Couto, M. C., & Carter, D. (2014). *Bilingual language use. Advances in the study of bilingualism* (pp. 93–110). Bristol: Multilingual Matters.

Dijkstra, T. (2005). Bilingual visual word recognition and lexical access. *Handbook of Bilingualism: Psycholinguistic Approaches*, 179–201.

Dijkstra, T., Grainger, J., & Van Heuven, W. J. (1999). Recognition of cognates and interlingual homographs: The neglected role of phonology. *Journal of Memory and Language, 41*(4), 496–518.

Dijkstra, T., Van Hell, J. G., & Brenders, P. (2015). Sentence context effects in bilingual word recognition: Cognate status, sentence language, and semantic constraint. *Bilingualism: Language and Cognition, 18*(4), 597–613.

Dijkstra, T., & Van Heuven, W. J. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition, 5*(3), 175–197.

Duyck, W., Van Assche, E., Drieghe, D., & Hartsuiker, R. J. (2007). Visual word recognition by bilinguals in a sentence context: Evidence for nonselective lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(4), 663.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior, 20*(6), 641–655.

Emmorey, K., Borinstein, H. B., Thompson, R., & Gollan, T. H. (2008). Bimodal bilingualism. *Bilingualism: Language and Cognition, 11*(1), 43–61.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology, 44*(4), 491–505.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language, 41*(4), 469–495.

Federmeier, K. D., McLennan, D. B., De Ochoa, E., & Kutas, M. (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.

Fernandez, C. B., Litcofsky, K. A., & Van Hell, J. G. (2019). Neural correlates of intra-sentential code-switching in the auditory modality. *Journal of Neurolinguistics, 51*, 17–41.

Fields, E. C. (2017). *Factorial mass univariate ERP toolbox*. Computer software]. Retrieved from https://github. com/ericcfields/FMUT/releases.

Fields, E. C. (2019). Using FMUT [Github Wiki Page]. Available at: https://github. com/ericcfields/FMUT/wiki/Using-FMUT.

Fields, E. C., & Kuperberg, G. R. (2020). Having your cake and eating it too: Flexibility and power with mass univariate statistics for ERP data. *Psychophysiology, 57*(2), Article e13468.

Fischler, I., & Bloom, P. A. (1979). Automatic and attentional processes in the effects of sentence contexts on word recognition. *Journal of Verbal Learning and Verbal Behavior, 18*(1), 1–20.

FitzPatrick, I., & Indefrey, P. (2014). Head start for target language in bilingual listening. *Brain research, 1542*, 111–130.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior, 12*(6), 627–635.

Foucart, A., & Frenck-Mestre, C. (2011). Grammatical gender processing in L2: Electrophysiological evidence of the effect of L1–L2 syntactic similarity. *Bilingualism: Language and Cognition, 14*(3), 379–399.

Foucart, A., & Frenck-Mestre, C. (2012). Can late L2 learners acquire new grammatical features? Evidence from ERPs and eye-tracking. *Journal of Memory and Language, 66* (1), 226–248.

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

Friederici, A. D. (2005). Neurophysiological markers of early language acquisition: From syllables to sentences. *Trends in Cognitive Sciences, 9*(10), 481–488.

Fromont, L. A., Steinhauer, K., & Royle, P. (2020). Verbing nouns and nouning verbs: Using a balanced design provides ERP evidence against "syntax-first" approaches to sentence processing. *PLoS One, 15*(3), Article e0229169.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and hierarchical/multilevel models*. New York, NY: Cambridge.

Gollan, T. H., & Acenas, L. A. R. (2004). What is a TOT? Cognate and translation effects on tip- of-the-tongue states in Spanish-English and tagalog-English bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(1), 246.

Grainger, J., & Beauvillain, C. (1987). Language blocking and lexical access in bilinguals. *The Quarterly Journal of Experimental Psychology Section A, 39*(2), 295–319.

Grainger, J., & Dijkstra, T. (1992). On the representation and use of language information in bilinguals. In *, vol. 83. Advances in psychology* (pp. 207–220). North-Holland.

Grainger, J., & Holcomb, P. J. (2009). Watching the word go by: On the time-course of component processes in visual word recognition. *Lang & Ling Compass, 3*(1), 128–156.

Grainger, J., Kiyonaga, K., & Holcomb, P. J. (2006). The time course of orthographic and phonological code activation. *Psychological Science, 17*(12), 1021–1026.

Green, D. W. (1998). Bilingualism and thought. *Psychologica Belgica, 38*(3–4), 251–276.

Grey, S., Schubel, L. C., McQueen, J. M., & Van Hell, J. G. (2018). Processing foreign-accented speech in a second language: Evidence from ERPs during sentence comprehension in bilinguals. *Bilingualism: Language and Cognition*, 1–18.

Grey, S., & van Hell, J. G. (2017). Foreign-accented speaker identity affects neural correlates of language comprehension. *Journal of Neurolinguistics, 42*, 93–108.

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011a). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology, 48*(12), 1711–1725.

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011b). Mass univariate analysis of event-related brain potentials/fields II: Simulation studies. *Psychophysiology, 48*(12), 1726–1737.

Grosjean, F. (2001). The bilingual's language modes. One mind, two languages. *Bilingual Language Processing, 122*.

Guajardo, L. F., & Wicha, N. Y. (2014). Morphosyntax can modulate the N400 component: Event related potentials to gender-marked post-nominal adjectives. *NeuroImage, 91*, 262–272.

Gumperz, J. J. (1982). *Discourse strategies* (vol. 1). Cambridge University Press.

Gunter, T. C., Friederici, A. D., & Schriefers, H. (2000). Syntactic gender and semantic expectancy: ERPs reveal early autonomy and late interaction. *Journal of Cognitive Neuroscience, 12*(4), 556–568.

Günther, F., Dudschig, C., & Kaup, B. (2015). LSAfun-an R package for computations based on latent semantic analysis. *Behavior Research Methods, 47*(4), 930–944.

Hagoort, P. (1993). Impairments of lexical-semantic processing in aphasia: Evidence from the processing of lexical ambiguities. *Brain and Language, 45*(2), 189–232.

Hahne, A., & Friederici, A. D. (1999). Electrophysiological evidence for two steps in syntactic analysis: Early automatic and late controlled processes. *Journal of Cognitive Neuroscience, 11*(2), 194–205.

Hale, J. (2001, June). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American chapter of the association for computational linguistics on language technologies* (pp. 1–8). Association for Computational Linguistics.

Hartsuiker, R. J., Pickering, M. J., & Veltkamp, E. (2004). Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals. *Psychological Science, 15*(6), 409–414.

Hasting, A. S., & Kotz, S. A. (2008). Speeding up syntax: On the relative timing and automaticity of local phrase structure and morphosyntactic processing as reflected in event-related brain potentials. *Journal of Cognitive Neuroscience, 20*(7), 1207–1219.

Heller, M. (Ed.). (2007). *Bilingualism: A social approach*. Springer.

Heredia, R. R., & Altarriba, J. (2001). Bilingual language mixing: Why do bilinguals code- switch? *Current Directions in Psychological Science, 10*(5), 164–168.

Holcomb, P. J., & Grainger, J. (2006). On the time course of visual word recognition: An event related potential investigation using masked repetition priming. *Journal of Cognitive Neuroscience, 18*(10), 1631–1643.

Holcomb, P. J., & Neville, H. J. (1991). Natural speech processing: An analysis using event- related brain potentials. *Psychobiology, 19*(4), 286–300.

Hoversten, L. J., & Traxler, M. J. (2020). Zooming in on zooming out: Partial selectivity and dynamic tuning of bilingual language control during reading. *Cognition, 195*, 104118.

Hut, S. C., & Leminen, A. (2017). Shaving bridges and tuning kitaraa: The effect of language switching on semantic processing. *Frontiers in Psychology, 8*, 1438.

Ito, A. (2016). *Prediction during native and non-native language comprehension: The role of mediating factors*. Doctoral dissertation. UK: University of Edinburgh.

Ito, A., Corley, M., Pickering, M. J., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language, 86*, 157–171.

Ito, A., Martin, A. E., & Nieuwland, M. S. (2017). On predicting form and meaning in a second language. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(4), 635.

Ito, A., & Pickering, M. J. (2021). Automaticity and prediction in non-native language comprehension. In T. Grüter, E. Kaan, & J. Benjamins (Eds.), *Prediction in second-language processing and learning*.

Ito, A., Pickering, M. J., & Corley, M. (2018). Investigating the time-course of phonological prediction in native and non-native speakers of English: A visual world eye-tracking study. *Journal of Memory and Language, 98*, 1–11.

Jordan, T. R., & Thomas, S. M. (2002). In search of perceptual influences of sentence context on word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(1), 34.

Kaan, E., Harris, A., Gibson, E., & Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Language & Cognitive Processes, 15*(2), 159–201.

Kaan, E., Kheder, S., Kreidler, A., Tomić, A., & Valdés Kroff, J. R. (2020). Processing code-switches in the presence of others: An ERP study. *Frontiers in Psychology, 11*, 1288.

Kambe, G., Rayner, K., & Duffy, S. A. (2001). Global context effects on processing lexically ambiguous words: Evidence from eye fixations. *Memory & Cognition, 29*(2), 363–372.

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49*(1), 133–156.

Kappenman, E. S., & Luck, S. J. (2012). ERP components: The ups and downs of brainwave recordings. In *The Oxford handbook of event-related potential components* (pp. 3–30).

Kaushanskaya, M., Blumenfeld, H. K., & Marian, V. (2020). The language experience and proficiency questionnaire (leap-q): Ten years later. *Bilingualism: Language and Cognition, 23*(5), 945–950.

Khamis-Dakwar, R., & Froud, K. (2007, February). Lexical processing in two language varieties. In *, vol. 290. Perspectives on Arabic linguistics: Papers from the annual symposium on Arabic linguistics. Volume XX: Kalamazoo, Michigan, March 2006* (p. 153). John Benjamins Publishing.

Kim, A., & Lai, V. (2012). Rapid interactions between lexical semantic and word form analysis during word recognition in context: Evidence from ERPs. *Journal of Cognitive Neuroscience, 24*(5), 1104–1112.

Kim, A., & Sikos, L. (2011). Conflict and surrender during sentence processing: An ERP study of syntax-semantics interaction. *Brain and Language, 118*(1–2), 15–22.

King, J. W., & Kutas, M. (1995). Who did what and when? Using word-and clause-level ERPs to monitor working memory usage in reading. *Journal of Cognitive Neuroscience, 7*(3), 376–395.

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.

Kolk, H., & Chwilla, D. (2007). Late positivities in unusual situations. *Brain and Language, 100*(3), 257–261.

Kotz, S. A., & Elston-Güttler, K. (2004). The role of proficiency on processing categorical and associative information in the L2 as revealed by reaction times and event-related brain potentials. *Journal of Neurolinguistics, 17*(2–3), 215–235.

Kroll, J. F., Dussias, P. E., Bogulski, C. A., & Kroff, J. R. V. (2012). Juggling two languages in one mind: What bilinguals tell us about language processing and its consequences for cognition. In *, vol. 56. Psychology of learning and motivation* (pp. 229–262). Academic press.

Kuperberg, G., Brothers, T., & Wlotko, E. (2019). A tale of two positivities (and the N400): Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *Journal of Cognitive Neuroscience*, 1–24.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research, 1146*, 23–49.

Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience, 31* (5), 602–616.

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences, 4*(12), 463–470.

Kutas, M., & Federmeier, K. D. (2009). N400. *Scholarpedia, 4*(10), 7790.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology, 62*, 621–647.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*(4427), 203–205.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*(5947), 161.

Kutas, M., & King, J. W. (1996). *The potentials for basic sentence processing: Differentiating integrative processes*.

Kutas, M., Moreno, E., & Wicha, N. (2009). Code-switching and the brain. In B. E. Bullock, & A. J. Toribio (Eds.)*, 2009. The Cambridge handbook of linguistic code-switching*. Cambridge University Press 2009.

Kutas, M., Van Petten, C. K., & Kluender, R. (2006). Psycholinguistics electrified II (1994– 2005). In *Handbook of psycholinguistics* (pp. 659–724). Academic Press.

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

Lau, E., Stroud, C., Plesch, S., & Phillips, C. (2006). The role of structural prediction in rapid syntactic analysis. *Brain and Language, 98*(1), 74–88.

Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *Journal of Cognitive Neuroscience, 25*(3), 484–502.

Lee, C. L., & Federmeier, K. D. (2006). To mind the mind: An event-related potential study of word class and semantic ambiguity. *Brain Research, 1081*(1), 191–202.

Lee, C. L., & Federmeier, K. D. (2009). Wave-ering: An ERP study of syntactic and semantic context effects on ambiguity resolution for noun/verb homographs. *Journal of Memory and Language, 61*(4), 538–555.

Lee, C. L., & Federmeier, K. D. (2012). Ambiguity's aftermath: How age differences in resolving lexical ambiguity affect subsequent comprehension. *Neuropsychologia, 50* (5), 869–879.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2020). emmeans: Estimated marginal means. R package version 1.4. 4. *The American Statistician, 34*(4), 216–221.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177.

Li, P. (1996). Spoken word recognition of code-switched words by Chinese–English bilinguals. *Journal of Memory and Language, 35*(6), 757–774.

Liao, C. H., & Chan, S. H. (2016). Direction matters: Event-related brain potentials reflect extra processing costs in switching from the dominant to the less dominant language. *Journal of Neurolinguistics, 40*, 79–97.

Litcofsky, K. A., & Van Hell, J. G. (2017). Switching direction affects switching costs: Behavioral, ERP and time-frequency analyses of intra-sentential codeswitching. *Neuropsychologia, 97*, 112–139.

Loebell, H., & Bock, K. (2003). Structural priming across languages. *Linguistics, 41*, 791–824 (5; ISSU 387).

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience, 8*, 213.

Luck, S. J. (2005). *Ten simple rules for designing ERP experiments. Event-related potentials: A methods handbook*, 262083337.

Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT press.

Macnamara, J., & Kushnir, S. L. (1971). Linguistic independence of bilinguals: The input switch. *Journal of Verbal Learning and Verbal Behavior, 10*(5), 480–487.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research, 50*, 940–967.

Marian, V., & Spivey, M. (2003). Competing activation in bilingual language processing: Within- and between-language competition. *Bilingualism: Language and Cognition, 6* (2), 97–115.

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

Martin, C. D., Thierry, G., Kuipers, J. R., Boutonnet, B., Foucart, A., & Costa, A. (2013). Bilinguals reading in their second language do not predict upcoming words as native readers do. *Journal of Memory and Language, 69*(4), 574–588.

Midgley, K. J., Holcomb, P. J., & Grainger, J. (2009). Masked repetition and translation priming in second language learners: A window on the time-course of form and meaning activation using ERPs. *Psychophysiology, 46*(3), 551–565.

Milroy, L., & Gordon, M. (2008). *Sociolinguistics: Method and interpretation* (vol. 13). John Wiley & Sons.

Milroy, L., & Muysken, P. (1995). Introduction: Code-switching and bilingualism research. In *One speaker, two languages: Cross-disciplinary perspectives on code-switching* (pp. 1–14).

Molinaro, N., Barber, H. A., Caffarra, S., & Carreiras, M. (2015). On the left anterior negativity (LAN): The case of morphosyntactic agreement. *Cortex, 66*, 156–159.

Molinaro, N., Barber, H. A., & Carreiras, M. (2011). Grammatical agreement processing in reading: ERP findings and future directions. *cortex, 47*(8), 908–930.

Moreno, E. M., Federmeier, K. D., & Kutas, M. (2002). Switching languages, switching palabras (words): An electrophysiological study of code switching. *Brain and Language, 80*(2), 188–207.

Moreno, E. M., Rodríguez-Fornells, A., & Laine, M. (2008). Event-related potentials (ERPs) in the study of bilingual language processing. *Journal of Neurolinguistics, 21* (6), 477–508.

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.

Nevins, A., Dillon, B., Malhotra, S., & Phillips, C. (2007). The role of feature-number and feature-type in processing Hindi verb agreement violations. *Brain Research, 1164*, 81–94.

Ng, S., Gonzalez, C., & Wicha, N. Y. (2014). The fox and the Cabra: An ERP analysis of reading code switched nouns and verbs in bilingual short stories. *Brain Research, 1557*, 127–140.

Nieuwland, M. S. (2019). Do "early"brain responses reveal word form prediction during language comprehension? A critical review. *Neuroscience & Biobehavioral Reviews, 96*, 367–400.

Nieuwland, M. S., Martin, A. E., & Carreiras, M. (2013). Event-related brain potential evidence for animacy processing asymmetries during sentence comprehension. *Brain and Language, 126*(2), 151–158.

Nieuwland, M. S., Otten, M., & Van Berkum, J. J. (2007). Who are you talking about? Tracking discourse-level referential processing with event-related brain potentials. *Journal of Cognitive Neuroscience, 19*(2), 228–236.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., … Mézière, D. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *ELife, 7*, Article e33468.

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

Ochshorn, R., & Hawkins, M. (2016). *Gentle, A robust yet lenient forced aligner built on Kaldi* [Computer Software]. https://lowerquality.com/ gentle/.

O'Rourke, P. L., & Van Petten, C. (2011). Morphological agreement at a distance: Dissociation between early and late components of the event-related brain potential. *Brain Research, 1392*, 62–79.

Osterhout, L. (1997). On the brain response to syntactic anomalies: Manipulations of word position and word class reveal individual differences. *Brain and Language, 59* (3), 494–522.

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

Osterhout, L., McLaughlin, J., Kim, A., Greenwald, R., & Inoue, K. (2004). Sentences in the brain: Event-related potentials as real-time reflections of sentence comprehension and language learning. In *The on-line study of sentence comprehension: Eyetracking, ERP, and beyond* (pp. 271–308).

Osterhout, L., & Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language, 34*(6), 739–773.

Otten, M., & Van Berkum, J. J. (2008). Discourse-based word anticipation during language processing: Prediction or priming? *Discourse Processes, 45*(6), 464–496.

Payne, B. R., Lee, C. L., & Federmeier, K. D. (2015). Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials. *Psychophysiology, 52*(11), 1456–1469.

Payne, B. R., Ng, S., Shantz, K., & Federmeier, K. D. (2020). Event-related brain potentials in language processing: The N's and the P's. *Psychology of Learning and Motivation, 72*, 75–118.

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin, 144*(10), 1002.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*(4), 329–347.

Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en Español: Toward a typology of code-switching1. *Linguistics, 18*(7–8), 581–618.

Poplack, S. (2018). Categories of grammar and categories of speech. *Questioning Theoretical Primitives in Linguistic Inquiry: Papers in Honor of Ricardo Otheguy, 76*, 7.

Poulisse, N., & Bongaerts, T. (1994). First language use in second language production. *Applied Linguistics, 15*(1), 36–57.

Proverbio, A. M., Čok, B., & Zani, A. (2002). Electrophysiological measures of language processing in bilinguals. *Journal of Cognitive Neuroscience, 14*(7), 994–1017.

Proverbio, A. M., Leoni, G., & Zani, A. (2004). Language switching mechanisms in simultaneous interpreters: An ERP study. *Neuropsychologia, 42*(12), 1636–1656.

R Core Team. (2020). *R foundation for statistical computing. 2013*. Vienna, Austria: 2014. R: A language and environment for statistical computing.

Roelofs, A., Meyer, A. S., & Levelt, W. J. (1998). A case for the lemma/lexeme distinction in models of speaking: Comment on Caramazza and Miozzo (1997). *Cognition, 69*(2), 219–230.

Royle, P., Drury, J. E., & Steinhauer, K. (2013). ERPs and task effects in the auditory processing of gender agreement and semantics in French. *The Mental Lexicon, 8*(2), 216–244.

Ruigendijk, E., Hentschel, G., & Zeller, J. P. (2016). How L2-learners' brains react to code- switches: An ERP study with Russian learners of German. *Second Language Research, 32*(2), 197–223.

Scarborough, D. L., Gerard, L., & Cortese, C. (1984). Independence of lexical access in bilingual word recognition. *Journal of Verbal Learning and Verbal Behavior, 23*(1), 84–99.

Schwanenflugel, P. J., & LaCount, K. L. (1988). Semantic relatedness and the scope of facilitation for upcoming words in sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(2), 344.

Sebba, M., Mahootian, S., & Jonsson, C. (Eds.).. (2012). *Language mixing and code-switching in writing: Approaches to mixed-language written discourse*. Routledge.

Severens, E., Jansma, B. M., & Hartsuiker, R. J. (2008). Morphophonological influences on the comprehension of subject–verb agreement: An ERP study. *Brain Research, 1228*, 135–144.

Soares, C., & Grosjean, F. (1984). Bilinguals in a monolingual and a bilingual speech mode: The effect on lexical access. *Memory & Cognition, 12*(4), 380–386.

Spivey, M. J., & Marian, V. (1999). Cross talk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science, 10*(3), 281–284.

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language, 82*, 1–17.

Steinhauer, K., & Drury, J. E. (2012). On the early left-anterior negativity (ELAN) in syntax studies. *Brain and Language, 120*(2), 135–162.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*(5217), 1632–1634.

Tanner, D. (2015). *On the left anterior negativity (LAN) in electrophysiological studies of morphosyntactic agreement: A commentary on "Grammatical agreement processing in reading: ERP findings and future directions" by Molinaro et al., 2014*.

Tanner, D., Grey, S., & Van Hell, J. G. (2017). Dissociating retrieval interference and reanalysis in the P600 during sentence comprehension. *Psychophysiology, 54*(2), 248–259.

Tanner, D., & Van Hell, J. G. (2014). ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia, 56*, 289–301.

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

Tokowicz, N., & MacWhinney, B. (2005). Implicit and explicit measures of sensitivity to violations in second language grammar: An event-related potential investigation. *Studies in Second Language Acquisition*, 173–204.

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(3), 443.

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

Van Der Meij, M., Cuetos, F., Carreiras, M., & Barber, H. A. (2011). Electrophysiological correlates of language switching in second language learners. *Psychophysiology, 48*(1), 44–54.

Van Hell, J. G., & Witteman, M. J. (2009). The neurocognition of switching between languages. *Multidisciplinary Approaches to Code Switching, 41*, 53.

Van Hell, J. G., & De Groot, A. M. (1998). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. *Bilingualism: Language and Cognition, 1*(3), 193–211.

Van Hell, J. G., & De Groot, A. M. (2008). Sentence context modulates visual word recognition and translation in bilinguals. *Acta Psychologica, 128*(3), 431–451.

Van Hell, J. G., Fernandez, C. B., Kootstra, G. J., Litcofsky, K. A., & Ting, C. Y. (2018). Electrophysiological and experimental-behavioral approaches to the study of intra-sentential code-switching. *Linguistic Approaches to Bilingualism, 8*(1), 134–161.

Van Hell, J. G., Litcofsky, K. A., & Ting, C. Y. (2015). Intra-sentential code-switching: Cognitive and neural approaches. *The Cambridge Handbook of Bilingual Processing*, 459–482.

Van Hell, J. G., & Tanner, D. (2012). Second language proficiency and cross-language lexical activation. *Language Learning, 62*, 148–171.

Van Petten, C. (1993). A comparison of lexical and sentence-level context effects in event-related potentials. *Language & Cognitive Processes, 8*(4), 485–531.

Van Petten, C. (1995). Words and sentences: Event-related brain potential measures. *Psychophysiology, 32*(6), 511–525.

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(2), 394.

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory & Cognition, 18*(4), 380–393.

Van Petten, C., & Kutas, M. (1991). Influences of semantic and syntactic context on open- and closed-class words. *Memory & Cognition, 19*(1), 95–112.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology, 83*(2), 176–190.

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019). Expectation-based comprehension: Modeling the interaction of world knowledge and linguistic experience. *Discourse Processes, 56*(3), 229–255.

Weber-Fox, C. M., & Neville, H. J. (1996). Maturational constraints on functional specializations for language processing: ERP and behavioral evidence in bilingual speakers. *Journal of Cognitive Neuroscience, 8*(3), 231–256.

Wicha, N. Y., Bates, E. A., Moreno, E. M., & Kutas, M. (2003). Potato not pope: Human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters, 346*(3), 165–168.

Wicha, N. Y., Moreno, E. M., & Kutas, M. (2003). Expecting gender: An event related brain potential study on the role of grammatical gender in comprehending a line drawing within a written sentence in Spanish. *Cortex, 39*(3), 483–508.

Wicha, N. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience, 16*(7), 1272–1288.

Wlotko, E. W., & Federmeier, K. D. (2015). Time for prediction? The effect of presentation rate on predictive sentence comprehension during word-by-word reading. *Cortex, 68*, 20–32.

Zeller, J. P. (2020). Code-switching does not equal code-switching. An event-related potentials study on switching from L2 German to L1 Russian at prepositions and nouns. *Frontiers in Psychology*, 11.

Zeller, J. P., Hentschel, G., & Ruigendijk, E. (2016). Psycholinguistic aspects of Belarusian-Russian language contact. An ERP study on code-switching between closely related languages. *Slavic Languages in Psycholinguistics. Chances and Challenges for Empirical and Experimental Research*, 257–278.