Brain-based Individual Differences in On-line L2 Sentence Comprehension

Darren Tanner
Pennsylvania State University

Kayo Inoue
University of Washington

Lee Osterhout
University of Washington

Author Note

Darren Tanner, Center for Language Science, Department of Psychology, Pennsylvania State University; Kayo Inoue, Integrated Brain Imaging Center, Department of Radiology, School of Medicine, University of Washington; Lee Osterhout, Department of Psychology, University of Washington.

This research was supported by NIDCD grant R01DC01947 to Lee Osterhout and NSF BCS-0951595 to Lee Osterhout and Darren Tanner. Darren Tanner also received support from the William Orr Dingwall Neurolinguistics Dissertation Fellowship.

Correspondence regarding this article should be addressed to Darren Tanner, Center for Language Science, Pennsylvania State University, 4F Thomas Building, University Park, PA 16802. Email dstanner@gmail.com
Abstract

We investigated individual differences in brain response profiles of proficient L1 Spanish-L2 English bilinguals during morphosyntactic processing. Event-related potentials (ERPs) were recorded while participants read English sentences, some of which contained a subject-verb agreement anomaly. Grand-mean ERP analyses revealed a biphasic N400-P600 response to agreement violations. However, subsequent correlation analyses showed that participants’ brain responses varied along a continuum between N400-dominant and P600-dominant responses. We further investigated this pattern by quantifying two aspects of individuals’ brain responses: overall response magnitude and relative response dominance (e.g., N400 or P600). Multiple regression results showed that larger brain response magnitude was associated with higher L2 proficiency, while the type of brain response was most strongly associated with individuals’ age of arrival in the English-speaking country and motivation to learn English. Moreover, approximately 61% of the variance in brain response type was accounted for by a coalition of five variables known from behavioral studies to impact L2 learning. We interpret our results in the context of cue-based sentence processing models. Additionally, by computing two new outcome measures for ERP analysis we also show that different aspects of the ERP signal (timing, amplitude, and polarity) can be differentially related aspects of learners’ backgrounds.

Keywords: ERP, N400, P600, individual differences, morphosyntax, second language acquisition
Brain-based Individual Differences in On-line L2 Sentence Comprehension

In an interconnected global society, knowledge of a second language (L2) is an increasingly indispensable skill. Growing globalization in trade, education, and politics requires large numbers of individuals with strong L2 skills, and increasing international immigration is creating a large population of individuals who find themselves needing to master a nonnative language rapidly. However, most people find learning an L2 after early childhood to be a difficult and laborious process, and most learners never achieve full mastery of subtle aspects of L2 grammar (Johnson & Newport, 1989). Nonetheless, there is a great deal of individual variation in the rate and ultimate success of L2 learning. Understanding what factors are associated with successful learning can therefore be of interest both to cognitive psychologists and applied researchers interested in identifying cognitive skills or strategies underlying learning, identifying gifted language learners for selection in language training programs, or more generally improving learner outcomes. Indeed, some research has sought to characterize predictive cognitive factors or learner strategies associated with rapid and successful learning (e.g., Carroll, 1962; Naiman, Fröhlich, Stern, & Todesco, 1996; Skehan, 1989). Other studies focusing on experiential factors associated with individuals’ long-term learning outcomes have shown that greater success is associated with early immersion, higher motivation to learn, and more frequent L2 use in daily life (e.g., Flege, Yeni-Komshian, & Liu, 1999).

These studies have largely relied on offline measures such as learners’ performance on large test batteries, or subjective measures such as teachers’ ratings of learner development over time. However, little is known about what factors relate to individual differences in the neural substrates of either first language (L1) or L2 real-time language comprehension. Methodologies like event-related brain potentials (ERPs) allow researchers to investigate the neural mechanisms
of language comprehension with millisecond-level temporal resolution. ERPs provide a measure of individuals’ brainwave activity that is time- and phase-locked to the presentation of stimuli, such as words in sentences. In studies of L1 sentence processing, ERPs have shown a neurocognitive dissociation between the processing of lexico-semantic and morphosyntactic anomalies. Lexico-semantic manipulations typically trigger an enhanced negativity, peaking around 400ms poststimulus (the N400 effect: Bentin, 1987; Kutas & Hillyard, 1980, 1984; Osterhout & Nicol, 1999), whereas violations of morphosyntactic rules typically elicit a large positive-going wave with a maximum around 600ms poststimulus (the P600 effect: Allen, Badecker, & Osterhout, 2003; Friederici, Hahne, & Mecklinger, 1996; Kaan & Swaab, 2003; Molinaro, Kim, Vespignani, & Job, 2008; Osterhout & Holcomb, 1992; Osterhout & Mobley, 1995).\(^1\) The N400 and P600 effects can be seen as indexes of two independent, but highly interactive ‘streams’ of processing which are differentially sensitive to linguistic cues (Kim & Osterhout, 2005; Kuperberg, 2007; Osterhout, Kim, & Kuperberg, 2012; see also Jackendoff, 2007; MacWhinney, Bates, & Kleigl., 1984; van de Meerendonk, Kolk, Vissers, & Chwillia, 2010).

ERPs’ multidimensional nature can also be useful in characterizing the cognitive mechanisms underlying L2 comprehension. Over the last decade there has been an enormous growth of interest in the neurocognitive substrates of L2 learning and processing (see McLaughlin et al., 2010; Mueller, 2005; Osterhout, McLaughlin, Pitkänen, Frenck-Mestre, & Molinaro, 2006; Steinhauer, White, & Drury, 2009; van Hell & Tokowicz, 2010 for reviews), and much of this research has focused on whether L2 syntactic comprehension is fundamentally different from or similar to native language processing. One of the driving questions in this line of research is therefore whether L2 learners can show P600 effects to L2 syntactic violations.
There is now growing consensus that there can be strong continuity between L1 and L2 processing systems, as P600 effects have been observed in relatively low proficiency L2 learners processing violations of syntactic rules common to the L1 and L2 (e.g., McLaughlin et al., 2010; Rossi, Gugler, Friederici, & Hahne, 2006; Tokowicz & MacWhinney, 2005), as well as high proficiency learners processing novel L2 features (e.g., Frenck-Mestre, Foucart, Carrasco, & Herschensohn, 2009; Gillon Dowens, T. Guo, J. Guo, Barber, & Carreiras, 2011; Gillon Dowens, Vergara, Barber, & Carreiras, 2010; Morgan-Short, Sanz, Steinhauer, & Ullman, 2010).2

However, most published ERP waveforms represent the central tendency after averaging across both trials and subjects, and therefore may not depict any individual's brain response to a stimulus on a particular trial. Given the findings of marked and systematic individual variation in learning rate, style, and success found in behavioral batteries mentioned above, it seems likely that learners’ brain responses during online comprehension may also show systematic differences. Relatively little research has used ERPs to study individual differences in either L1 or L2 processing. Some studies have used grouped designs, where groups are determined by splits on background measures (e.g., working memory span, comprehension performance, L2 proficiency, or age of arrival: King & Kutas, 1995; Rossi et al., 2006; Vos, Gunter, Kolk, & Mulder, 2001; Weber-Fox & Neville, 1996; Weckerly & Kutas, 1999), though this approach provides little information as to the nature of the relationship between the background and outcome measures (e.g., ERP amplitude, latency, or polarity), which may be linear and graded.

One possibility for overcoming this limitation is to apply correlation and regression techniques to the study of language ERPs, but only very recently have these methods begun to be used. Some studies have focused on individual differences in L2 lexico-semantic processing by measuring correlates of N400 amplitude and latency. Moreno & Kutas (2005) used multiple
regression to investigate N400 peak latency in Spanish-English bilinguals processing semantic anomalies. Results showed that age of acquisition and vocabulary proficiency each predicted unique variance in effect latency, with earlier and more proficient bilinguals showing earlier N400 effect peaks. Newman and colleagues (Newman, Tremblay, Nichols, Neville, & Ullman, 2012) used linear mixed models to investigate effects of proficiency on semantic processing in L1 and late L1 Spanish-L2 English speakers. They argued that individual differences in N400 amplitude can be accounted for by language proficiency, while N400 onset may be more fundamentally related to L1-L2 processing differences. Individual differences in L1 syntactic processing have also recently been reported. Pakulak & Neville (2010) showed that the laterality of an early negative component and the amplitude of the P600 component correlated linearly with native English speakers’ proficiency in their L1.

Other studies have shown that not only can the amplitude or latency of ERPs vary systematically across individuals, but also the type (N400 vs. P600). Nakano and colleagues (Nakano, Saron, & Swaab, 2010) showed that brain response type to verb-argument animacy violations correlated with individuals’ working memory span. N400 effect magnitudes were negatively correlated with speaking span measures, while P600 effect magnitudes were positively correlated with span measures. Inoue and Osterhout (in prep; see also Inoue & Osterhout, 2007; Osterhout, McLaughlin, Kim, Greewald, & Inoue, 2004) investigated the processing of anomalous case markings in L1 Japanese speakers. Their grand mean results showed a biphasic N400-P600 response, while further investigation revealed that this was an artifact of averaging over individuals, some of whom showed an N400 and some of whom showed a P600 (see also Nieuwland & Van Berkum, 2008; Osterhout, 1997 for similar findings). Moreover, within and across individuals, N400 and P600 effect magnitudes were negatively
correlated, such that as one increases in magnitude, the other decreases to a similar degree (Inoue & Osterhout, in prep). Tanner, McLaughlin, Herschensohn, & Osterhout (accepted) showed that a similar function characterized brain responses to subject-verb agreement violations in novice L1 English learners of L2 German. Further correlational analyses revealed that larger N400 and P600 effects were associated with poorer and better performance on a grammaticality judgment task, respectively. This suggests that relative reliance on a lexico-semantic or combinatorial-syntactic stream when processing inflectional anomalies can be associated with different stages of grammatical acquisition in beginning learners (see also McLaughlin et al., 2010).

What individual factors are associated with morphosyntactic processing profiles in proficient L2 learners? Here we investigate individual-level correlates of grammatical processing in late L1 Spanish-L2 English bilinguals with long-term immersion in an L2 English environment. We focus on ERP correlates of processing subject-verb agreement, a rule shared by both Spanish and English. Agreement is well-studied using ERPs in both L1 and L2 populations, and has shown to be useful in identifying stages of L2 learning (McLaughlin et al., 2010; Tanner et al., accepted; see Molinaro, Barber, & Carreiras, 2011 for a review of L1 studies on agreement processing). Based on previous L2 research we predicted that agreement violations would elicit large P600 effects in our highly experienced learners. However, results showed a great deal of variability in both the type and magnitude of learners’ brain responses. We investigated this variability by assessing the impact of several factors which are either known from behavioral studies to impact long-term learning outcomes or which have been shown to impact language processing: age of arrival, length of residence, frequency of L2 use, language proficiency, and learner motivation (Birdsong, 2006; Birdsong & Molis, 2001; Dörnyei, 2005; Dörnyei & Skehan, 2003; Flege et al., 1999; Moreno & Kutas, 2005; Newman et al., 2012; Pakulak & Neville, 2010).
Multiple regression results showed that a majority of variability in brain responses was systematically associated with these factors. Additionally, by computing two new outcome measures for ERP analysis we also show that different aspects of the ERP signal (timing, amplitude, and polarity) can be differentially related aspects of learners’ backgrounds.

**Method**

**Participants**

Our participants included 20 native Spanish speakers (7 male) who had acquired English as an L2. All participants were strongly right-handed as assessed by an abridged version of the Edinburgh Handedness Inventory, and had normal or corrected-to-normal vision. Participants were screened such that they had not been exposed to English in the home, had first moved to an English-speaking country at age 15 or later, and had lived immersed in an English-speaking environment for a minimum of five years. Participants completed a language background questionnaire, which included self-reports on age of initial exposure to English (AoE), age of arrival in an English-speaking environment (AoA), and total length of residence in an English-speaking environment (LoR), as well as proficiency self-ratings for their L1 Spanish and L2 English on a Likert-scale between 1 (no proficiency) and 7 (perfect proficiency). Other questions asked participants about their frequency of use of English in various contexts in daily life, an overall estimate of English use between 1 (never use English) and 7 (always use English), as well as their motivation to speak English like a native speaker between 1 (not important to sound like a native speaker at all) and 7 (extremely important to sound like a native speaker). Participants also completed a pen-and-paper proficiency test consisting of 50 questions selected from the Michigan Examination for the Certificate of Proficiency in English (ECPE).
Participants’ responses are reported in Table 1. Participants provided informed consent and received a small amount of cash for taking part.

---------------------
Insert Table 1 about here
---------------------

Materials

Stimuli were sentences that were either grammatically correct or contained a violation of subject-verb agreement. All sentences contained a singular subject noun, followed by a prepositional phrase modifier containing another singular noun, followed by an auxiliary verb that either agreed or disagreed with the subject noun in number (is/are, was/were, or has/have), followed by a short predicate (e.g., The winner of the big trophy has/*have proud parents). Two hundred forty sentence frames were constructed with eight versions of each sentence. Two versions corresponded to the grammatical-ungrammatical sentence pairs reported here; the other six versions investigated agreement in more complex syntactic configurations (see Tanner, 2011; Tanner, Nicol, Herschensohn, & Osterhout, 2012). Sentences were distributed across eight experimental lists in a Latin square design, such that each listed contained only one version of each sentence frame. Each list thus contained 30 grammatical and 30 ungrammatical sentences in the experiment reported here. Experimental sentences were pseudo-randomized among other sentences belonging to other experiments not reported here. Each list contained 540 sentences, half of which contained either a syntactic or semantic anomaly.

Procedure

Participants took part in three sessions. The first session involved completion of the background questionnaires and proficiency test; ERPs were recorded during the two subsequent sessions. Because of the large number of experimental sentences, participants saw half of the
sentences at each session. ERPs were first averaged within sessions. No between-session differences were found, so ERPs were subsequently averaged across sessions. All reported results reflect these cross-session averages. During ERP recording participants were seated in a comfortable recliner in front of a CRT monitor. Participants were instructed to relax and minimize movements while reading and to read each sentence as normally as possible. Each trial consisted of the following events: each sentence was preceded by a blank screen for 1000ms, followed by a fixation cross, followed by a stimulus sentence, presented one word at a time. The fixation cross and each word appeared on the screen for 400ms followed by a 200ms ISI. Sentence-ending words appeared with a full stop followed by a “yes/no” prompt asking for a sentence acceptability judgment. Participants were instructed to respond “yes” to sentences that were well-formed and semantically coherent and “no” to sentences that were ungrammatical or semantically incoherent. Participants were randomly assigned to use either their left or right hand for the “yes” response.

**Data Acquisition and Analysis**

Continuous EEG was recorded from 19 tin electrodes attached to an elastic cap (Electro-cap International) in accordance with the 10-20 system (Jasper, 1958). Eye movements and blinks were monitored by two electrodes, one placed beneath the left eye and one placed to the right of the right eye. Electrodes were referenced to an electrode placed over the left mastoid. EEG was also recorded from an electrode placed on the right mastoid to determine of there were experimental effects detectable on the mastoids. No such effects were found. EEG signals were amplified with a bandpass of 0.01-100Hz (3db cutoff) by an SAI bioamplifier system. ERPs were filtered off-line below 15hz. Impedances at scalp and mastoid electrodes were held below 5 kΩ and below 15 kΩ at eye electrodes.
Continuous analog-to-digital conversion of the EEG and stimulus trigger codes was performed at a sampling frequency of 250Hz. ERPs, time-locked to the onset of the critical verb, were averaged off-line for each participant at each electrode site in each condition. Trials characterized by eye blinks, excessive muscle artifact, or amplifier blocking were not included in the averages; 5.8% and 5.5% of grammatical and ungrammatical trials were rejected, respectively. ERPs were quantified as mean amplitude within a given time window. Because of a small amount of noise in the prestimulus interval, a 50ms prestimulus to 50ms poststimulus baseline was used. In accordance with previous literature and visual inspection of the grand mean waveforms, the time windows 400-500ms and 500-1000ms were chosen, as they correspond roughly to the N400 and P600 effects, respectively. For grand mean analyses, ANOVAs were calculated within each time window with grammaticality (grammatical, ungrammatical) as a within-subjects factor. Data from midline (Fz, Cz, Pz), medial-lateral (right hemisphere: Fp2, F4, C4, P4, O2; left hemisphere: Fp1, F3, C3, P3, O1), and lateral-lateral (right hemisphere: F8, T8, P8; left hemisphere: F7, T7, P7) electrode sites were treated separately in order to identify topographic and hemispheric differences. ANOVAs on midline electrodes included electrode as an additional within-subjects factor (3 levels), ANOVAs on medial-lateral electrodes included hemisphere (2 levels) and electrode pair (5 levels) as additional within-subjects factors, and ANOVAs over lateral-lateral electrodes included hemisphere (2 levels) and electrode pair (3 levels) as additional within-subjects factors. The Greenhouse-Geisser correction for inhomogeneity of variance was applied to all repeated measures on ERP data with greater than one degree of freedom in the numerator. In such cases, the corrected p-value is reported. Regression models will be described with the results.

**Results**
Grand Mean Results

Grand mean ERP results for grammatical and ungrammatical verbs are depicted in Figure 1. Visual inspection of the waveforms showed that, relative to grammatical verbs, ungrammatical verbs elicited a small biphasic waveform characterized by a centrally-distributed negativity between approximately 400-500ms poststimulus (N400), followed by a broadly-distributed positivity (P600). Statistical analysis confirmed these observations. In the 400-500ms window there was an effect of grammaticality that was strongest over central electrodes [grammaticality x electrode interaction: midline, F(2, 38) = 4.967, p < .03; medial-lateral, F(4, 76) = 4.721, p < .02], and which showed a stronger left-hemisphere distribution over lateral-lateral sites [grammaticality x hemisphere interaction, F(1, 19) = 9.190, p < .01]. Between 500-1000ms the positivity reached significance over a broad portion of the scalp [main effect of grammaticality: midline, F(1, 19) = 8.719, p < .01; medial-lateral: F(1, 19) = 10.292, p < .01; lateral-lateral: F(1, 19) = 7.519, p < .02]. Over lateral-lateral sites the positivity showed a stronger posterior [grammaticality x electrode interaction: F(2, 38) = 4.299, p < .03] and right-hemisphere distribution [grammaticality x hemisphere interaction: F(1, 19) = 10.776, p < .01].

Individual Differences Analyses

Despite the significant biphasic N400-P600 results in the omnibus ANOVA, inspection of individuals’ ERP waveforms showed that most individuals showed primarily either an N400 or P600 to ungrammatical verbs, but not both. Following Inoue & Osterhout (in prep) and Tanner et al. (accepted), we regressed individuals' N400 effect magnitude averaged over three midline electrodes onto their P600 effect magnitude and showed them to be highly negatively
correlated, $r = -0.777, p < .0001$, Figure 2. Brain responses showed a continuous distribution across individuals such that as one effect increased, the other decreased to a similar degree. Figure 3 shows effects in those who showed primarily a negativity or positivity to agreement violations separately. Those in the negativity-dominant group showed a large, centrally-distributed negativity between 400-500ms (an N400), but no effects between 500-1000ms, while those in the positivity-dominant group showed no effects between 400-500ms and a large, posteriorly distributed positivity between 500-1000ms (a P600). Thus, grand mean waveforms depicted an effect that was not representative of most individuals’ ERP profiles.

To investigate what factors might determine the type and magnitude of ERP response to agreement violations in this group of learners, we computed two new measures to serve as the dependent variables (DVs) in multiple regression models. For the first measure, the Response Magnitude Index (RMI), we calculated the summed ungrammaticality effect across both the N400 and P600 time windows (i.e., an individual’s Euclidian distance from zero in Figure 2). Larger values of the RMI indicate relatively greater neural responses to agreement violations across both time windows, regardless of the type of response. For the second measure we subtracted the N400 effect magnitude from the P600 effect magnitude for each individual. This measure, the Response Dominance Index (RDI), gives an indication of the relative dominance (N400 or P600) of an individual’s anomaly response. Values near zero reflect relatively equal-sized P600 and N400 effects, while more positive and negative values reflect relatively more dominant positive or negative effects across both time windows, respectively. Equations (1) and
(2) show how the RMI and RDI were computed, where N400 and P600 refer to mean amplitude between 400-500ms and 500-1000ms, respectively, averaged over midline electrodes.

\[
RMI = \sqrt{(N400_{Gram} - N400_{Ungram})^2 + (P600_{Ungram} - P600_{Gram})^2} \quad (1)
\]

\[
RDI = (P600_{Ungram} - P600_{Gram}) - (N400_{Gram} - N400_{Ungram}) \quad (2)
\]

We fit two multiple regression models with the RMI and RDI serving as the respective DVs. Five background measures known to affect learning outcomes and processing profiles were selected as independent variables for the models: AoA, LoR, frequency of L2 use, L2 proficiency (ECPE scores), and motivation (see Table 1). Because of right-skewed distributions, AoA and LoR were log-transformed prior to entry into the regression models. Additionally, since all participants reported having relatively high motivation to speak like a native English speaker, this measure was dichotomized into those reporting very high motivation (i.e., those reporting 7 on the Likert-scale, n=11) and those reporting high motivation (i.e., those reporting 5 or 6 on the Likert-scale, n=9). ‘High’ motivation set as the reference level (0) and ‘very high’ motivation set to 1 in the models.4 Distributional statistics for the DVs and background measures entered into the models are reported in Table 2; the correlation matrix for the measures is reported in Table 3. Collinearity diagnostics showed that multicollinearity among predictor variables was not a problem (tolerances > .7, VIFs < 1.5). Residuals of both models were approximately normally distributed.

-------------
Insert Table 2 about here
-------------
-------------
Insert Table 3 about here
-------------

The total model predicting the RMI was not significant ($R^2 = .418$, Adjusted $R^2 = .210$, $F(5, 14) = 2.008, p = .14$), though one individual predictor reached significance. After
controlling for the effects of AoA, LoR, frequency of use, and motivation, English proficiency showed a positive association with the RMI (partial-r = .490, p = .05; at electrode site Cz, where N400 and P600 effects intersected most strongly, partial-r = .504, p < .05). This shows that those who were more proficient in an offline pen-and-paper test also showed larger overall neural sensitivity to agreement violations during online processing. The model predicting brain response dominance was highly significant (Table 4), such that the linear combination of AoA, LoR, frequency of use, L2 proficiency and motivation accounted for approximately 61% of the variance in response dominance. Two individual predictors also reached significance. AoA and motivation each uniquely predicted response dominance: those who experienced earlier immersion in an English-speaking environment and who reported higher motivation to sound like a native speaker showed stronger P600-dominant brain responses, while those who arrived later and reported lower motivation were more likely to show N400-dominant responses.

Discussion

We investigated individual differences in neural activity to violations of subject-verb agreement in late, but highly experienced L1 Spanish immigrants to an L2 English environment. Although grand-mean analyses showed a reliable biphasic N400-P600 pattern, individual analyses showed that this pattern was not representative of most individuals' brain responses. Individuals’ brain responses varied along a continuum between N400 and P600 effects, with most participants showing a dominance in one response or the other. In addition, multivariate analyses showed that a large portion of this between-subject variability was systematic. By differentially quantifying the type and magnitude of brain responses, we showed that different aspects of the ERP signal (e.g., timing, amplitude, and polarity) were associated with different
aspects of a learner’s background. Earlier age of arrival and higher motivation to master English were associated with greater dominance of the P600 effect, relative to the N400 effect. Proficiency, however, had impact on relative response dominance. Higher proficiency was instead associated with a larger overall brain response, regardless of the type. Additionally, the linear combination of age of arrival, length of residence, frequency of use, English proficiency, and motivation accounted for a majority of the overall variance in the type of brain response individuals’ showed to agreement violations. It is important to note that we found these partial correlations and multivariate effects despite a relatively small sample size. This indicates exceptionally strong and systematic relationships, but also introduces a cautionary note in interpreting the null effects seen in the regression analyses.

Our findings add to a growing number of L2 studies reporting N400 responses to syntactic violations and extend these findings in important ways. Several longitudinal studies have reported that syntactic anomalies elicit N400 effects in early-stage L2 learners, but elicit P600 effects after increased L2 instruction (McLaughlin et al., 2010; Morgan-Short et al., 2010; Morgan-Short, Steinhauer, Sanz, & Ullman, 2012; Osterhout et al., 2006; Pitkänen, Tanner, McLaughlin, & Osterhout, in prep). Furthermore, Tanner and colleagues (Tanner et al., accepted) report cross-sectional results from novice L2 learners who show a nearly identical distribution of brain responses to that reported here. In most of these studies, N400 effects are associated with the earliest stages of learning and are replaced by P600 effects within a single year of classroom L2 instruction. However, in the present study N400 effects to L2 agreement violations persisted in some of the bilinguals, despite long-term L2 immersion and relatively high L2 proficiency. There is therefore a seeming inconsistency between the current results and reports of low- and intermediate-proficiency classroom-educated L2 learners showing robust P600 (and no N400)
effects to inflectional anomalies (e.g., Bond, Gabriele, Fiorentino, & Alemán Bañón, 2011; Foucart & Frenck-Mestre, 2011; Frenck-Mestre, Osterhout, McLaughlin, & Foucart, 2008; McLaughlin et al., 2010; Tokowicz & MacWhinney, 2005). Since N400 effects to syntactic violations have been associated with the earliest stage of grammatical learning, it is conceivable that the N400-dominant learners in the current study may have somehow fossilized at an earlier stage of acquisition. N400 effects in late-stage bilinguals might therefore represent instances of poor L2 learning, at least at a neurocognitive level. The significant positive partial correlations between the response dominance index and learners’ age of arrival and motivation to learn English are consistent with this hypothesis, as behavioral research has long noted relationships between age of arrival, learner motivation, and long-term learning outcomes (Birdsong & Molis, 2001; Dörnyei, 2005; Flege et al., 1999; Johnson & Newport, 1989).

However, a broader view of the relevant literature raises doubts about the “fossilization” interpretation of the N400 effect. Syntactic violations have sometimes elicited N400 or biphasic N400-P600 responses similar to those reported here in L1 sentence processing studies (e.g., Deutsch & Bentin, 2001; Mancini, Molinaro, Rizzi, & Carreiras, 2011; Severens, Jansma, & Hartsuiker, 2008). In some studies biphasic responses have been shown to be artifacts resulting from averaging over individual differences (Inoue & Osterhout, in prep; Nieuwland & Van Berkum, 2008; Osterhout, 1997; Osterhout et al., 2004; see also Nakano, et al., 2010). Moreover, the “response dominance continuum” reported here in bilinguals has also been observed in studies of L1 syntactic processing involving Japanese- and English-speaking adults (Inoue & Osterhout, in prep; Tanner, 2011). Collectively, these results indicate that the individual differences reported here exist in early-stage L2 learners, later-stage bilinguals, and native speakers. Although age and motivation were implicated in predicting brain response type in the
current study, explanations for individual differences in response dominance must therefore move beyond these usual suspects in accounting for the similar pattern across L1 and L2 populations.

Individual differences in adult sentence comprehension have been linked to variation in working memory capacity, neural efficiency, and cortical connectivity, among other variables (e.g., Bates, Devescovi, & Wulfeck, 2001; Just & Carpenter, 1992; Prat, 2011; Prat, Keller, & Just, 2007). Although these and other possibilities should be explored further, the sentence processing models proposed by MacWhinney, Bates and colleagues (e.g., MacWhinney et al., 1984) and Ferreira, Christianson and colleagues (e.g., Christianson, Hollingworth, Halliwell, & Ferreira, 2001; Ferreira, 2003; Ferreira, Bailey, & Ferraro, 2002) might prove to be particularly efficacious for explaining the results reported here. Central to the MacWhinney and Bates model is the notion of cue strength, that is, the reliability of the linguistic cues that indicate the mapping between linguistic form and linguistic meaning. Languages differ in this regard (MacWhinney et al., 1984). Ferreira, Christianson and colleagues argue that, in certain cases, comprehenders fail to compute detailed syntactic representations of incoming linguistic information. In these circumstances individuals may instead rely on shallower ‘good enough’ processing heuristics. For example, in complex syntactic configurations such as passives or garden path sentences, individuals may place greater weight on lexical or semantic cues than the syntactic cues, especially when the two sets of cues conflict (Ferreira, 2003; see also Kim & Osterhout, 2005; Kuperberg, 2007; Osterhout et al., 2012)

Our results suggest that cue strength and reliance on ‘good enough’ processing heuristics may differ across individuals as well. Some individuals seem to rely more heavily on lexical information (i.e., they are N400-dominant) and others on syntactic information during processing
(i.e., they are P600-dominant). Under this interpretation, the response dominance index represents a metric of cue strength, a value that can vary across individuals at all levels of language proficiency. For morphosyntactic dependencies like agreement, the relative weighting of the cues (i.e., response dominance) seems to be especially malleable in many early-stage L2 learners, such that additional instruction can produce a within-learner shift from an N400 response to a P600 response to the same syntactic anomalies (McLaughlin et al., 2010). Presumably this reflects the gradual acquisition of a grammatical rule and subsequent reduced reliance on lexical information for dealing with that particular aspect of the L2. In L1 adults and later-stage L2 learners, response dominance may reflect more stable processing biases, although this interpretation has yet to be verified empirically.

From this perspective, between-subject variability in the response-dominance continuum would be minimized by narrowing the set of relevant cues and maximized by broadening the set of cues. For example, the syntactic violations in the current study were realized with suppletive lexical alternations between short, high-frequency verbs (e.g., is/are, has/have). Those showing N400-dominant effects may have relied upon lexical form-based expectations of verbal agreement features, while those showing P600-dominant responses may have been more sensitive to the syntactic features carried on the verbs. The availability of multiple cues to subject-verb agreement violations (wordform- versus syntactic feature-based) may allow for variability in the exact cue that individuals attend to. Conversely, in contexts where syntactic relationships are marked morphologically, the lexical cue to the agreement dependency would be weakened and violations should be more likely to elicit P600 effects across all individuals. One potential test of this hypothesis is provided by Foucart and Frenck-Mestre (2012) who investigated grammatical gender processing in L1 English learners of L2 French. In L2 learners,
gender agreement violations elicited P600 effects when the anomaly was detectable on adjectives, which are morphologically marked for gender, but an N400 effect when the anomaly was detectable on nouns, where gender is encoded lexically and not morphologically. Indeed, this may explain the discrepancy between the current findings and other L2 studies which have found P600 (but no N400) effects to morphosyntactic anomalies in learners across the proficiency spectrum. These studies have generally used grammatical dependencies marked by decomposable inflectional morphemes (e.g., Frenck-Mestre et al., 2008; McLaughlin et al., 2010; Rossi et al., 2006; Sabourin & Stowe, 2008; Tokowicz & MacWhinney, 2005; see also Hahne, Mueller, & Clahsen, 2006). Our findings suggest that later and less motivated learners may prioritize lexical cues to syntactic relationships when they are available. However, given that P600 effects are readily found in L2 learners across the proficiency spectrum, the broader L2 processing literature suggests that late learners are in no way restricted to such cues (contra, e.g., Clahsen & Felser, 2006). Instead, our findings are consistent with the notion that individuals may modulate their relative reliance on the lexical/semantic or morphosyntactic processing streams based on the cues provided by the target stimulus (see Tanner, 2011 for further evidence in this regard).

Studies of L1 processing further corroborate this proposal. For example, syntactically anomalous function words (which play a grammatical role but have little referential meaning) elicit P600 effects in most participants. By contrast, syntactically anomalous content words (which have rich referential meanings as well as grammatical roles) can, at least under certain circumstances, elicit robust P600 effects in some subjects and N400 effects in others (Osterhout, 1997). Similarly, anomalous Japanese case markers, which simultaneously indicate the thematic and grammatical roles of each noun, elicit robust P600 effects in some Japanese-speaking adults
and N400 effects in others (Inoue & Osterhout, in prep; Osterhout et al., 2004). An important implication of all of these results is that the grand-mean brain response observed in any given experiment will be a function of not only the manipulated properties of the stimulus (e.g., its linguistic category, morphological complexity, well-formedness) but also systematic properties of the participants (e.g., each subject’s weighting of the relevant linguistic cues), and an interaction between the two. Repeated sampling from a population would by chance produce different outcomes. Collectively, these results illustrate the dangers inherent in the exclusive use of grand-average ERPs to characterize L2 sentence processing. In some cases (as in the present study), a thorough investigation of between-learner variability can be more informative than inspection of grand mean waveforms.

In all of these experiments (including the present study), we have observed a robust negative correlation between N400 and P600 magnitudes across individuals: as the amplitude of one effect increased in size, the amplitude of the other effect decreased to a similar extent (Fig. 2). We have observed the negative correlation in L2 learners and in native speakers (Inoue & Osterhout, in prep; Tanner, 2011; Tanner et al., accepted; see also Nakano et al, 2010; Nieuwland & Van Berkum, 2005; Osterhout 1997). Conceivably, the negative correlation manifests a dynamic interaction between negatively correlated processing mechanisms involved in different aspects of sentence comprehension (for example, lexico-semantic and grammatical aspects). Evidence to support this notion has been acquired recently in an experiment that combined an ERP linguistic anomaly paradigm with distributed source analysis (Inoue & Osterhout, in prep). This experiment demonstrated a negative correlation in activity within anterior and posterior regions of the left-hemisphere perisylvian cortex when participants read well-formed sentences. Individual differences in the “balance of activation” within these areas
when reading well-formed sentences predicted individual differences in the ERP response to anomalous versions of those sentences, providing a neurobiological explanation for the individual differences.

A final implication of the present findings follows from the observation that our two ERP indices, the response magnitude index and the response dominance index, are uncorrelated. That is, variation in the quantity of the ERP response to linguistic anomalies is uncorrelated with the quality of the response. However, we have shown here that the variability in each of these measures independently can be systematically associated with subject-level covariates. It remains to be seen how some of the individual variables mentioned above (e.g., working memory, neural efficiency, cortical connectivity, etc) map onto the neurocognitive correlates of language processing that we report here (see Bond et al., 2011, for a first attempt to link individuals' specific language aptitude and non-verbal reasoning ability with L2 ERP effects). Although further research is needed to better understand these relationships, the present study provides yet another indication of the unique sensitivities of event-related potentials as tools for studying language learning.
References


Publishing.


Pitkänen, I., Tanner, D., McLaughlin, J., & Osterhout, L. (in prep). Use it or lose it: Second language attrition in the brain looks like acquisition in reverse.


Tanner, D. (2011). Agreement mechanisms in native and nonnative language processing:


Footnotes

1 Some studies of morphosyntactic processing have reported an additional negative-going wave prominent over left-anterior portions of the scalp, with an onset around between 100-400ms poststimulus (the LAN effect: e.g., Friederici et al., 1996; Osterhout & Holcomb, 1992; Osterhout & Mobley, 1995; Rossi, Gugler, Hahne, & Friederici, 2005). However, the LAN has been inconsistent across studies (e.g. Allen et al., 2003; Ditman, Holcomb, & Kuperberg, 2007; Kaan, Harris, Gibson, & Holcomb, 2000; Kaan & Swaab, 2003; Kuperberg et al., 2003; Molinaro et al., 2008; Osterhout & Mobley, 1995). We therefore focus on the P600 as an index of morphosyntactic processing.

2 There are, however, some exceptions to this generalization (e.g., Chen, Shu, Y. Liu, Zhao, & Li, 2007; Sabourin & Haverkort, 2003; Sabourin & Stowe, 2008).

3 Data from four additional participants were excluded due to excessive movement, eye blink, or alpha wave artifact in the recorded EEG.

4 We also fit regression models with the motivation variable undichotomized. Results were qualitatively similar. However, we chose to include the dichotomized variable in the final model because the distribution of the variable was highly skewed and there were a large number of observations at one end of the distribution. Dichotomization is an appropriate transformation in such circumstances (MacCallum, Zhang, Preacher, & Rucker, 2002).
Table 1

*Background information for L1 Spanish-L2 English participants*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at testing (years)</td>
<td>35.2</td>
<td>6.9</td>
<td>24 – 49</td>
</tr>
<tr>
<td>AoE (years)</td>
<td>11.6</td>
<td>7.1</td>
<td>5 – 30</td>
</tr>
<tr>
<td>AoA (years)</td>
<td>23.9</td>
<td>6.4</td>
<td>15 – 40</td>
</tr>
<tr>
<td>LoR (years)</td>
<td>10.6</td>
<td>6.4</td>
<td>5 – 27</td>
</tr>
</tbody>
</table>

*Self-rated Proficiency (Spanish)*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaking</td>
<td>7</td>
<td>0</td>
<td>7 – 7</td>
</tr>
<tr>
<td>Listening</td>
<td>7</td>
<td>0</td>
<td>7 – 7</td>
</tr>
<tr>
<td>Reading</td>
<td>7</td>
<td>0</td>
<td>7 – 7</td>
</tr>
<tr>
<td>Writing</td>
<td>6.95</td>
<td>0.2</td>
<td>6 – 7</td>
</tr>
</tbody>
</table>

*Self-rated Proficiency (English)*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaking</td>
<td>5.8</td>
<td>0.7</td>
<td>4 – 7</td>
</tr>
<tr>
<td>Listening</td>
<td>6.2</td>
<td>0.7</td>
<td>4 – 7</td>
</tr>
<tr>
<td>Reading</td>
<td>6.3</td>
<td>0.6</td>
<td>5 – 7</td>
</tr>
<tr>
<td>Writing</td>
<td>5.8</td>
<td>0.9</td>
<td>4 – 7</td>
</tr>
</tbody>
</table>

*English ECPE Proficiency Scores*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary (30 possible)</td>
<td>27.7</td>
<td>2.7</td>
<td>20 – 30</td>
</tr>
<tr>
<td>Grammar (20 possible)</td>
<td>17.4</td>
<td>2.3</td>
<td>11 – 20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Total (50 possible)</td>
<td>45.1</td>
<td>4.2</td>
<td>36 – 50</td>
</tr>
<tr>
<td>L2 Motivation</td>
<td>6.4</td>
<td>0.8</td>
<td>5-7</td>
</tr>
<tr>
<td>Frequency use English (overall)</td>
<td>4.6</td>
<td>0.9</td>
<td>3-7</td>
</tr>
</tbody>
</table>
Table 2

Mean, standard deviation, and range for DVs and background measures used in regression models

<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDI</td>
<td>0.40</td>
<td>4.63</td>
<td>-7.41 – 12.14</td>
</tr>
<tr>
<td>RMI</td>
<td>4.02</td>
<td>2.11</td>
<td>0.96 – 8.40</td>
</tr>
<tr>
<td>LogAoA</td>
<td>1.36</td>
<td>0.11</td>
<td>1.18 – 1.60</td>
</tr>
<tr>
<td>LogLoR</td>
<td>1.00</td>
<td>0.22</td>
<td>0.70 – 1.43</td>
</tr>
<tr>
<td>Freq.</td>
<td>4.60</td>
<td>0.88</td>
<td>3 – 7</td>
</tr>
<tr>
<td>Proficiency</td>
<td>45.05</td>
<td>4.17</td>
<td>36 – 50</td>
</tr>
<tr>
<td>Motivation</td>
<td>0.55</td>
<td>0.51</td>
<td>0 – 1</td>
</tr>
</tbody>
</table>

Note: RDI = Response Dominance Index; RMI = Response Magnitude Index; LogAoA = Log-Age of Arrival; LogLoR = Log-Length of Resience; Freq. = Frequency of L2 Use; Proficiency = Total ECPE Score; Motivation = Dichotomized Motivation self-report.
Table 3

*CORRELATION MATRIX FOR DVs AND BACKGROUND MEASURES USED IN REGRESSION MODELS

<table>
<thead>
<tr>
<th></th>
<th>RDI</th>
<th>RMI</th>
<th>LogAoA</th>
<th>LogLoR</th>
<th>Freq.</th>
<th>Proficiency</th>
<th>Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDI</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMI</td>
<td>.187</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LogAoA</td>
<td>-.310</td>
<td>-.327</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LogLoR</td>
<td>.440⁺</td>
<td></td>
<td>-.367</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Freq.</td>
<td>.361</td>
<td>.275</td>
<td>-.164</td>
<td>.235</td>
<td>--</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proficiency</td>
<td>-.012</td>
<td>.438⁺</td>
<td>-.003</td>
<td>.264</td>
<td>.077</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>Motivation</td>
<td>.511⁺</td>
<td>.111</td>
<td>.385</td>
<td>.016</td>
<td>.047</td>
<td>.209</td>
<td>--</td>
</tr>
</tbody>
</table>

Note: RDI = Response Dominance Index; RMI = Response Magnitude Index; LogAoA = Log-Age of Arrival; LogLoR = Log-Length of Residency; Freq. = Frequency of L2 Use; Proficiency = Total ECPE Score; Motivation = Dichotomized Motivation self-report. ⁺p < .06, *p < .05.
Table 4

Multiple regression coefficients for model with RDI as DV

<table>
<thead>
<tr>
<th>Predictor</th>
<th>B</th>
<th>SE</th>
<th>β</th>
<th>t</th>
<th>Partial-r</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>LogAoA</td>
<td>-18.62</td>
<td>7.06</td>
<td>-0.45</td>
<td>-2.64</td>
<td>-.576</td>
<td>.019</td>
</tr>
<tr>
<td>LogLoR</td>
<td>5.86</td>
<td>3.43</td>
<td>0.28</td>
<td>1.71</td>
<td>.415</td>
<td>.110</td>
</tr>
<tr>
<td>Freq.</td>
<td>1.09</td>
<td>0.78</td>
<td>0.21</td>
<td>1.39</td>
<td>.348</td>
<td>.186</td>
</tr>
<tr>
<td>Proficiency</td>
<td>-0.283</td>
<td>0.17</td>
<td>-0.26</td>
<td>-1.67</td>
<td>-.408</td>
<td>.117</td>
</tr>
<tr>
<td>Motivation</td>
<td>6.57</td>
<td>1.47</td>
<td>0.73</td>
<td>4.48</td>
<td>.767</td>
<td>&lt; .001</td>
</tr>
</tbody>
</table>

Note: $R^2 = .711$, Adjusted $R^2 = .608$, $F(5, 14) = 6.897$, $p = .002$. 
Figure 1. Upper panel shows grand mean waveforms for grammatical (black line) and ungrammatical (red line) verbs from nine representative electrodes. Onset of the critical verb is indicated by the vertical bar. Calibration bar shows 3µV of activity; each tick mark represents 100ms of time. Positive voltage is plotted down. Lower panel shows scalp topographies of the main effect of grammaticality in the 400-500ms and 500-1000ms time window. Voltages represent mean amplitude in each time window in the ungrammatical minus grammatical condition. Topographic map scale is in microvolts.
Figure 2. Scatterplot showing the distribution of N400 and P600 effect magnitudes across participants, averaged across three midline electrodes (Fz, Cz, and Pz). Each dot represents a data point from a single participant. The solid line shows the best-fit line for the data from the regression analysis. The dashed line represents equal N400 and P600 effect magnitudes: individuals above/to the left of the dashed line showed primarily an N400 effect to agreement violations, while individuals below/to the right of the dashed line showed primarily a P600 effect.
Figure 3. Grand mean waveforms and scalp topographies of grammaticality effects (i.e., mean amplitude in ungrammatical minus grammatical condition) presented separately for those who
showed primarily N400 (Panel A) and P600 (Panel B) effects to verb agreement violations. Onset of the critical verb is indicated by the vertical bar. Calibration bar shows 3µV of activity; each tick mark represents 100ms of time. Positive voltage is plotted down. Topographic map scale is in microvolts.