frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 20 September 2011
doi: 10.3389/fpsyg.2011.00219

Error-related activity and correlates of grammatical plasticity

Doug J. Davidson[1]* and Peter Indefrey[2]

[1] *Basque Centre on Cognition, Brain and Language, Donostia, Basque Country, Spain*
[2] *Institut für Sprache und Information, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany*

Edited by:
Judith F. Kroll, Penn State University, USA

Reviewed by:
Natasha Tokowicz, University of Pittsburgh, USA
Susan C. Bobb, Georg-August-Universität Göttingen, Germany
Janet McDonald, Louisiana State University, USA

****Correspondence:***
Doug J. Davidson, Basque Centre on Cognition, Brain and Language, Mikeletegi Pasealekua 69, 2, 20009 Donostia, Basque Country, Spain.
e-mail: ddavidson@bcbl.eu

Cognitive control involves not only the ability to manage competing task demands, but also the ability to adapt task performance during learning. This study investigated how violation-, response-, and feedback-related electrophysiological (EEG) activity changes over time during language learning. Twenty-two Dutch learners of German classified short prepositional phrases presented serially as text. The phrases were initially presented without feedback during a pre-test phase, and then with feedback in a training phase on two separate days spaced 1 week apart. The stimuli included grammatically correct phrases, as well as grammatical violations of gender and declension. Without feedback, participants' classification was near chance and did not improve over trials. During training with feedback, behavioral classification improved and violation responses appeared to both types of violation in the form of a P600. Feedback-related negative and positive components were also present from the first day of training. The results show changes in the electrophysiological responses in concert with improving behavioral discrimination, suggesting that the activity is related to grammar learning.

Keywords: error-related activity, language learning, plasticity, bilingualism, morphosyntax

1. INTRODUCTION

Grammatical learning has been subject to extensive debate in linguistics, psychology, and neuroscience. One reason for this is the widely discussed sensitive period hypothesis (Lenneberg, 1967; Johnson and Newport, 1989, 1991; Newport et al., 2001), which maintains that the adult-onset learning of grammatical features or rules is less effective than child-onset learning. A second reason is the practical relevance of adult grammar learning for second language (L2) learning and bilingualism, both of which have a profound impact on group social organization and work productivity (Knudsen et al., 2006). However, the dynamic patterns of change in grammar learning are much less discussed. Learning dynamics could be important in these debates because grammatical features are learned over multiple, embedded time scales: The learning that occurs in the span of days or hours of a single lecture is embedded within the years or months that make up a second language course. In principle, learning must include brain activity at even shorter time scales, such as the time span of individual sentences or words, because the learning events that make up the longer time scales of the lecture or the course consist of individual sentences or words. For these reasons, a potentially effective method to study the cortical mechanisms involved in grammatical learning would be to relate the cortical activity present at the shorter temporal scales of individual sentences to activity at longer temporal scales.

The event-related potential (ERP) is an effective tool to understand this process because unlike other measures of physiology such as fMRI or PET, it has sufficient temporal resolution to separate responses to individual words within a sentence, as well as to a classification response, if it is used in a learning task, or the response to the feedback that might occur after the classification response. In studies of first language comprehension, ERP grammatical violation responses have been observed to a variety of morphosyntactic errors within a sentence (Hagoort et al., 1993; Osterhout and Holcomb, 1995; for review see Kutas et al., 2006). These responses have also been observed in adult language learners who have obtained a relatively high level of second language proficiency, but usually not in those who have not (Weber-Fox and Neville, 1996; Hahne, 2001; Hahne and Friederici, 2001; Friederici et al., 2002; Rossi et al., 2006; see Kotz, 2009 for a recent review). These L2 studies indicate that at a time scale that is sufficiently long to attain (behavioral) L2 grammatical competence, the electrophysiological responses to L2 violations have also changed to resemble those of the native L1 response. However, many of these studies have compared cross-sections of learner groups (Weber-Fox and Neville, 1996; Hahne and Friederici, 2001; Rossi et al., 2006), or relatively long time scales of learning in longitudinal designs (Osterhout et al., 2006, 2008; cf. Mueller et al., 2005). Strikingly, the work of Osterhout and collegues has shown a sequence of ERP violation effects such that grammatical violations elicit an N400 effect early in learning, but in contrast a P600 more like the native response later on. Morgan-Short et al. (2010) have recently compared explicit and implicit training conditions for an artificial grammar-learning task, observing either an N400-like pattern or P600-like pattern depending on proficiency levels and the type of instruction. It is, however, less clear which types of EEG activity are present at shorter time scales when learners are acquiring knowledge, or perhaps more importantly, how this activity is related to behavioral change.

One way this could be done is to examine explicit learning. This might occur, for instance, when learners judge sentences

as grammatical or not (e.g., providing a classification response), in the case where feedback is provided to indicate whether the classification is correct or not (e.g., comprehending a feedback signal). The feedback would allow learners to establish, over a number of trials, which features of the sentences are relevant to obtain a correct classification, so that over trials more of the behavioral responses would be correct. In the electrophysiological literature, there is a second type of ERP effect observed in choice–response tasks related to this process of explicit learning which can be viewed as error-related activity (Falkenstein et al., 1991; Gehring et al., 1993). The response effect, termed the error-related negativity (N_e) is obtained by subtracting the average ERP for correct choices from error choices in a time window of −150 ms before and +150 ms after participants make a behavioral response, appearing on centro-frontal electrodes. An error-related positivity (P_e) is observed after the N_e, in the time window from +150 to 300 ms post-response on a similar set of electrodes (Overbeek et al., 2005). There are also error-related ERP effects seen in response to feedback that informs subjects that their previous responses were incorrect. In this paper we will refer to these as feedback-N_e and feedback-P_e, so that the terminology is symmetrical to the behavioral response terminology. It should be noted, however, that the positive component is often described as a P300 effect in the literature. Obtained by a similar subtraction of correct feedback-related ERPs from error feedback-related ERPs, the time windows for the feedback-N_e (+150 to +300 ms post-feedback) and feedback-P_e (+300 to +500 ms post-feedback) reflect the average electrophysiological response to the feedback stimulus.

There is an extensive body of research in cognitive neuroscience that shows that brain areas in medial frontal cortex are involved in cognitive control, including performance monitoring, response errors, conflict, and uncertainty about correct responses, as well as responses to feedback indicating that performance is correct or incorrect in learning tasks (Ridderinkhof et al., 2004a,b; Overbeek et al., 2005). This work seeks to provide a unified account of brain activity in situations of response conflict as well as during adaptive adjustments to the environment. Earlier work by Holroyd and Coles (2002) provided a theoretical account of the changing relationship between the feedback-N_e and response-N_e in learning tasks. In their account, the initial large-amplitude feedback-N_e (but not response-N_e) reflects a learning process in which the internal state of learners is modified to predict the likely outcome of behavioral choices. Later in training, this modification is reflected in a larger response-N_e (and not feedback-N_e) at the point in time where the response choice is made. It is in this sense that the components can be seen as correlates of cognitive control: Learners adjust their internal state according to the constraints of the experimental task. The interpretation of the response- or feedback-P_e is less clear in the literature. It has been argued that the response P_e is a type of P300 that is related to error awareness (Leuthold and Sommer, 1999; Frank et al., 2007), which predicts that it could be a mediating factor in learning-related improvement.

This study therefore investigated how violation-, response-, and feedback-related activity correlates with behavioral grammatical learning of declension, following a previous study investigating the same topic (Davidson and Indefrey, 2009). In particular, we focus on the dynamics of these ERP components to better understand how events at shorter time scales of individual learning sessions are related to the stability of grammatical knowledge at the time scale of weeks or months.

Several recent studies of second language usage or artificial grammar learning have also employed error-related activity. Sebastian-Gallés et al. (2006) showed that error-related negativity was reduced or absent in Spanish-dominant early Spanish–Catalan bilinguals making a difficult non-word decision. Although this study employed multiple ERP effects (N400, N_e) to examine the lexical decision making process, it did not directly concern learning processes linked to these components. Also studying already-proficient learners, Ganushchak and Schiller (2009) showed that an N_e effect present on error trials of a phonome-monitoring task was larger with increased time pressure in Dutch–German bilinguals. Finally, Opitz et al. (2011) have shown using an artificial grammar-learning task with visually presented strings that a feedback-related negativity (N_e in the present paper) remained constant over learning, while the amplitude of a feedback-related positivity decreased with learning. Like the present study, this study employed both classification and feedback, although the task was not that of L2 language learning. The present study, like Davidson and Indefrey (2009) attempts to use multiple ERP components present in a learning task (P600, response-N_e/P_e, feedback-N_e/P_e) to try to understand how brain activity that is linked to discrimination or learning changes over time.

The experiment reported here investigates how Dutch learners acquire German morphosyntactic distinctions related to gender and declension within short prepositional phrases (see the example in **Table 1**). The task for the learners was the same as in Davidson and Indefrey (2009). In brief, they were asked to judge the correctness of prepositional phrases in which we manipulated whether or not the adjective carried syntactic feature information and whether or not there was gender agreement between the head noun and the preceding determiner or adjective. In German, the expression of case, number, and gender features on an adjective depends on the preceding elements of the noun phrase. This dependency is considered to be a syntactic dependency (Zwicky, 1986) and, following Schlenker (1999), it can be described as a rule according to which syntactic features are to be expressed only on the first inflectable element of a noun phrase. If the adjective

Table 1 | Example phrases illustrating the experimental conditions.

Correct, 3 word	mit kleinem$_{[+Dat,\ -F,\ -Pl]}$ Kind$_{[-F,\ -M]}$ with small child *"with a small child"*
Correct, 4 word	mit dem$_{[+Dat,\ -F,\ -Pl]}$ kleinen$_{[]}$ Kind$_{[-F,\ -M]}$ with the small child *"with the small child"*
Declension violation, 3 word	mit *kleinen$_{[]}$ Kind$_{[-F,\ -M]}$
Declension violation, 4 word	mit dem$_{[+Dat,\ -F,\ -Pl]}$ *kleinem$_{[+Dat,\ -F,\ -Pl]}$ Kind$_{[-F,\ -M]}$
Gender violation, 3 word	mit kleinem$_{[+Dat,\ -F,\ -Pl]}$ *Frau$_{[+F]}$
Gender violation, 4 word	mit dem$_{[+Dat,\ -F,\ -Pl]}$ kleinen$_{[]}$ *Frau$_{[+F]}$

Dat, dative; F, feminine; M, masculine; Pl, plural.

is the first inflectable element of a noun phrase, it takes on a suffix of the "strong" declension paradigm. The suffix *−em* in "mit kleinem$_{[+Dat, -F, -Pl]}$ Kind$_{[-F, -M]}$" ("with a small child"), for example, specifies dative case, non-feminine gender, and singular. By contrast, if the adjective is preceded by a definite determiner that expresses the feature information, the adjective has a suffix from the weak declension paradigm that is compatible with the feature specification of the determiner but does not express the features itself ["mit dem$_{[+Dat, -F, -Pl]}$ kleinen$_{[]}$ Kind$_{[-F, -M]}$" (also "with the small child")]. According to previous linguistic analyses of German adjective declension the weak *−en* suffix can be seen as a default or "elsewhere" form (Bierwisch, 1967; Zwicky, 1986; Blevins, 1995, 2003; Cahill and Gazdar, 1997; Wunderlich, 1997; Schlenker, 1999; Clahsen et al., 2001; Penke et al., 2004). For our stimuli, we used a subset of the full German paradigm involving dative case singular noun phrases, in order to restrict the learning problem. Please see Davidson and Indefrey (2009) for further details, as well as the primary linguistic work (Zwicky, 1986; Schlenker, 1999) for a description of the full paradigm. Examples for correct noun phrases and the declension and gender violation conditions are given in **Table 1**.

In Davidson and Indefrey (2009), both native German speakers and Dutch L2 learners of German responded to declension violations with P600 effects, but for gender violations, only native speakers showed P600 effects. In that study, after an initial pre-test phase in which no instructions or feedback was provided, we provided explicit instructions for classifying these phrases, and feedback immediately after the classification response. In the present study, we again used a pre-test phase without explicit instructions or feedback, and in the training phase provided the feedback, but with some delay, after the classification response. We also changed the procedure by not providing instructions about the grammatical rules. These two changes to the EEG experiment (slightly delayed feedback and no explicit instructions) were designed to reveal changes in the different aspects of the error-related activity over time. The separation of the behavioral response and the feedback was designed to examine differences in the dynamical behavior of the response-related and feedback-related activity, to see whether the predicted changes in activity derived from Holroyd and Coles' (2002) account apply in this task. Also, without explicit instructions, it was hypothesized that participants would take longer to reach the comparable levels of proficiency. This was based on the assumption that the previously used instructions had informed participants about which aspects of the phrases to attend and remember during the task. Without instruction, there should be a slower evolution of the changes in behavior, and allow us to see more clearly how the ERP responses are related to behavior.

1.1. SUMMARY AND HYPOTHESES

Based on our previous work, we hypothesized that changes in violation- and error-related responses will occur in conjunction with morphosyntactic learning, to the extent that this learning is revealed by classification performance. Our linking assumption is that ensemble electrophysiological activity can be recorded with EEG in language learners related to the following: (*i*) recognizing grammatical constraints, (*ii*) making correct and incorrect choices, and (*iii*) processing feedback signals. With respect to recognition of grammatical constraints, averaging the single trials of EEG and comparing violation ERPs to their controls should reveal a P600 violation response in the learners. Our assumption is that the synchronized ensemble activity giving rise to the violation ERP reflects the recognition or repair of grammatical violations. A prediction from this is that the P600 amplitude to the violations will be greater after learning than before learning to the extent that the learners can employ the knowledge they have acquired in real time. With respect to the electrophysiological response to feedback signals, it is assumed that comparing the ERP to feedback indicating an incorrect choice to the ERP to feedback indicating a correct choice will reveal difference components such as the N_e (and possibly the P_e), because this has been observed in previous EEG work with two-alternative forced-choice responses. The prediction is that the feedback N_e or P_e effect will be present early during learning but decrease in amplitude over learning trials, as predicted by the Holroyd and Coles (2002) account. With respect to the behavioral classification, participants' discrimination should improve over trials. In concert with this, a response-related N_e could appear, as this is also predicted by the Holroyd and Coles (2002) account. Finally, individual learner variation in the violation- and/or error-related ERP magnitudes should be statistically related to variation in grammatical classification, to the extent that there is a simple and direct (linear) relationship between the activity and the classification performance (see also van der Helden et al., 2010).

The present experiment also included additional behavioral tasks, including an *n*-back test of working memory and a computerized version of the Wisconsin card sort task. These behavioral tasks were used in an attempt to measure components of individual variation which might be related to the learning task. It was hypothesized that differences in working memory ability, for example, might be related to participants' ability to remember the outcome of previous trials while processing the phrases. The card-sorting task was hypothesized to relate to participants' tendency to change classification rules in response to feedback.

2. METHOD

2.1. PARTICIPANTS

Twenty-two native Dutch speakers (16 female, all right-handed, average age $M = 23.1$ years, SD $= 3.1$ years, range 19–29 years) were recruited with posted advertisements from Radboud University in Nijmegen, The Netherlands, a city near a border with Germany. The advertisements described a generic EEG experiment, and did not refer to language learning or to German instruction. As shown in **Table 2**, most participants had previous coursework in German during high school, and their average self-rated proficiency (5-point scale, $max = 5$ indicating high proficiency) was near the midpoint of the scale, or below the midpoint. This was true for most language skill components: speaking ($M = 2.1$), listening ($M = 2.8$), writing ($M = 1.7$), reading ($M = 2.8$), grammar ($M = 1.4$), and expression ($M = 2.0$). Also, recent exposure (self-reported number of hours in the last 3 months) was relatively low. Before completing the EEG tasks, all participants completed a European Reference Frame multiple choice assessment of German (maximum 30 possible correct) prepared by the Goethe Institute[1].

[1] www.goethe.de

Table 2 | Participant variables related to knowledge of German or valence toward German.

Variable	Mean	SD	Range (min–max)
Self-rated proficiency (average)	2.1	1.1	1.3–3.0
Age of initial German language education (years)	13.5	5.0	12–23
Duration of German language education (years)	3.2	2.4	0–6
German language proficiency, global test	14.7	3.5	9–22
German language proficiency, specific	3.7	1.3	1–6
Comfortable–uncomfortable using German	1.9	1.4	1–3
Like–dislike using German	3.5	1.8	1–5
Important–non-important to use German	2.5	2.5	1–5
Easy-difficult to use German	3.1	1.5	2–4
Recent exposure to German	0.5	1.0	0–1

Six of the questions on the test concerned morphosyntactic properties, the average score for this subset ($max = 6$) was slightly higher than chance. The Goethe Test scores indicate relatively low levels of German proficiency, and in addition, the test scores were not significantly correlated with self-rated proficiency, $r = 0.276$.

2.2. DESIGN AND PROCEDURE

The design of the experiment (see **Table 1**) was similar to that of Davidson and Indefrey (2009). In the pre-test, training, and follow-up phases, there were three repeated measures factors: prepositional phrase Grammaticality (violation, control), number of words in the phrase (3 or 4, corresponding to strong and weak forms of the adjective, respectively), and sentence Type (declension, gender).

The procedure consisted of two experimental EEG sessions and a third behavioral-only follow-up. The first EEG session included several behavioral measures, a pre-test and the first part of the training phase. The second EEG session, approximately 1 week after the first session ($M = 8.0$ days, SD = 1.9, $min = 5$, $max = 13$), included behavioral measures and the second and third parts of the training phase. All participants completed both EEG sessions. Approximately three and a half months after the second EEG session ($M = 109.6$ days, SD = 17.6, $min = 83$, $max = 148$), 15 participants returned for a behavioral follow-up (the others did not respond to the follow-up request, or were not available).

During both EEG sessions, the main experimental task was the classification task. This task consisted of a series of trials in which phrases were presented on a CRT monitor at 300 ms/word with an ISI of 600 ms between words. The words were presented in 26 point white Arial characters on a black background in a dimly lit room. Each trial began with a yellow fixation cross for 600 ms, followed by the first word of the phrase. The last word of the phrase was followed by a white fixation cross, which remained on screen until 1 s after participants responded.

The pre-test was conducted to assess participants' grammatical knowledge of and performance on the materials used in the experiment at the start of the experiment. During this pre-test (as well as in the behavioral-only follow-up), participants classified phrases without feedback on their response choices. Participants classified phrases as acceptable or not acceptable by pressing one of two keys with the index or ring fingers of their right hand.

The learning phase on the first day followed the pre-test after a short (5 min) break. During the learning phase, the classification task was presented just as in the pre-test, but with feedback after each response. Specifically, 1.0 s after participants' response choice, a small green (red) square was presented on the center of the screen for 0.25 s, indicating that their classification was correct (incorrect). Note that the feedback did not indicate the source of the participants' correct response or error, and it did not indicate the correct version of the presented phrase, only whether the classification was correct or not. After the feedback, the next trial began 1 s after the feedback signal.

In addition to the classification task, several other behavioral measures were provided during the application of the electrodes before the classification task in both EEG sessions. These included an *n*-back test of working memory (Owen et al., 2005), and a computerized version of the Wisconsin card-sorting task. These tasks were included to provide different measures of individual differences in general and language-related performance. EEG was not recorded during their administration.

2.3. MATERIALS

Four common German adjectives (*klein, groß, alt, neu*; respectively "small," "large," "old," "new") and 40 common German nouns were chosen to serve as stimulus materials, as well as the preposition *mit* (with) and the determiners *dem* and *der*; corresponding to dative case neuter and feminine forms of the definite determiner ("the"). Dutch has a two-gender system with a neuter gender corresponding to the neuter gender of most German cognate nouns and a so-called common gender corresponding to the masculine or feminine gender of most German cognate nouns. The nouns were chosen so that they had the corresponding gender of the Dutch translation (neuter: *Fenster, Haus, Pferd, Gleis, Schaf, Buch, Glas, Bett, Messer, Institut, Museum, Hemd, Hotel, Gebäude, Bild, Dorf, Büro, Schloss, Schiff, Auge*; corresponding to (respectively) "window," "house," "horse," "track," "sheep," "book," "glass," "bed," "knife," "institute," "museum," "shirt," "hotel," "building," "picture," "village," "office," "castle," "ship," "eye"; and feminine: *Tür, Schule, Kuh, Katze, Straße, Geschichte, Tasse, Couch, Gabel, Universität, Ausstellung, Hose, Garage, Wohnung, Zeichnung, Stadt, Bank, Kirche, Bahn, Nase*; corresponding to "door," "school," "cow," "cat," "street," "story," "cup," "couch," "fork," "university," "exhibition," "pants," "garage," "house," "drawing," "city," "bank," "church," "train," "nose"). To ensure that determiner and adjective forms unambiguously predicted the gender of the following nouns, we did not use masculine nouns, because the masculine forms of the determiner and adjectives are identical to neuter forms in German. The critical words (CW) included the adjective and noun for the various conditions (see **Table 1**). The phrases were created by pairing each noun to two of the adjectives in one set of stimuli and to the remaining two adjectives in a second set of stimuli. In each phase (Pre-test, Training 1–3), there were 240 stimuli. These consisted of 40 phrase stimuli presented for each violation and each control condition for the three- and the four-word versions of the

phrases, for the gender and the declension contrasts. Thus, on the first day of the experiment, there were 240 items presented in the pre-test, involving six repetitions of a particular noun and three repetitions of a particular article–noun pair. A distinct set of 240 items was presented in Training 1. In the second day (1 week later), the same items were presented again, 240 in Training 2, and 240 in Training 3. The number of trials did not vary between sessions. Additional practice items (10) were presented before the Pre-test to insure that participants understood the task.

2.4. APPARATUS

EEG was recorded from 64 electrodes using battery-powered BrainVision BRAINAMP Series amplifiers (Brain Products GmbH, München, Germany). Signals were sampled at 500 Hz, with a low-pass filter at 200 Hz and a high-pass filter with a time constant of 159 s during acquisition. Electrodes were applied to an equivalent inter-electrode distance Easy-Cap (Brain Products; see **Figure 1** for the electrode arrangement). Impedance levels were kept below 10 kΩ at the electrode–skin interface, with input impedance at the amplifiers at 10 MΩ (see Ferree et al., 2001). The data were recorded with respect to a left mastoid reference, and later re-referenced to an average reference including all electrodes before analysis. An additional electrode was placed below the left eye to record activity related to vertical eye movements referenced to an electrode above the eye. Lateral eye movement activity was recorded as the difference between channels near the left and right canthus.

2.5. DATA ANALYSIS

Fixed and random effects for the behavioral measures, in addition to several covariates, were modeled using a general linear mixed-effects model approach (Bagiella et al., 2000; Pinheiro and Bates, 2000; Friston et al., 2005; Baayen et al., 2008). Beta weights for the regression (b) and a t-test for the parameter values (t) are provided in the text in order to show the magnitude of the effects. Uncertainty in the parameter estimates was evaluated using highest posterior density (*HPD*) intervals, which can be treated as 95% confidence intervals for the regression parameters. The outcome measure in the regression analysis was discrimination performance (average d'), calculated as the (z-transformed) average proportion of correctly rejected violation stimuli minus the (z-transformed) average proportion of control stimuli incorrectly classified as a violation. This use of d' was intended as a measure of the ability of participants to reject violations, while at the same time correcting for false alarms.

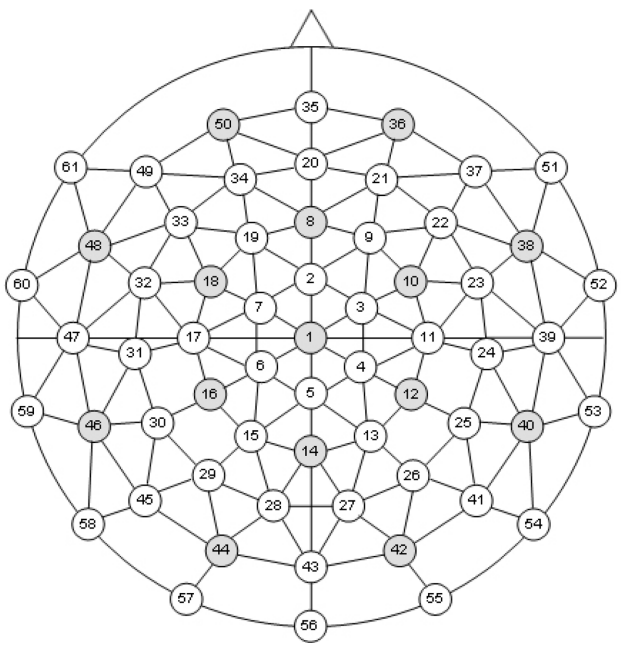

FIGURE 1 | Electrode array with locations in gray indicating approximate 10–20 locations.

In addition to the EEG measures, the covariates were taken as six loadings (accounting for 80% of the variance, identified by plotting variance captured by the loadings) from a principle components analysis (PCA) of a set of variables (see **Table 2**), including average self-rated proficiency (5-point scale for speaking, listening, writing, reading, grammar, and expression), age of initial German language education, German language proficiency as measured by the Goethe Institute Test (Goethe-Institut, 2005; both the global test, and specific items concerning case and gender), number of errors on the Wisconsin card sort task ($M = 21.7$ errors, range 3–64, out of 136 trials), average RT following an error on the Wisconsin card sort task, the slope of each participant's error curve on the n-back test (average proportion correct for $n = 1,2$, and 3 was 0.86, 0.76, and 0.68, respectively), several 5-point scale measures of valence toward learning German (comfortable–uncomfortable using German, like–dislike, important–non-important, easy–difficult), as well as an indicator of recent exposure to German (number of hours in the last 3 months). The questions for the valence toward German scales asked participants to rate whether, e.g., they liked to use German. The selection of these variables was based on a preliminary inspection of the individual difference measures, which suggested that several variables were correlated with each other. Those measures which appeared to have substantive variability across subjects were entered into the PCA analysis in order to identify a collection of linearly independent factors for the regression analysis. Together, the loadings on these principle components were hypothesized to reflect individual variation in general task performance, German proficiency, attitude toward learning German, and recent German exposure. The principle components PC1 to PC6 were most highly loaded, respectively, on the following single factors: comfortable–uncomfortable, Goethe Institute specific test, like–dislike, Wisconsin card sort number of errors, n-back task slope, and Goethe Institute global test. Note that only in the case of the n-back loading (PC5) was a single factor clearly related to the loading. In all other cases, multiple factors were related.

For the ERP analyses, trials of the recorded EEG data containing eye movement, muscle, and other noise artifacts were excluded using Matlab-based preprocessing functions[2] (Oostenveld et al., 2011), filtered with a low-pass filter (two-pass

[2]www.ru.nl/donders/fieldtrip/

6th-order Butterworth finite impulse response) with square-root half-maximum attenuation at 20 Hz, re-referenced to an average reference (please see Nunez and Srinivasan, 2006 for discussion of different re-referencing schemes), and segmented into 1 s epochs consisting of 100 ms before the onset of the CW and 900 ms following the CW. The resulting epochs were baselined with respect to the 100-ms baseline interval before CW onset and averaged according to experimental condition. Only trials with correct responses were included in the violation–control ERP contrasts, and only those participants with at least 10 observations in both violation and control conditions (two participants were excluded on this basis). The time interval for the P600 effect was defined as the range from 500 to 900 ms. On average, the numbers of non-error observations contributing to the average ERPs for the gender (declension) violation and control conditions were: $M = 17.5, 17.8, (18.7, 18.8)$ for the four-word versions, and $M = 18.1, 18.0, (18.7, 19.1)$ for the three-word versions. Response-locked data were averaged to quantify activity related to correct and incorrect responses in two time windows based on inspection of the grand average response-locked waveforms: −150 to 150 ms (N_e) and 150 to 300 ms (P_e). In both cases, the baseline interval was −300 to −150 ms. The response-locked epochs were baselined with respect to the interval from −400 to −200 ms before response onset. Feedback trials were time-locked to the feedback stimulus. The feedback N_e was defined over the interval from 100 to 300 ms, while the feedback-P_e was defined over 300–500 ms. For the feedback N_e/P_e ERPs, there were $M = 114.2$ and $M = 70.1$ trials for correct and incorrect for Training 1; $M = 132.0$ and $M = 66.8$ for Training 2, and $M = 149.0$ and $M = 45.8$ for Training 3. For the response-N_e/P_e ERPs, there were $M = 122.2$ and $M = 122.8$ for correct and incorrect in the pre-test; $M = 144.6$ and $M = 98.4$ for Training 1; $M = 344.3$ and $M = 159.1$ for Training 2 and 3 combined (see Section 3). The statistical significance of observed differences in the electrophysiological data was assessed using a clustering and randomization test (Maris, 2004; Maris and Oostenveld, 2007). As it is used here, for the average potential in a time window, the clustering and randomization test first computes a contrast statistic between conditions which is thresholded and clustered for observations in adjacent electrodes. A cluster-level statistic (sum of t-statistic) is computed for the samples in this joint set. The maximum is taken from this set, and a p-value is calculated using Monte Carlo resampling in the randomization test. The contrast values for the ERP measures were taken as the average effect over the electrodes within the significant clusters.

3. RESULTS

3.1. GRAMMATICAL CLASSIFICATION

Figure 2 shows the average classification (d') for each phase. During the pre-test, classification was near chance ($b = 0.0834$, $t = 0.5661$, HPD = −0.2057, 0.3713), but improved during the first block of the training for the trials that included the declension violations and controls (TRN1, $b = 0.6434$, $t = 2.2781$, HPD = 0.0891, 1.2116). There was no statistical evidence for improvement with the gender trials in the first block of training (TRN1, $b = 0.2068$, $t = 1.0267$, HPD = −0.2057, 0.3713, interval includes zero). Over the entire three training blocks (TRN1–3), classification improved for the gender trials ($b = 0.5098$, $t = 4.112$, HPD = 0.2693, 0.7647), but improved more for the declension trials ($b = 0.3851$, $t = 2.210$, HPD = 0.0290, 0.7309). In the follow-up phase, classification was better than during the pre-test ($b = 0.7225$, $t = 2.4194$, HPD = 0.1386, 1.3277), but lower than during the final block of training ($b = -0.8490$, $t = -2.382$, HPD = −1.5828, −0.1236). This did not depend on whether the follow-up trials were part of the declension or gender contrast ($b = -0.1084$, $t = -0.213$, HPD = −1.7538, 0.3112, interval includes zero). Note that during the pre-test and follow-up phases, there was no feedback on performance.

3.2. VIOLATION-RELATED EVOKED ACTIVITY

Figure 3 shows the isovoltage topographical distribution of the average difference between violation and control conditions for the gender contrast in the P600 time window (500–900 ms) at the CW noun in the four-word version of the phrases. A P600 effect was present on posterior electrodes in Training 3, but not in Training 1, or the Pre-test (see **Table 3**). The P600 effect was marginally significant in Training 2 for this condition. **Figure 4** shows the corresponding distribution for the declension contrast at the CW adjective, in the four-word version. A P600 effect on posterior electrodes was present in all Training sessions, but not the Pre-test (**Table 3**). The trace plots of the violation and control conditions at electrode Cz/1 for both the gender and declension contrasts indicate that the difference was primarily in the 500- to 900-ms time window. There were no significant differences in the three-word versions of the phrases, in any phase (see **Figures 5** and **6**).

3.3. ERROR-RELATED EVOKED ACTIVITY

3.3.1. Feedback-related activity

Figure 7 shows the feedback-locked average isovoltage contours for the difference between error- and correct-response trials in the feedback-N_e and -P_e time windows. During Training 1 and 2 there was an N_e effect, as indicated by the negative difference over centro-frontal electrodes and a positive difference on peripheral electrodes (see **Table 4**). In Training 3 the amplitude of this difference was reduced and the statistical effect was no longer present. In the P_e time window, there was a significant positive difference in Training 1, and in Training 2 and 3 this difference became larger, on a similar set of electrodes as in Training 1. Finally, a correlation analysis of the ERP effects did not reveal any statistically significant relationships (positive or negative) between the P600 and error-related components.

3.3.2. Response-related activity

The difference between error- and correct-response trials in the N_e and P_e time windows in the time interval near the classification response showed little evidence of a response-related N_e (time window −150 to 150 ms) or response-related P_e (time window 150–300 ms). There were no significant N_e or P_e effects during the Pre-test or any of the Training sessions. When Training 2 and 3 were pooled to increase statistical power, there was some evidence of a small centro-frontal negativity (−0.44 μV, Sum-$t = -24.68$, $p = 0.019$, on eight electrodes: 2–3, 6–8, 10–12, and 17), shown in **Figure 8**. However, this negativity was sustained into the later P_e time window (−0.54 μV, Sum-$t = -14.53$, $p = 0.037$, on five electrodes: 2–3, 10–11, and 22).

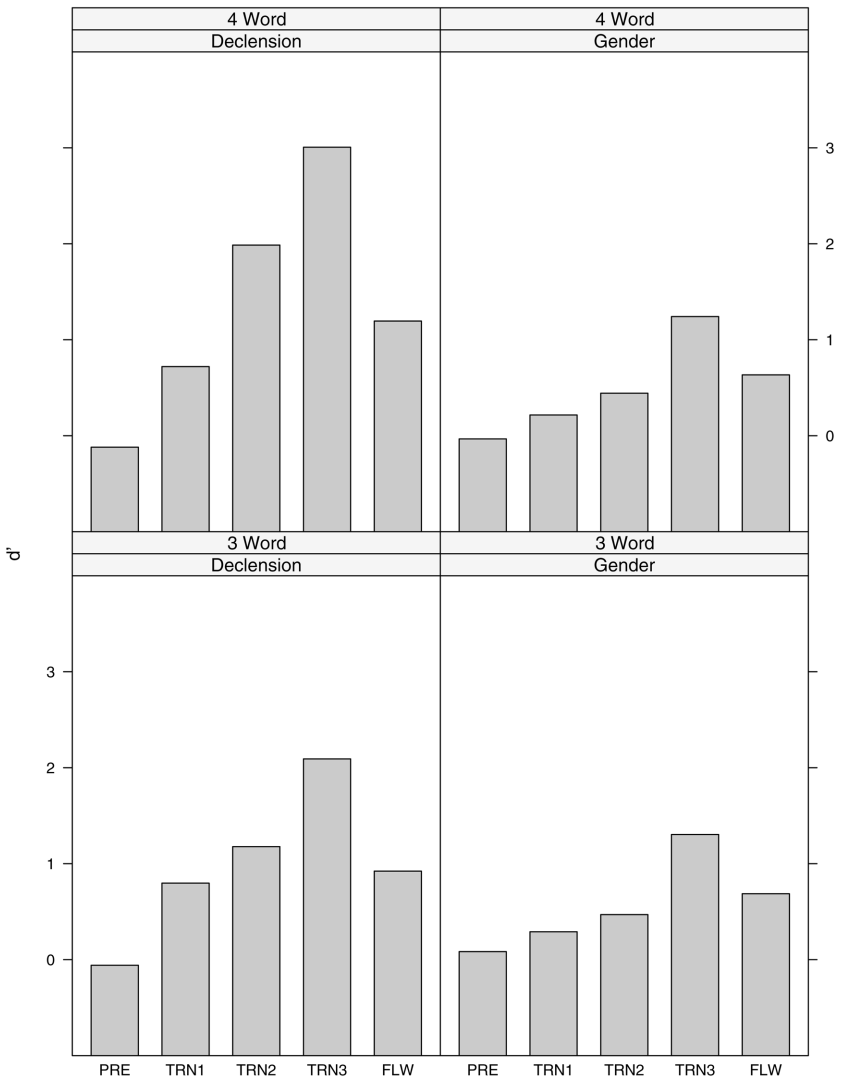

FIGURE 2 | Average classification performance (*d′*) over pre-test, training (1–3), and follow-up blocks for the three- and four-word phrases, for the declension and gender contrasts.

3.4. RELATIONSHIP BETWEEN PERFORMANCE AND ERP ACTIVITY

To investigate the relationship between EEG activity and the relatively short-term behavioral changes during training, the change in discrimination performance (average d') was modeled as a function of Training session (1–3), the ERP difference wave amplitude for the violation- and feedback-error components in sessions 1–3, as well as the individual subject loadings of the six principle components summarizing the individual difference measures (see Section 2.5). Recall that the principle components were most highly loaded, respectively, on the factors (1) comfort using German, (2) Goethe Institute specific test, (3) whether participants like using German, (4) Wisconsin card sort number of errors, (5) *n*-back task slope, and (6) Goethe Institute global test. Only in the case of the *n*-back loading (PC5) was a single factor clearly related to the loading. In all other cases, multiple factors were related. The violation components included the P600 effect amplitudes for the (four-word) gender and declension violation effects, and the error-related components included the feedback N_e and P_e effects, in all cases for each of the Training sessions 1–3. The interaction of session with each of the ERP effects and each of the principle components were included as predictors for the classification response. Each of the predictors was scaled to a mean of zero and unit SD.

The regression indicated a significant interaction of feedback-P_e effect magnitude with session such that participants with a larger P_e effect improved more over the sessions (e.g., the increasing slope in **Figure 9**, $b = 0.1498$, $t = 2.6149$, HPD $= 0.0453$, 0.2793). Please note that the first training session took place 1 week before sessions 2 and 3. None of the other ERP measures predicted performance alone, or in interaction with session. In particular, N_e effect magnitude did not predict performance, in contrast to the results reported in Davidson and Indefrey (2009).

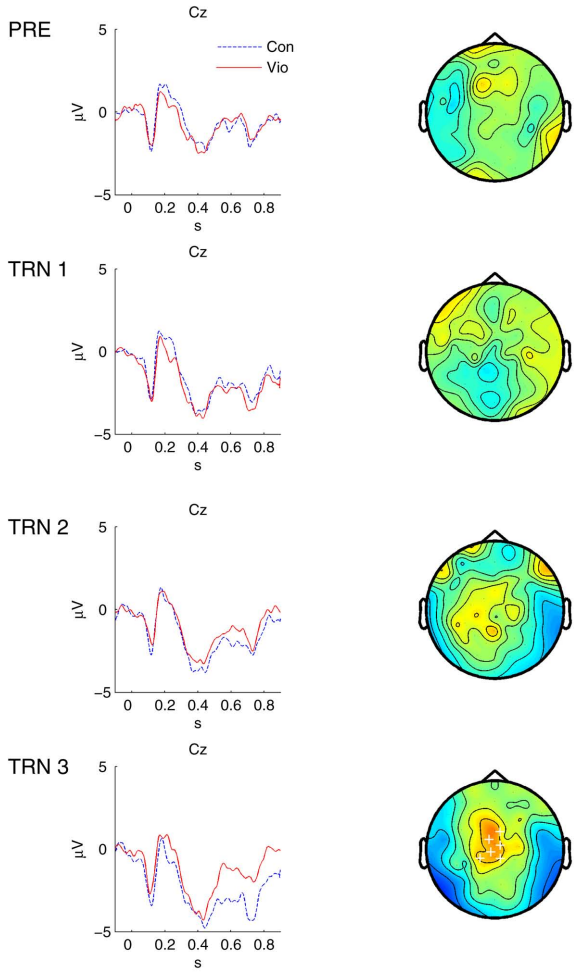

FIGURE 3 | Average event-related potentials for gender violation contrast (4 word). Trace plots indicate average voltage (in μV) as a function of time at Cz/1; control = Con (blue, dotted line), violation = Vio (red, solid line). Symbols plotted on the topographical plots indicate membership within a statistically significant negative ("−") or positive ("+") cluster of electrodes. Topographical plots are viewed from the top (left is on the left-hand side).

There were two other interactions of session with principle components of the auxiliary measures: Session*PC3, $b = -0.1387$, $t = -2.2594$, HPD = −0.2796, −0.0081; and Session*PC5, $b = 0.1744$, $t = 2.5839$, HPD = 0.0345, 0.3353. PC3 was positively related to valence toward German (the scale like–dislike learning German) as well as the number of languages reported to be known by the subjects, but negatively related to hours of recent exposure to German. PC5 was related most strongly to *n*-back performance, and it was not strongly related to other variables.

To investigate discrimination forgetting (or performance decline), the difference in discrimination from the last phase of training (Training 3) to the follow-up phase was modeled as a function of the EEG and principle component measures that were significant predictors in the Training analysis. In this analysis, P_e magnitude positively predicted discrimination in Training 3 ($b = 0.6233$, $t = 4.7330$, HPD = 0.3553, 0.9017), as was shown in the previous analysis, and there was a negative relationship between P_e magnitude and discrimination in the follow-up ($b = -0.4852$, $t = -3.1530$, HPD = −0.9551, −0.0316). This result indicates that P_e magnitude predicted which participants gained during training, but also which participants lost discrimination 3 months later. In this analysis, the principle components which were significant predictors in the Training sessions, were not predictors for the follow-up loss. In summary, the P_e effect amplitude was a consistent predictor of discrimination gain (and loss, over the longer term), even when adjusting for individual differences in a variety of performance tasks.

Table 3 | Statistics for violation-related EEG activity.

Phase	Type	Ave	Sum-*t*	*p*	Electrodes
Training 1	Decl	+1.53	38.02	0.0013	4, 6, 12–18, 25–29 (14)
	Decl	−1.68	−26.50	0.0131	45–46, 53–58 (8)
	Gen	+0.94	–	–	–
	Gen	−0.76	–	–	–
Training 2	Decl	+1.93	36.10	0.0020	1, 3–6, 10, 12–17, 28 (12)
	Decl	−1.43	−24.79	0.0116	33, 46–49, 58, 61, 63 (8)
	Gen	+0.87	9.10	0.0609	14–17 (4)
	Gen	−1.31	−8.71	0.0662	39–40, 53–54 (4)
Training 3	Decl	+2.09	16.32	0.0218	3–5, 11–13, 25 (7)
	Decl	−2.27	−40.11	0.0001	41, 48, 53–60, 63 (11)
	Gen	+1.56	17.26	0.0307	1–2, 5–7, 16 (6)
	Gen	−1.40	−13.02	0.0593	39–40, 52–54 (5)

Average difference (Ave, in μV) calculated as the average within the P600 interval (500–900 ms) after the CW onset for the two violation contrasts (Type). The summary statistics (Sum-t) and p-values (p) for the clustering and randomization tests are provided, along with the approximate electrode locations according to ***Figure 1,*** *as well as the number of electrodes in the cluster.*

4. DISCUSSION

We expected violation- and error-related ERP effects to appear in concert with the learners' discrimination improvement. The experiment reported here provided evidence for several of these effects: As participants' grammatical discrimination improved over time, P600 responses to grammatical violations emerged, error-related activity was observed in response to feedback, and the amplitude of one of the feedback-related responses was related to improved grammatical discrimination. Finally, the magnitude of the error-related activity predicted later retention of the discrimination ability.

The observed P600 responses to violations of the syntactically determined declension of German adjectives replicated findings of an earlier study (Davidson and Indefrey, 2009) in which participants had been provided with explicit instruction on the rules of German adjective declension. In the present study, no explicit instruction was given and the learners had to rely on positive and negative feedback for the learning of adjective declension rules. The behavioral data showed a gradual increase from grammaticality judgment performance near-chance level in the pre-test to high performance in the range of a native speaker control group participating in the previous study (the range of the hit rate – false alarm rate measure in native speakers was approximately 0.6–0.9,

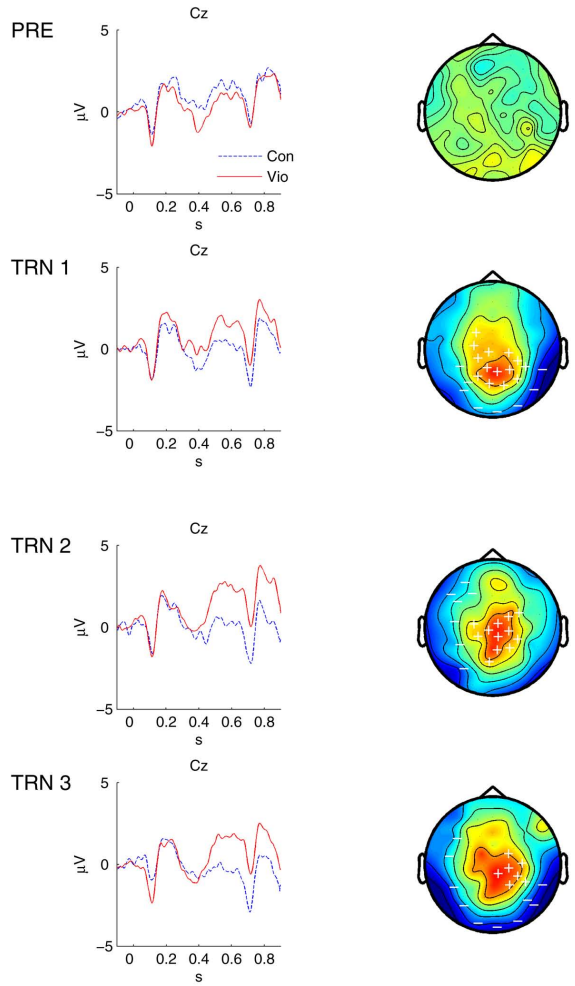

FIGURE 4 | Average event-related potentials for declension violation contrast (4 word). Trace plots indicate average voltage (in μV) as a function of time at Cz/1; control = Con, violation = Vio. Symbols plotted on the topographical plots indicate membership within a statistically significant negative ("−") or positive ("+") cluster of electrodes. Topographical plots are viewed from the top (left is on the left-hand side).

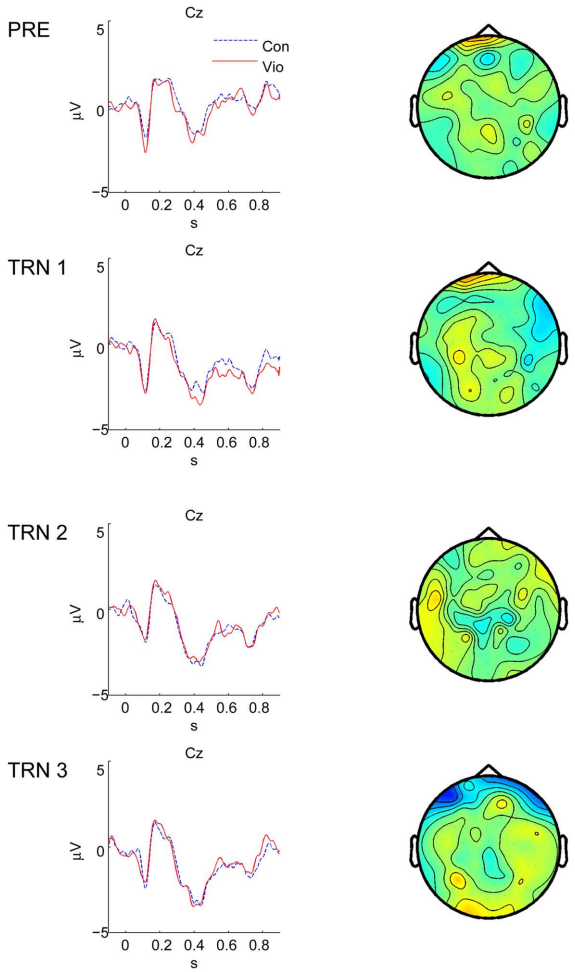

FIGURE 5 | Average event-related potentials for gender violation contrast (3 word). Trace plots indicate average voltage (in μV) as a function of time at Cz/1; control = Con (blue, dotted line), violation = Vio (red, solid line). There were no statistically significant clusters. Topographical plots are viewed from the top (left is on the left-hand side).

see Figure 2 of Davidson and Indefrey, 2009). In the previous study with explicit instruction, there was a steep performance increase in the first training session compared to the pre-test. Despite the absence of explicit instruction in the present study, classification performance also improved rapidly and there was a significant P600 responses to declension violations already in the first training session. This gained in strength over the following training sessions. These findings indicate that the changes in the neural responses are related to the acquired grammatical knowledge *per se* (i.e., adjective declension rules) rather than how this knowledge is acquired, i.e., by receiving explicit rule instruction or by finding the rules themselves.

Like Davidson and Indefrey (2009), we only found declension violation responses in the four word prepositional phrases (e.g., *mit dem *kleinem Kind*) where the violating adjective carries a strong inflectional suffix redundantly specifying case, number, and gender information but not in three word prepositional phrases (e.g., *mit *kleinen Kind*) where the violating adjective carries a default suffix that does not specify syntactic feature information. Although both types of incorrect prepositional phrases violate the syntactic rule according to which case, number, and gender features must be specified on the first (and only on the first) inflectable element of the noun phrase, the neural violation response thus seems to hinge on the presence of (incorrect) positively specified syntactic features. Hierarchical feature specification analyses of the German adjective paradigm predict a differential response to strong forms positively specifying syntactic features and weak default forms and hence are supported by our data (see Clahsen et al., 2001; Penke et al., 2004 for corresponding psycholinguistic evidence).

In contrast to the experiment reported in Davidson and Indefrey (2009), which observed P600 responses to grammatical gender violations in native speakers but not in Dutch learners of German,

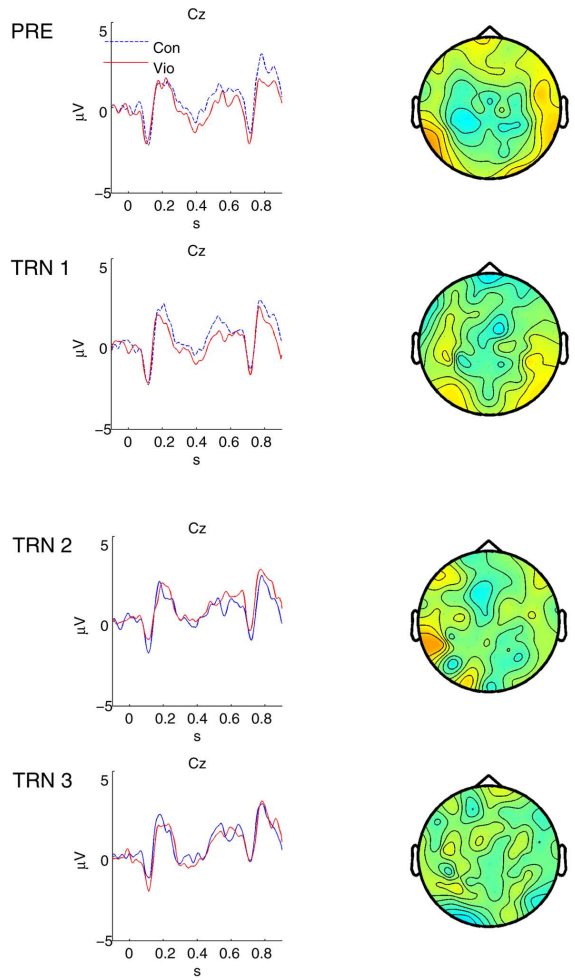

FIGURE 6 | Average event-related potentials for declension violation contrast (3 word). Trace plots indicate average voltage (in μV) as a function of time at Cz/1; control = Con, violation = Vio. There were no statistically significant clusters. Topographical plots are viewed from the top (left is on the left-hand side).

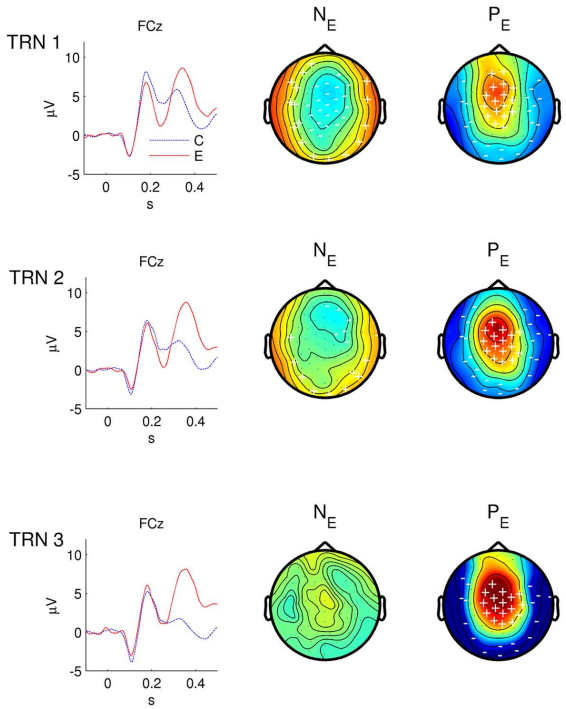

FIGURE 7 | Average event-related potentials for feedback-related activity. Trace plots indicate average voltage (in μV) over time at FCz/8; correct classification = C (blue, dotted line), error classification = E (red, solid line). Symbols plotted on the topographical plots indicate membership within a statistically significant negative cluster ("−") or positive cluster ("+") of electrodes.

in the present study we found significant P600 gender violation responses in the learners in the last training session. The observation that P600 violation effects can be observed to gender violations is in line with other studies reporting P600 responses to gender violations in a second language (Sabourin, 2003; Tokowicz and MacWhinney, 2005; Sabourin et al., 2006). One possible reason for the absence of a significant gender violation response in Davidson and Indefrey (2009) could be the fact that there was only one training session, providing no opportunity for a relatively late emergence of a response as in the present study. As indicated in Davidson and Indefrey (2009), learning to retrieve and apply grammatical gender information may be a more difficult learning task than learning to apply a declension rule because the gender category must be associatively linked to every single noun. Even though we chose the nouns such that their Dutch translation equivalents also fell into two different gender classes, the Dutch participants may have been insecure about the German gender (masculine or feminine) of the nouns whose Dutch translation equivalents belong to the common gender. A suggested by an anonymous reviewer, the Dutch participants might also have had some residual knowledge of German nominative determiner forms and were initially confused by the dative forms used in our experiment. Taken together, there is reason to assume that learning the gender class of German nouns may require more training, teaching, exposure, or usage. In Davidson and Indefrey (2009), there was one training phase with feedback, whereas in the present experiment there were three phases with feedback. See Blom et al. (2008) and Sabourin and Stowe (2008) for recent discussions of L2 gender learning and factors which contribute to learning variability.

Another important factor in the rate of grammatical learning might be the size of the set of lexical items used in the learning task. In the present experiment, a relatively small number of adjectives and nouns were used, in order to reduce the chance that participants would not know the meanings of the words. However, as suggested by an anonymous reviewer, the diversity of the item set may play an important role in modulating the appearance of the violation effects. At one extreme, if only a few carrier phrases are used, then the rate of learning might be relatively fast because the repetition of items would highlight morphosyntactic patterns (and/or reduce the lexical knowledge burden on learners). At the other extreme, if each trial had different carrier items (at both

Table 4 | Statistics for feedback error-related EEG activity (Rsp) in the three training phases.

Phase	Rsp	Win	Ave	Sum-*t*	*p*	Electrodes
Training 1	N_e	100–300	+1.17	62.33	0.0073	46–64 (18)
	N_e	100–300	−0.85	−72.20	0.0018	1–17, 19, 21–22, 28–29 (22)
	P_e	300–500	+1.49	49.37	0.0002	1–9, 14–20, 34 (17)
	P_e	300–500	−1.44	−73.56	0.0001	23–24, 38–40, 43–44, 46–47, 51–64 (22)
Training 2	N_e	100–300	+0.84	29.24	0.0144	41, 53–60, 63 (10)
	N_e	100–300	−0.79	−26.58	0.0185	3, 8–11, 20–22 (8)
	P_e	300–500	+2.18	86.35	0.0001	1–10, 13–19 (18)
	P_e	300–500	−1.86	−87.22	0.0001	37–40, 43–49, 51–64 (24)
Training 3	N_e	100–300	+0.09	–	–	–
	N_e	100–300	−0.29	–	–	–
	P_e	300–500	+3.79	105.22	0.0001	1–19 (19)
	P_e	300–500	−3.12	−114.02	0.0001	38–39, 42–43, 47–48, 52–61, 63–64 (18)

Average difference (Ave, in μV) calculated as the mean value within a time interval (Win, in ms) after the feedback onset. The summary statistics (Sum-t) and p-values (p) for the clustering and randomization tests, along with the approximate electrode locations are listed according to ***Figure 1,*** *along with the number of electrodes in the cluster.*

adjective and noun positions), then it may take a substantially higher number of trials for participants to successfully apply their grammatical knowledge, especially if it is likely that they do not know the meanings of all the words. This factor, the diversity of the item set, would be expected to have more of an impact for the gender contrast than the declension contrast, for the reasons outlined earlier. Also, in future studies it would be better to test for generalization by including new items that were not seen in the training set in an explicit test of generalization. There are potentially other explanations of the declension–gender difference seen here, but perhaps it would be advisable to find ERP evidence using a wider variety of items before elaborating different predictions.

In addition to using a small lexicon, we also simplified the German determiner and adjective declension system by only using dative case and avoiding syncretism. Our results suggest that the full system most likely would not have been learned in the available number of sessions. These simplifications, however, do not mean that our participants merely learned to associate particular items or item combinations with particular responses. Firstly, just because a relatively small set of adjectives were repeated many times in all possible forms, the identification of a particular adjective stem did not provide any cue as to its appropriate ending. Secondly, as can be seen in **Table 1**, due to the presence of both gender and declension violations our stimulus set contained an equal number of trials in which a specific adjective form with the same preceding context (e.g., mit kleinem) required a correct response (as in mit kleinem Kind) and an incorrect response (as in mit kleinem *Frau). Given that our feedback did not distinguish between incorrect responses due to declension or gender errors, participants could not learn to base their declension class decision on a particular combination of word forms such as mit kleinem. For the same reason, correct gender agreement could not be learned based on simple associations between particular determiner–noun or adjective–noun combinations and constant response requirements. Taken together, these properties of our stimulus set mean that in order to respond correctly at the observed performance level our participants had to learn the grammatical rules of the reduced system.

There have been relatively few previous EEG studies of practice-related improvement with similar tasks. A natural question is whether the P600 responses that were observed in the present study might be more generally related to practice-related improvement. In a non-linguistic domain, both Romero et al. (2008) and Pauli et al. (1994) found that practice on tasks requiring mathematical knowledge reduces the amplitude of frontal positive potentials. Both studies found that a non-selective frontal P300 component was attenuated with practice, while Romero et al. also found that a later posterior P500 component selective for correct equations than incorrect equations became larger after practice. In contrast, Pauli et al. (1994) found that the posterior positive potential was not attenuated with practice. Romero et al. attributed the difference to the fact that their experiment used a task involving the verification of alphabet–arithmetic equations, which were likely to be unfamiliar to participants before practice, while the Pauli et al. task involved producing the answers to ordinary multiplication equations, which were likely to be known before practice. If the positive responses are general task-related effects, then the results from the equation-processing experiments would suggest that late positive components seen in the present experiment should have either decreased as a function of practice (Pauli et al., 1994), or become larger for correct-string as compared to incorrect-string stimuli (Romero et al., 2008).

However, these predictions are not in agreement with the results for the P600 component. In the present experiment (also in Davidson and Indefrey, 2009), the P600 violation effect was absent initially, and only appeared after practice. This pattern after practice was similar to the native speaker control group in Davidson and Indefrey (2009). Together, these findings suggest that the emergence of the grammatical violation effect seen in the present study is likely to be related to grammatical processing, rather than

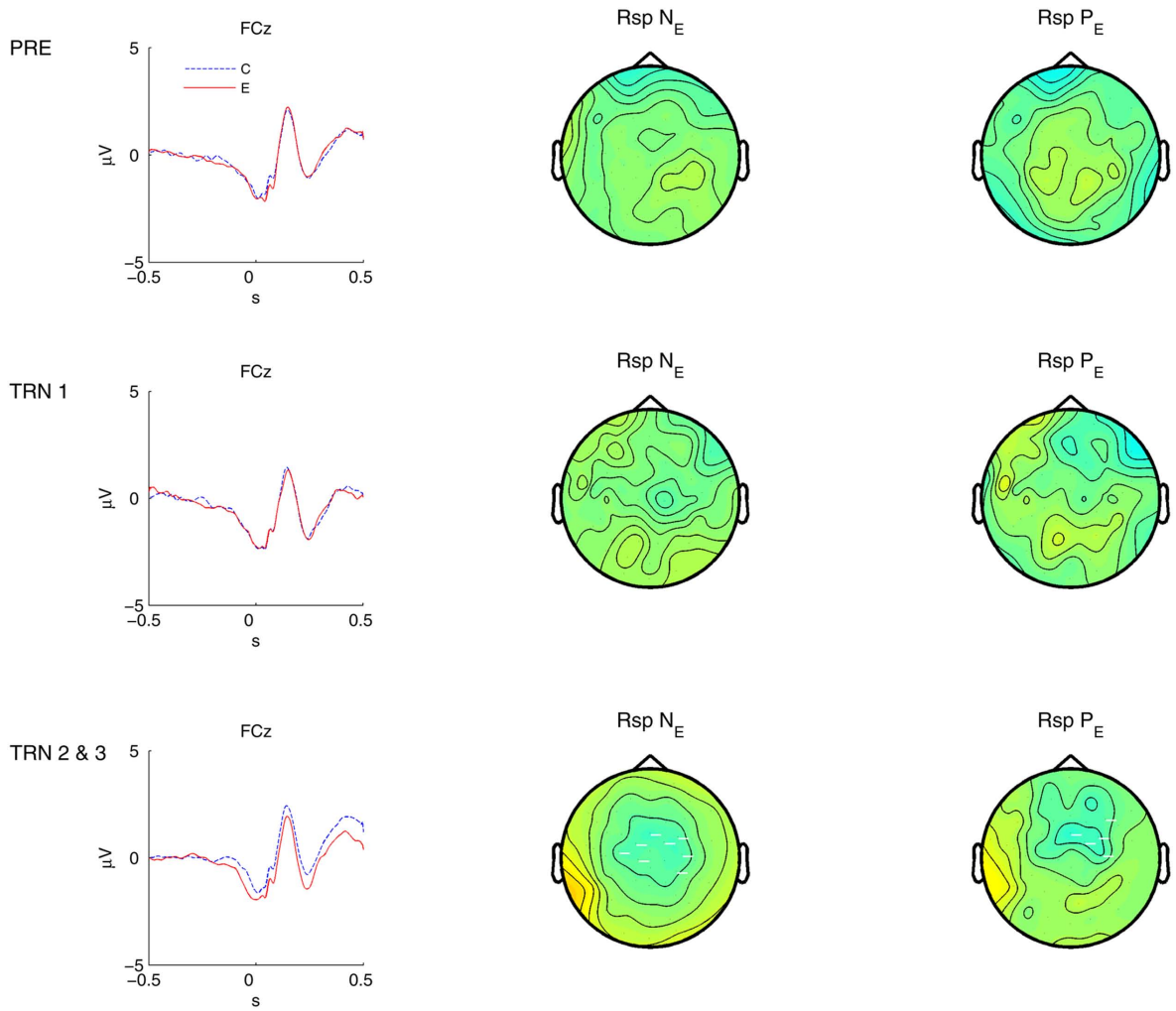

FIGURE 8 | Average event-related potentials for response-related activity. Trace plots indicate average voltage (in μV) over time at FCz/8; correct classification = C (blue, dotted line), error classification = E (red, solid line). Symbols plotted on the topographical plots indicate membership within a statistically significant negative cluster ("−") of electrodes.

a general task-related P300 effect. The changing pattern of the P600 responses seen here is consistent with the results of Davidson and Indefrey (2009), with the exception that a P600 response was observed for gender violations in the present experiment and not the previous study (see Davidson and Indefrey, 2009, as well as Sabourin, 2003; Sabourin et al., 2006).

One of the main aims of the present work was to disentangle the contributions of response- and feedback-related activity as the results in Davidson and Indefrey (2009) were not able to distinguish these, and to test differential predictions for the two error responses, suggesting that the magnitude of the feedback-N_e decreases over time, in concert with an increase in the magnitude of the response-N_e (Holroyd and Coles, 2002). In the present study we found a clear feedback negativity whereas the amplitude of the response-related N_e activity, although statistically significant, was weak and unreliable, suggesting that the error negativity observed in Davidson and Indefrey (2009) was likely due to feedback-related activity.

With respect to changes of feedback and response negativities as a function of time, during learning, our data confirmed the first prediction: The feedback-N_e indeed decreased in magnitude from the first to the last training session. This finding supports Holroyd and Coles (2002) hypothesis that the feedback-N_e reflects a learning process in which the internal state of learners is modified. More specifically, this modification likely involved the learners representation of the declension regularities on which they based their grammaticality decisions. The learners, starting with no or very little knowledge as indicated by their near-chance performance in the pre-test, had to extract the relevant grammatical knowledge from the information provided by the feedback. This means that initially this feedback must have had a relatively large impact on the learners representations as indicated by a performance increase to a high level. Even though at higher performance levels negative feedback arguably constituted a stronger conflict with the participants expectation (they knew the probability of having made a correct decision was higher than at the beginning), the amplitude

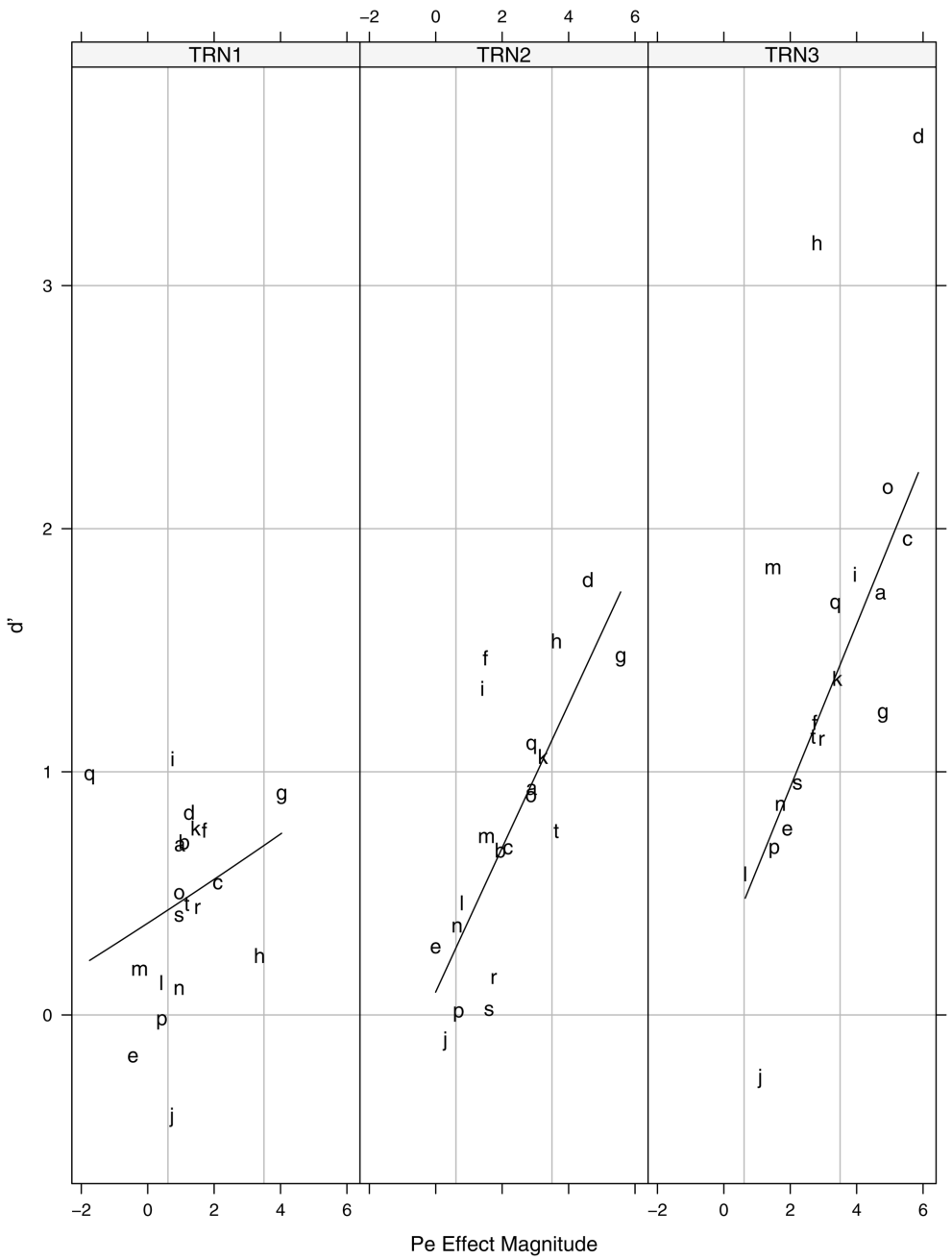

FIGURE 9 | Relationship between feedback-P_e and discrimination performance during Training blocks 1–3 (subjects indicated by letters).

of the ERP response to negative feedback decreased. In line with Holroyd and Coles (2002) prediction this may be interpreted as showing that in spite of negative feedback the learners were less prone to change their internal representations at later stages of training.

Our data do not allow any conclusion with respect to a possible response negativity. A plausible explanation for the weak response-N_e might be the task parameters. Unlike the speeded response time tasks used in previous studies of the response-N_e, participants in our study were not under substantial time pressure to provide their responses. This may have contributed to the relative weakness of the response-N_e effect, and future work investigating the response-N_e in grammatical learning might impose a shorter response deadline to boost the effect.

The second component of the feedback response, the feedback positivity (P_e), contrary to the feedback negativity became larger as a function of training and individual variation in the feedback-P_e amplitude was related to discrimination performance.

As mentioned in the Introduction, the response P_e has been suggested to be a type of P300 reflecting error awareness (Leuthold and Sommer, 1999; Frank et al., 2007). An extension of this functional characterization of the response P_e to the feedback-P_e would be in accordance with our data as the awareness of a conflict between the participants expectation and the actual feedback quite plausibly increased with the participants performance level. The better their performance level, the larger the perceived conflict would be and hence the corresponding feedback positivity.

The changing P600 and N_e/P_e ERP responses in this experiment suggest that the presentation of a series of phrases (along with the feedback) affects the behavioral classification of phrases presented at a later point in time. The relationship between the feedback activity and the violation-related activity appears to be complex, however, for at least two reasons. First, the error-related activity was both changing and multi-phasic. Over the course of training from the first to the second day, the feedback-N_e amplitude decreased while the feedback-P_e amplitude increased. Second, while the P600 amplitude increased with training, it was not itself statistically related to the N_e/P_e activity, possibly due to too much variability in the responses. While this pattern of activity precludes a simple account of the relationship between feedback and discrimination improvement, the results do provide additional support for the claim that feedback-related activity (N_e and/or P_e) can be related to grammatical learning under some circumstances. Nevertheless, given these findings, future work might employ experimental designs which are better optimized to estimate the P600 and N_e/P_e relationship, perhaps by focusing on a single type of violation with more trials and more training. The regression results also indicate that valence toward learning German or languages generally (like–dislike scale, number of languages known), as well as working memory span may be important modulating variables. Given the sample size of the present experiment, perhaps behavioral experiments on learning, which can be run with substantially larger sample sizes, could better elucidate whether these factors strongly modulate learning.

While error-related activity was related to discrimination improvement like the previous study, one notable exception was that the present study did not show a direct relationship between N_e amplitude and discrimination performance. The main differences in design were the absence of explicit instruction in the present study, and the temporal separation of the classification response from the feedback. In addition, feedback was presented on both days of training in the present study, but only during the first day of the previous study. It was hypothesized that slower learning would slow the evolution of the error-related activity over a longer time scale, but the absence of the instruction may have altered the task dynamics in such a way to make the experiments less than fully comparable. As expected, performance on the present experiment improved more slowly than in Davidson and Indefrey (2009), most likely because participants in the present study had to determine how to classify the phrases by trial and error, rather than by relying on their memory for the explicitly provided rules.

It may be that in this task, the feedback-N_e reflects recognition that the current hypothesis about grammatical classification needs to be changed. With little initial knowledge (in the present experiment), the large N_e may reflect a new, updated hypothesis, but this new hypothesis may not have been correct. Participants who updated to a better hypothesis early would in fact have shown smaller N_e subsequently. Those who changed several times before they got it right might have shown in total larger N_e responses. This would explain the present data well, but in turn raises the question about the relationship found in the previous study. In the previous study, the instruction may have made it more likely that the initial change in hypothesis was effective, because it could have strengthened the memory of the instructions. Although speculative, this account of the N_e/P_e contribution to the effects observed here could be investigated in future experiments by including a variable that would affect participants' ability to apply rules, or the number of rules to be applied. The results also suggest independent roles for the feedback-N_e and feedback-P_e effects, which have not been extensively investigated previously (see also Overbeek et al., 2005).

Finally, the results suggest that future models of grammatical plasticity should include not only an account of the learning of grammatical knowledge, but also an account of how grammatical knowledge is lost, or otherwise made unavailable after a period of disuse. The present results, along with several other recent findings (Mueller et al., 2005; Osterhout et al., 2005, 2006) suggest that the learning of grammatical knowledge can occur quite rapidly in adults, at least when acquired explicitly. The learning of this knowledge does not imply that it is stable, however. Without maintenance or usage to reinforce learning, adult grammar knowledge appears to be vulnerable to decay or interference. Future work might investigate whether the dominant factor(s) determining the effects of the hypothesized sensitive period in adult grammar learning are more related to retention than learning.

5. CONCLUSION

The experiment described here has shown that there are several electrophysiological correlates of learning in grammar-learning tasks with feedback. The results showed that these ERP measures are *dynamic*, in the sense that they can change within the span of one or two experimental sessions, at least with Dutch participants learning German as studied here. The results were largely congruent with the pattern of data reported in Davidson and Indefrey (2009), despite the absence of instructions in the present experiment. The response- and feedback-ERP results can be taken as evidence that cognitive control mechanisms function during explicit learning to help modify the knowledge state of second language learners, and/or enable the memory of this knowledge so that it can be put to use during real-time language comprehension.

ACKNOWLEDGMENTS

This work was supported by the Dutch science foundation Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) and the German science foundation Max Planck Gesellschaft (MPG). Esther Meeuwissen and Daniel von Rhein recruited participants, helped create the stimuli, and recorded the EEG with assistance from the first author.

REFERENCES

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412.

Bagiella, E., Sloan, R. P., and Heitjan, D. F. (2000). Mixed-effects models in psychophysiology. *Psychophysiology* 37, 13–20.

Bierwisch, M. (1967). "Syntactic features in morphology: general problems of so-called pronominal inflection in German," in *To Honor Roman Jakobson* (The Hague: Mouton), 239–270.

Blevins, J. P. (1995). Syncretism and paradigmatic opposition. *Linguist. Philos.* 18, 113–152.

Blevins, J. P. (2003). Stems and paradigms. *Language*, 79, 737–767.

Blom, E., Polisenská, D., and Unsworth, S. (2008). The acquisition of grammatical gender in Dutch. *Second Lang. Res.* 24, 259–265.

Cahill, L., and Gazdar, G. (1997). The inflectional phonology of German adjectives, determiners, and pronouns. *Linguistics* 35, 211–245.

Clahsen, H., Sonnenstuhl, I., Hadler, M., and Eisenbeiss, S. (2001). Morphological paradigms in language processing and language disorders. *Trans. Philol. Soc.* 99, 247–277.

Davidson, D. J., and Indefrey, P. (2009). An ERP study on changes of violation and error responses during morphosyntactic learning. *J. Cogn. Neurosci.* 21, 433–446.

Falkenstein, M., Hohnsbein, J., Hoormann, J., and Blanke, L. (1991). Effects of crossmodal divided attention on late ERP components 2. Error processing in choice reaction tasks. *Electroencephelogr. Clin. Neurophysiol.* 78, 447–455.

Ferree, T. C., Luu, P., Russell, G. S., and Tucker, D. M. (2001). Scalp electrode impedance, infection risk, and EEG data quality. *Clin. Neurophysiol.* 112, 536–544.

Frank, M. J., D'Lauro, C., and Curran, T. (2007). Cross-task individual differences in error processing: neural, electrophysiological, and genetic components. *Cogn. Affect. Behav. Neurosci.* 7, 297–308.

Friederici, A. D., Steinhauer, K., and Pfeifer, E. (2002). Brain signatures of artificial language processing: evidence challenging the critical period hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* 99, 529–534.

Friston, K. J., Stephan, K. E., Lund, T. E., Morcom, A., and Kiebel, S. (2005). Mixed-effects and fMRI studies. *Neuroimage* 24, 244–252.

Ganushchak, L. Y., and Schiller, N. O. (2009). Speaking one's second language under time pressure: an ERP study on verbal self-monitoring in German–Dutch bilinguals. *Psychophysiology* 46, 410–419.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., and Donchin, E. (1993). A neural system for error-detection and compensation. *Psychol. Sci.* 30, 306–315.

Goethe-Institut. (2005). *Placement Test German.* Available at: http://www.goethe.de [retrieved March 5, 2005].

Hagoort, P., Brown, C., and Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP-measure of syntactic processing. *Lang. Cogn. Process.* 8, 439–483.

Hahne, A. (2001). What's different in second-language processing? Evidence from event-related brain potentials. *J. Psycholinguist. Res.* 30, 251–266.

Hahne, A., and Friederici, A. D. (2001). Processing a second language: late learners' comprehension mechanisms as revealed by event-related brain potentials. *Biling. (Camb. Engl.)* 4, 123–141.

Holroyd, C. B., and Coles, M. G. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychol. Rev.* 109, 679–709.

Johnson, J. S., and Newport, E. L. (1989). Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. *Cogn. Psychol.* 20, 60–99.

Johnson, J. S., and Newport, E. L. (1991). Critical period effects on universal properties of language: the status of subjacency in the acquisition of a second language. *Cognition* 39, 215–258.

Knudsen, E. I., Heckman, J. J., Cameron, J. L., and Shonkoff, J. P. (2006). Economic, neurobiological, and behavioral perspectives on building America's future workforce. *Proc. Natl. Acad. Sci. U.S.A.* 103, 10155–10162.

Kotz, S. A. (2009). A critical review of ERP and fMRI evidence on L2 syntactic processing. *Brain Lang.* 109, 68–74.

Kutas, M., Van Petten, C., and Kluender, R. (2006). "Psycholinguistics electrified II: 1995–2005," in *Handbook of Psycholinguistics*, eds M. Traxler and M. Gernsbacher (Amsterdam: Academic Press), 659–724.

Lenneberg, E. H. (1967). *Biological Foundations of Language.* New York: Wiley.

Leuthold, H., and Sommer, W. (1999). ERP correlates of error processing in spatial S-R compatibility tasks. *Clin. Neurophysiol.* 110, 342–357.

Maris, E. (2004). Randomization tests for ERP topographies and whole spatio-temporal data matrices. *Psychophysiology* 41, 142–151.

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

Morgan-Short, K., Sanz, C., Steinhauer, K., and Ullman, M. T. (2010). Second language acquisition of gender agreement in explicit and implicit training conditions: an event-related potential study. *Lang. Learn.* 60, 154–193.

Mueller, J. L., Hahne, A., Fujii, Y., and Friederici, A. D. (2005). Native and nonnative speakers' processing of a miniature version of Japanese as revealed by ERPs. *J. Cogn. Neurosci.* 17, 1229–1244.

Newport, E. L., Bavelier, D., and Neville, H. J. (2001). "Critical thinking about critical periods: perspectives on a critical period for language acquisition," in *Language, Brain, and Development*, ed. E. Dupoux (Cambridge, MA: MIT Press), 481–502.

Nunez, P. L., and Srinivasan, R. (2006). *Electric Fields of the Brain: The Neurophysics of EEG.* New York, NY: Oxford University Press.

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J. M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* 2011.

Opitz, B., Ferdinand, N. K., and Mecklinger, A. (2011). Timing matters: the impact of immediate and delayed feedback on artificial language learning. *Front. Hum. Neurosci.* 5:8. doi: 10.3389/fnhum.2011.00008

Osterhout, L., and Holcomb, P. J. (1995). Event-related brain potentials elicited by syntactic anomaly. *J. Mem. Lang.* 31, 785–806.

Osterhout, L., McLaughlin, J., Kim, A., Greenwald, R., and Inoue, K. (2005). "Sentences in the brain: real-time reflections of sentence comprehension and language learning," in *The On-line Study of Sentence Comprehension: Eyetracking, ERP, and Beyond,* eds M. Carreiras and J. C. Clifton (New York, NY: Psychology Press), 271–308.

Osterhout, L., McLaughlin, J., Pitkänen, I., Frenck-Mestre, C., and Molinaro, N. (2006). Novice learners, longitudinal designs, and event-related potentials: a means for exploring the neurocognition of second language processing. *Lang. Learn.* 56, 199–230.

Osterhout, L., Poliakov, A., Inoue, K., McLaughlin, J., Valentine, G., Pitkanen, I., Frenck-Mestre, C., and Hirschensohn, J. (2008). Second language learning and changes in the brain. *J. Neurolinguistics* 21, 509–521.

Overbeek, T. J. M., Nieuwenhuis, S., and Ridderinkhof, K. R. (2005). Dissociable components of error processing: on the functional significance of the vis-á-vis the ERN/N_e. *J. Psychophysiol.* 19, 319–329.

Owen, A. M., McMillan, K. M., Laird, A. R., and Bullmore, E. (2005). N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. *Hum. Brain Mapp.* 25, 46–59.

Pauli, P., Lutzenberger, W., Rau, H., Birbaumer, H., Rickard, T. C., Yaroush, R. A., and Bourne, L. E. (1994). Brain potentials during mental arithmetric: effects of extensive practice and problem difficulty. *Brain Res. Cogn. Brain Res.* 2, 21–29.

Penke, M., Janssen, U., and Eisenbeiss, S. (2004). Psycholinguistic evidence for the underspecification of morphosyntactic features. *Brain Lang.* 90, 423–433.

Pinheiro, J. C., and Bates, D. M. (2000). *Mixed-effects Models in S and S-PLUS.* New York: Springer-Verlag.

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., and Nieuwenhuis, S. (2004a). The role of the medial frontal cortex in cognitive control. *Science* 306, 443.

Ridderinkhof, K. R., Van Den Wildenberg, W. P. M., Segalowitz, S. J., and Carter, C. S. (2004b). Neurocognitive mechanisms of cognitive control: the role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning. *Brain Cogn.* 56, 129–140.

Romero, S. G., McFarland, D. J., Faust, R., Farrell, L., and Cacace, A. T. (2008). Electrophysiological markers of skill-related neuroplasticity. *Biol. Psychol.* 78, 221–230.

Rossi, S., Gugler, M. F., Friederici, A., and Hahne, A. (2006). The impact of proficiency on syntactic second-language processing of German and Italian: evidence from event-related potentials. *J. Cogn. Neurosci.* 18, 2030–2048.

Sabourin, L., and Stowe, L. A. (2008). Second language processing: when are first and second languages processed similarly? *Second Lang. Res.* 24, 397–430.

Sabourin, L., Stowe, L. A., and de Haan, G. J. (2006). Transfer effects in learning a second language grammatical gender system. *Second Lang. Res.* 22, 1–19.

Sabourin, L. L. (2003). *Grammatical Gender and Second Language Processing: an ERP Study*. Unpublished PhD, Rijksuniversiteit Groningen, Groningen.

Schlenker, P. (1999). La flexion de lB'adjectif en allemand: la morphologie de haut en bas [The inflection of adjectives in German: Morphology from top to bottom]. *Recherches linguistiques de Vincennes*, 28, 115–132.

Sebastian-Gallés, N., Rodríguez-Fornells, A., de Diego-Balaguer, R., and Díaz, B. (2006). First- and second-language phonological representations in the mental lexicon. *J. Cogn. Neurosci.* 18, 1277–1291.

Tokowicz, N., and MacWhinney, B. (2005). Implicit and explicit measures of sensitivity to violations in second language grammar: an event-related potentials study. *Stud. Second Lang. Acquisition* 27, 173–204.

van der Helden, J., Boksem, M. A. S., and Blom, J. H. G. (2010). The importance of failure: feedback-related negativity predicts motor learning efficiency. *Cereb. Cortex* 20, 1596–1603.

Weber-Fox, C., and Neville, H. J. (1996). Maturational constraints on functional specializations for language processing: ERP and behvioral evidence in bilingual speakers. *J. Cogn. Neurosci.* 8, 231–256.

Wunderlich, D. (1997). "Der unterspezifizierte Artikel. [The underspecified article]," in *Sprache im Fokus*, M. Schwarz, K.-H. Ramers and C. Düjrscheid (Tübingen: Niemeyer), 47–55.

Zwicky, A. M. (1986). German adjective agreement in GPSG. *Linguistics* 24, 957–990.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 March 2011; paper pending published: 16 May 2011; accepted: 21 August 2011; published online: 20 September 2011.

Citation: Davidson DJ and Indefrey P (2011) Error-related activity and correlates of grammatical plasticity. Front. Psychology ***2****:219. doi: 10.3389/fpsyg.2011.00219*

This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.

Copyright © 2011 Davidson and Indefrey. This is an open-access article subject to a non-exclusive license between the authors and Frontiers Media SA, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and other Frontiers conditions are complied with.