

# An Event-related Potential Study on Changes of Violation and Error Responses during Morphosyntactic Learning

Douglas J. Davidson<sup>1,2</sup> and Peter Indefrey<sup>1,2</sup>

## Abstract

■ Based on recent findings showing electrophysiological changes in adult language learners after relatively short periods of training, we hypothesized that adult Dutch learners of German would show responses to German gender and adjective declension violations after brief instruction. Adjective declension in German differs from previously studied morphosyntactic regularities in that the required suffixes depend not only on the syntactic case, gender, and number features to be expressed, but also on whether or not these features are already expressed on linearly preceding elements in the noun phrase. Violation phrases and matched controls were presented over three test phases (pretest and training on the first

day, and a posttest one week later). During the pretest, no electrophysiological differences were observed between violation and control conditions, and participants' classification performance was near chance. During the training and posttest phases, classification improved, and there was a P600-like violation response to declension but not gender violations. An error-related response during training was associated with improvement in grammatical discrimination from pretest to posttest. The results show that rapid changes in neuronal responses can be observed in adult learners of a complex morphosyntactic rule, and also that error-related electrophysiological responses may relate to grammar acquisition. ■

## INTRODUCTION

Many have proposed that grammatical processing is highly resistant to reorganization during adulthood, or possibly immutable after the closing of hypothesized sensitive periods for grammar acquisition (Sakai, 2005; DeKeyser, 2000; Johnson & Newport, 1989, 1991; Lenneberg, 1967). However, recent observations show that electrophysiological signatures of grammatical violation detection can change rapidly in adult learners during the course of grammatical learning (Mueller, Hahne, Fujii, & Friederici, 2005; Osterhout, McLaughlin, Kim, Greenwald, & Inoue, 2005; Petersson, Forkstam, & Ingvar, 2004; Friederici, Steinhauer, & Pfeifer, 2002). The present work investigates whether event-related potential (ERP) effects associated with the violation of a complex morphosyntactic rule of German can be seen early in adult learning, and whether there are other ERP components related to the grammar acquisition process.

The focus of past research on second-language (L2) sentence processing has been on L1–L2 group comparisons on tasks that assess sensitivity to grammatical or semantic errors. This past work has shown that L2 electrophysiological responses are essentially similar to

those of native speakers, but depend on variables such as the age of L2 learning onset, recency of L2 exposure, and the L2 proficiency of the learner at the time of test. For example, the P600 response, a positive potential difference observed in violation–control contrasts approximately 500–800 msec after a violation word primarily on posterior sensors, has been shown in response to a wide variety of L1 grammatical violations (Kutas, Van Petten, & Kluender, 2006; Osterhout & Holcomb, 1995; Hagoort, Brown, & Groothusen, 1993), but appears to be more variable in L2 comprehenders. Weber-Fox and Neville (2001) showed an inverse relationship between the onset age of second-language (L2) learning and the size and variability of a P600 response to L2 grammatical violations in Chinese learners of English. Similarly, Hahne (2001) demonstrated a P600 response to L2 phrase structure violations in proficient Japanese learners of German, but did not observe an early anterior negativity sometimes seen in L1 studies of the same violation. An atypical P600 response to a phrase structure violation was observed by Hahne and Friederici (2001) in late L2 learners, and Ojima, Nakata, and Kakigi (2005) did not observe a P600 to English (L2) subject–verb agreement violations in Japanese (L1) late learners, although an N400 response was observed in higher-proficiency learners. Tokowicz and MacWhinney (2005) observed a P600 response in early English learners of Spanish to Spanish tense and gender violations. Rossi, Gugler, Friederici, and Hahne (2006)

<sup>1</sup>F. C. Donders Centre for Cognitive Neuroimaging, <sup>2</sup>Max Planck Institute for Psycholinguistics

observed both early LAN and P600 responses in highly proficient German-L2 and Italian-L2 learners, but an absence of LAN responses and a delayed P600 in lower proficient learners. In sum, responses to grammatical violations are variable in second-language comprehenders, but it does appear that neurophysiological correlates of grammatical error detection can be observed in those who have acquired some degree of L2 proficiency.

More recent work has focused on the outcome of L2 learning over shorter periods of acquisition using longitudinal or learning-based experimental designs. These designs control some of the confounding factors that are present when groups of language users with different onset ages or proficiencies are compared in cross-sectional comparisons (Osterhout, McLaughlin, Pitkanen, Frenck-Mestre, & Molinaro, 2006). For example, Mueller et al. (2005) observed P600 violation effects for word category and case marking violations after German (L1) speakers learned to distinguish between grammatical and ungrammatical sentences of a restricted subcomponent of Japanese (L2) grammar. P600 effects were observed in the learners for both types of violations. Although the case marking violations led to an N400-like violation response in native Japanese speakers, this response was not present in the learners. Using the same design, Mueller, Hirotsu, and Friederici (2007) reported an earlier negativity for Japanese ungrammatical nominative case marking in German learners, but not for ungrammatical accusative case marking violations. Finally, Osterhout et al. (2005) observed an N400-like response to French (L2) morphosyntactic violations in English (L1) adult learners. Later in learning, a P600 violation response was observed to the same type of violation (see also Osterhout et al., 2006, in press). These results show that longitudinal studies of learners can effectively detect changes in grammatical violation responses that occur within months or weeks of L2 learning.

Note, however, that the types of morphosyntactic violations that have been studied so far (tense, case, number) all involved a simple mismatch of syntactic feature specifications. It is therefore unclear what type of violation responses can be seen when learners acquire more complex morphosyntactic regularities involving contingencies between two different types of syntactic information. Such a regularity exists in German adjective declension, where the required suffixes depend not only on the syntactic case, gender, and number features to be expressed but also on an additional syntactic rule according to which these syntactic features are to be expressed only on the first inflectable element of a noun phrase [NP] (Schlenker, 1999). German determiners and adjectives show agreement with their head nouns for the features of case, gender, and number. The degree to which these features are expressed by an adjective suffix differs between three classes termed the strong, weak, and mixed declension classes. Unlike noun declension classes in German (and many other languages), adjective

declension classes in German are not (arbitrary) lexical properties of the adjective. An adjective may take suffixes of all three classes, depending on the preceding elements of the NP. This dependency is considered to be a syntactic dependency rather than a semantic or phonological 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 NP. If the adjective is the first inflectable element, it takes on a strong suffix. The suffix *-em* in “mit kleinem<sub>+Dat, -F, -Pl</sub> Fenster<sub>-F-M</sub>” [with a small window], for example, specifies dative case, non-feminine gender, and singular number. By contrast, if the adjective is preceded by a definite determiner that expresses the feature information, the adjective has a weak suffix that is compatible with the feature specification of the determiner but does not express the features itself (“mit dem<sub>+Dat, -F, -Pl</sub> kleinen<sub>∅</sub> Fenster<sub>-F-M</sub>” [with the little window]), in this case, the suffix *-en*.

The first goal of the present experiment was to investigate whether Dutch native speakers, who were taught a subset of the German adjective declension system that, nonetheless, preserved the essential properties of the full system, would show violation responses similar to those observed with other morphosyntactic violations. Although a closely related West Germanic language, Dutch lacks most of the morphosyntactic marking that is found in German. Similar to English, residual case marking is found in pronouns but not in nouns or adjectives. There is a two-way gender distinction (neuter gender and nonneuter, so-called common gender) that is marked on singular definite determiners and adjectives. Similar to German, however, gender marking on the adjective is contingent upon the preceding context. Gender is only marked following indefinite determiners that do not themselves mark gender (e.g., “een groot<sub>neuter</sub> huis<sub>neuter</sub>” [a big house], “een grote jongen<sub>neuter</sub>” [a big boy], “het<sub>neuter</sub> grote huis<sub>neuter</sub>” [the big house], “de grote jongen<sub>neuter</sub>” [the big boy]). Although this contingency affects a single form rather than full case, gender, and number paradigms as in German, it might be described as conforming to the same rule according to which syntactic features (in this case, gender only) are to be expressed only on the first inflectable element of an NP. We therefore expected Dutch learners to acquire the context dependence of German adjective inflection.

A second goal of the experiment was to investigate whether violation responses would be sensitive to the feature specifications of incorrect adjective forms. German adjective forms show a high degree of syncretism with only five different suffixes expressing the 72 possible combinations of case, gender, number, and declension class. Linguistic analyses (Clahsen, Sonnenstuhl, Hadler, & Eisenbeiss, 2001; Schlenker, 1999; Cahill & Gazdar, 1997; Wunderlich, 1997; Blevins, 1995; Zwicky, 1986; Bierwisch, 1967) have dealt with the syncretism

by using a hierarchical feature specification, such that some suffixes are specified for case, number, and gender features and others are treated as default or “elsewhere” forms with reduced or no specification (see also Penke, Janssen, & Eisenbeiss, 2004). We investigated whether the feature specification of adjective forms would modulate electrophysiological declension class violation responses by separately analyzing declension class violations involving incorrect strong adjective forms having rich feature specifications and incorrect weak adjective forms having poor feature specifications.

## EXPERIMENT

The interaction of gender and the strong and weak forms of the adjectives was studied using simple case-marked NPs. We chose the subset of dative case neuter-singular and dative case feminine-singular NPs (see Table 1) to investigate how Dutch subjects learn these contingencies. All NPs were presented in the context of the dative-requiring preposition “mit” ([with]), so that participants would not be required to learn grammatical case alternations. Rather, the rules to be learned included the contingency of whether or not the adjective suffix should express gender information. This depended on the presence or absence of a definite determiner.

Participants read short prepositional phrases (e.g., “mit dem kleinen Fenster” [with the little window]) presented one word at a time and judged whether the phrases formed acceptable German phrases or not. For each participant, there were two EEG recording sessions. On the first day, subjects first took part in a pretest phase without feedback, and then on the same day learned the grammatical contrasts with feedback in a training phase. In the second session 1 week later, they applied the knowledge that they had acquired in posttest without feedback.

Declension violations were created by using adjective suffixes compatible with the required case, number, and gender but from the wrong declension class (see Table 2A). Given the rule that feature information should be specified only if not already specified on a preceding element of the NP, we distinguished two types of declension violations occurring on either four-word or three-word phrases. In the four-word declension violation, the adjective carried an incorrect strong suffix, thereby redundantly specifying gender. In the three-word

declension violation, the adjective carried an incorrect weak suffix resulting in a lack of gender specification.

In addition to declension violations, we also tested for ERP responses to gender violations (see Table 2B). Gender violations were created by using a noun whose grammatical gender did not match the preceding determiner/adjective combination which was correctly inflected with respect to case, number, and adjective declension.

Additional filler phrases contained determiners and adjectives that were incorrectly specified for nominative/accusative case (“mit \*das kleine Fenster”; “mit \*die kleine Maus”; “mit \*kleines Fenster”; “mit \*kleine Maus”). Note that German native speakers would perceive these as case violations, whereas the Dutch learners were not presented with correct nominative/accusative marked phrases, so that they simply learned that certain forms of determiners and adjectives were incorrect but could not identify a specific type of violation. Finally, we added correct filler phrases to balance the number of correct and incorrect phrases.

## METHODS

### Participants

Twenty-two native Dutch speakers (21 women, 17 right-handed, average age = 23.0 years) were recruited with posted advertisements from the university community at Radboud University in Nijmegen, The Netherlands. The advertisements described a generic EEG experiment, and did not refer to language learning or to German instruction. Twenty out of 22 participants completed both sessions (the two missing subjects can be considered missing at random: one canceled due to a power failure, the second due to a forgotten appointment). Most participants had previous coursework experience with German during high school. Before completing the EEG tasks, all participants completed a European Reference Frame multiple choice assessment of German prepared by the Goethe Institute ([www.goethe.de](http://www.goethe.de); see Table 3). In addition to the Dutch speakers, a control group of 23 native German speakers (20 women, 22 right-handed, average age = 26.0 years) recruited from the same community completed the same task with the same materials (equivalent to the posttest phase; see Design and Procedure).

### Design and Procedure

As described earlier, the first of the two experimental sessions included three parts: a pretest phase, a learning task, and a training phase. The pretest was conducted to assess participants’ knowledge of German when they started the experiment. During the pretest phase, participants performed the judgment task by classifying the phrases as acceptable or not following a response cue by pressing one of two buttons on a keypad with their right

**Table 1.** German Singular Dative Adjectival Suffixes

	<i>Neuter</i>	<i>Feminine</i>
Strong	–em	–er
Weak	–en	–en

hand. Participants saw half of the total experimental items, but did not receive feedback about their responses. During the learning task (duration from 15 to 20 min), participants studied the entire set of nouns (40) and adjectives (4) on a sheet of paper in a list, completed two paper-and-pencil tests concerning gender and the Dutch–German translation of the nouns, and also read a short description of the grammatical rules involved in the construction of the experimental phrases. In the paper-and-pencil tasks, participants were able to accurately match the nouns and adjectives between Dutch and German, as well as choose the appropriate determiner for the gender of the German terms (both performance levels  $>0.95$ ; see Table 3). The description of the grammatical rules informed participants that the gender of the nouns in the experiment would match the gender of the corresponding Dutch translations, and that the determiner needed to match the gender of the noun, just as in Dutch (“het papier” → “das Papier,” “de soep” → “die Suppe”). Further, they were informed that the determiner also depended on case, and that the phrases presented in the experiment would always start with the preposition “mit,” which assigns dative case in German. Participants were informed that the form of the determiner with dative case would depend on the gender of the noun (e.g., “mit dem Papier,” “mit der Suppe”). Lastly, participants were informed that depending on whether or not the determiner was present in the prepositional phrase, the form of the adjective would be different. If the determiner was not present, a form that depends on gender would be required (“mit rotem Papier,” “mit roter Suppe”). If the determiner *was* present, then rather than the adjective, a form for the determiner that depends on gender was required (“mit dem roten Papier,” “mit der roten Suppe”).

During the training phase, participants again performed the judgment task, but in addition, received feedback after every response. The feedback indicated whether their response was correct (green square with a short high-pitched tone, presented at the response for 200 msec) or incorrect (red square with a short dissonant tone presented at the response for 1000 msec). The feedback did not provide any additional information about the source of the error, or the correct version of the phrase.

The second experimental session consisted of the judgment task like the pretest (i.e., without feedback), but with all 40 nouns and 4 adjectives. Thus, the posttest phase assessed performance on items that had each been observed once in the pretest or training phases. For half of the subjects, 20 items were seen in the pretest and the other 20 were seen in the training phase, but all 40 were seen again in the posttest phase. The other half of the subjects saw the reverse assignment. The posttest phase was conducted 1 week after the initial session (average = 8.4 days,  $SD = 4.2$ , range = 1–15 days).

## Materials and Procedure

Four common German adjectives (klein, gross, gut, schlecht) and 40 common German nouns were chosen to serve as stimulus materials. The nouns were chosen so that they had the same corresponding gender of the Dutch noun translation (*neuter*: Fenster, Haus, Kind, Pferd, Schaf, Schwein, Ohr, Bein, Herz, Buch, Glas, Kaninchen, Bett, Messer, Dorf, Institut, Mädchen, Museum, Geschenk, Hemd; *feminine*: Tür, Schule, Frau, Kuh, Ziege, Katze, Nase, Hand, Niere, Zeitung, Tasse, Maus, Couch, Gabel, Stadt, Universität, Schwester, Ausstellung, Spende, Hose). The critical words included the adjective and noun for the various conditions (Table 2).

In the first experimental session, participants first filled out paper-and-pencil tests while the electrodes were applied, and this was followed by the pretest phase, then the learning test, and then the training phase. In the second experimental session, after electrode application, participants completed the posttest. EEG was recorded during each phase of the pretest, training, and posttest. During these recordings, the words in each phrase were presented on a CRT monitor for 250 msec, with an ISI of 500 msec using 26-point Arial white characters on a black background. Each trial began with a white fixation cross (duration = 1 sec), and the last word of the phrase was followed by the same white fixation cross (500 msec), followed by a yellow fixation cross that remained on the screen until participants provided their classification response.

## Apparatus

EEG was recorded from 62 channels using battery-powered BrainVision BRAINAMP Series amplifiers (Brain Products GmbH, Munich, Germany). Signals were sampled at 500 Hz, low-pass filtered at 200 Hz (3 dB reduction). Electrodes were applied to an equivalent interelectrode distance Easy-Cap (Brain Products; see Figure 1 for the electrode arrangement). Impedance levels were kept below 50 k $\Omega$  at the electrode–skin interface, with input impedance at the amplifiers at 10 M $\Omega$  (see Ferree et al., 2001). 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.

## Data Analysis

Behavioral response data were summarized by computing measures of a corrected hit rate, as well as discrimination ( $d'$ ). The participants' classification of a grammatical

**Table 2.** Experimental Conditions (Average Number of Trials for Pretest, Training, Posttest, and Native-speaker Controls in Parentheses)

(A) Declension Class Violation Contrast

Condition 1: Four-word declension violation–control pairs

*Violation (8.6, 15.6, 32.5, 35.6)*

mit dem<sub>[neut]</sub> \*kleinem<sub>[neut]</sub> Fenster<sub>[neut]</sub>

mit der<sub>[fem]</sub> \*kleiner<sub>[fem]</sub> Maus<sub>[fem]</sub>

*Control (9.4, 17.9, 32.4, 28)*

mit dem<sub>[neut]</sub> kleinen<sub>[]</sub> Fenster<sub>[neut]</sub>

mit der<sub>[fem]</sub> kleinen<sub>[]</sub> Maus<sub>[fem]</sub>

Condition 2: Three-word declension violation–control pairs

*Violation (12.4, 15.9, 33.5, 30.1)*

mit \*kleinen<sub>[]</sub> Fenster<sub>[neut]</sub>

mit \*kleinen<sub>[]</sub> Maus<sub>[fem]</sub>

*Control (10.3, 18.1, 33.7, 31.7)*

mit kleinem<sub>[neut]</sub> Fenster<sub>[neut]</sub>

mit kleiner<sub>[fem]</sub> Maus<sub>[fem]</sub>

(B) Gender Violation Contrast

Condition 3: Gender violation–control pairs

*Violation (11.6, 14.4, 30.4, 28.9)*

mit dem<sub>[neut]</sub> kleinen<sub>[]</sub> \*Maus<sub>[fem]</sub>

mit der<sub>[fem]</sub> kleinen<sub>[]</sub> \*Fenster<sub>[neut]</sub>

*Control (10.0, 17.9, 33.0, 30.1)*

mit dem<sub>[neut]</sub> kleinen<sub>[]</sub> Fenster<sub>[neut]</sub>

mit der<sub>[fem]</sub> kleinen<sub>[]</sub> Maus<sub>[fem]</sub>

*Violation (12.7, 14.7, 29.4, 24.9)*

mit kleinem<sub>[neut]</sub> \*Maus<sub>[fem]</sub>

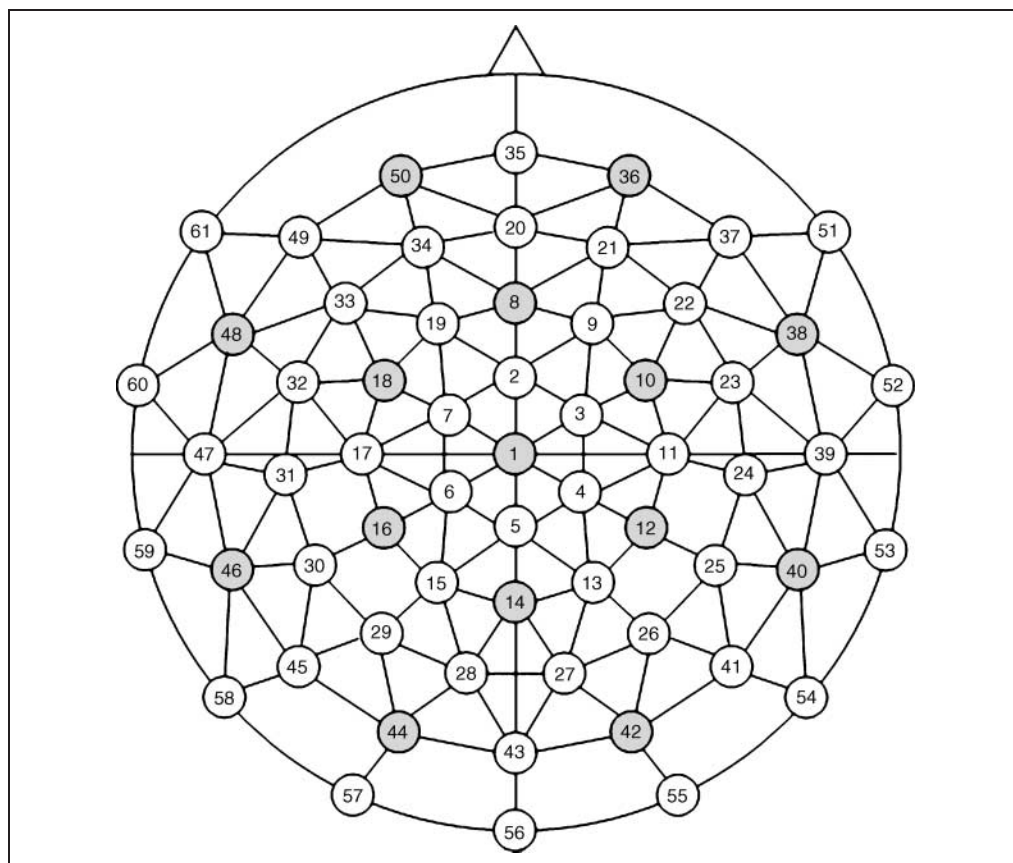
mit kleiner<sub>[fem]</sub> \*Fenster<sub>[fem]</sub>

*Control (11.4, 18.2, 33.8, 31.7)*

mit kleinem<sub>[neut]</sub> Fenster<sub>[neut]</sub>

mit kleiner<sub>[fem]</sub> Maus<sub>[fem]</sub>

**Figure 1.** Electrode array layout. Electrode numbers corresponding to approximate 10–20 locations are shown in gray.



phrase as acceptable was coded as a hit, whereas the classification of an ungrammatical phrase as grammatical was coded as a false alarm. The corrected hit rate was computed as proportion of hits – proportion of false alarms. The  $d'$  measure was computed as  $z(\text{hit}) - z(\text{false alarm})$ , where  $z$  is the inverse of a cumulative Gaussian distribution with a zero mean and unit standard deviation.

Fixed and random effects for the behavioral measures, in addition to several covariates (see Table 3), were modeled using a general linear mixed effects model approach on the basis of restricted maximum likelihood parameter estimates (Pinheiro & Bates, 2000). Likelihood ratio tests (LRT) were used to test for main effects and interactions. The covariates included average self-rated proficiency, age of initial German language education, German language proficiency as measured by the Goethe Institute Test (Goethe-Institut, 2005), nonlinguistic reasoning as measured by Raven's Advanced Progressive Matrices (Raven, 1962), performance on the gender-matching component of the learning period of the first session, and the difference in average response time between incorrect and correct responses in the training phase (error RT).

The recorded EEG data were screened for eye movement, muscle, and other noise artifacts, filtered with a

low-pass filter (two-pass 6th-order Butterworth finite impulse response) with square-root half-maximum attenuation at 20 Hz, re-referenced to an average reference, and segmented into 700-msec epochs consisting of 200 msec before the onset of the critical word (CW) and 500 msec following the CW. The proportion of correct-answer trials excluded by artifact rejection was 0.02 (0.01, 0.03) for the pretest, 0.04 (range 0.02, 0.06) for training, and 0.04 (range 0.02, 0.07) for the posttest. The resulting epochs were baselined with respect to the 200 msec baseline interval and averaged according to experimental condition. Only trials with correct responses were included in the violation–control ERP contrasts (see Table 2), and only those participants with at least 10 observations in both violation and control conditions. 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: 0–120 msec (an interval we will term the  $N_E$ ) and 150–250 msec ( $P_E$ ). The statistical significance of observed differences in the electrophysiological data was assessed using a clustering and randomization test (Maris & Oostenveld, 2007; Maris, 2004). In this approach, a randomization distribution of cluster statistics is constructed and used to evaluate statistically

**Table 3.** Participant Characteristics

<i>Participant Variable</i>	<i>Component</i>	<i>Score</i>
Self-rated skill <sup>a</sup>	Speaking	2
	Listening	3
	Writing	2
	Reading	3
	Grammar	2
	Pronunciation	3
Self-rated attitude <sup>a</sup>	Valence toward L2 usage	2
	Confidence in L2 usage	2
	Importance of L2 usage	3
	Valence toward L2 learning in general	4
	Ease of L2 learning in general	3
Age/exposure	Age (mean, range in years)	23.0, 11
	Age of acquisition (mean, years)	11.8
	L2 learning duration (median, years)	4
	Expose rate (mean; hr/month)	0.3
Test performance	Goethe Institute Test (mean; out of 30)	16.8 <sup>b</sup>
	NL–DE noun/adjective matching (mean; out of 44)	43.7
	Gender choice (mean; out of 40)	38.6

<sup>a</sup>Median ratings on a 5-point scale with 5 as highest level of skill, or most positive attitude.

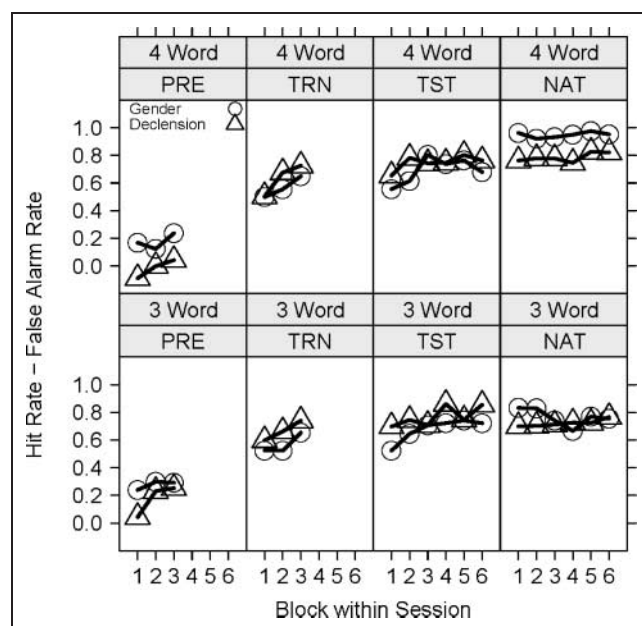
<sup>b</sup>European Reference Frame Level B2 (Independent User).

significant differences between conditions. In particular,  $t$  statistics are computed for each channel, and a clustering algorithm forms groups of channels based on these tests. The sum of the  $t$  statistics in an electrode group is then used as a cluster-level statistic ( $sum-T$ ), which is then tested for significance using a randomization test (using 4000 runs). For clusters of activity, the average ERP effect is reported for groups of channels in a cluster (see Figure 1 for the channel locations), and in addition, the average ERP effect in the clusters is related to the behavioral measures using mixed effects multiple regression.

## RESULTS

### Behavioral Performance

Figure 2 shows that the corrected hit rate increased as a function of experiment phase, and also as a function of trial block within each phase. In order to examine training effects within phase, pretest and training phases were divided into three blocks (the posttest phase contained twice as many stimuli, so it was divided into six blocks; see Figure 2). A mixed effects analysis for this measure of discrimination as the dependent measure showed main effects of phase [training higher than pretest,  $t(682) = 5.143, p < .001$ ; test higher than pretest,  $t(682) = 6.265, p < .001$ ], and block [discrimination improved with block,  $t(682) = 4.739, p < .001$ ], as well as an interaction of phase and type of violation [advantage of training over pretest higher for strength violations,  $t(682) = 2.424, p = .016$ ; advantage of posttest over pretest higher for strength violations,  $t(682) = 2.439, p = .015$ ]. Although knowledge of German (as



**Figure 2.** Discrimination (hit rate – false alarm rate) over blocks within session by type of violation.

measured by the Goethe test) predicted performance in the pretest [ $t(17) = 2.081, p = .053$ ], there was no interaction between the pre–post gain and scores on the Goethe test ( $p > .5$ ). A comparison of posttest performance on the items that were seen during the pretest versus those that were seen during training showed no main effect, and there were no interactions with the type of violation.

A comparison of the learner performance on the posttest phase with the native classification revealed a main effect of block [ $t(433) = 2.075, p = .038$ ], a main effect of type of violation [ $t(433) = 2.146, p = .032$ ], a main effect of group [ $t(38) = 2.026, p = .050$ ], as well as an interaction with the type of violation with group [ $t(433) = -3.110, p = .002$ ], such that the corrected hit rate for the native speakers for gender violations was higher than for the learners.

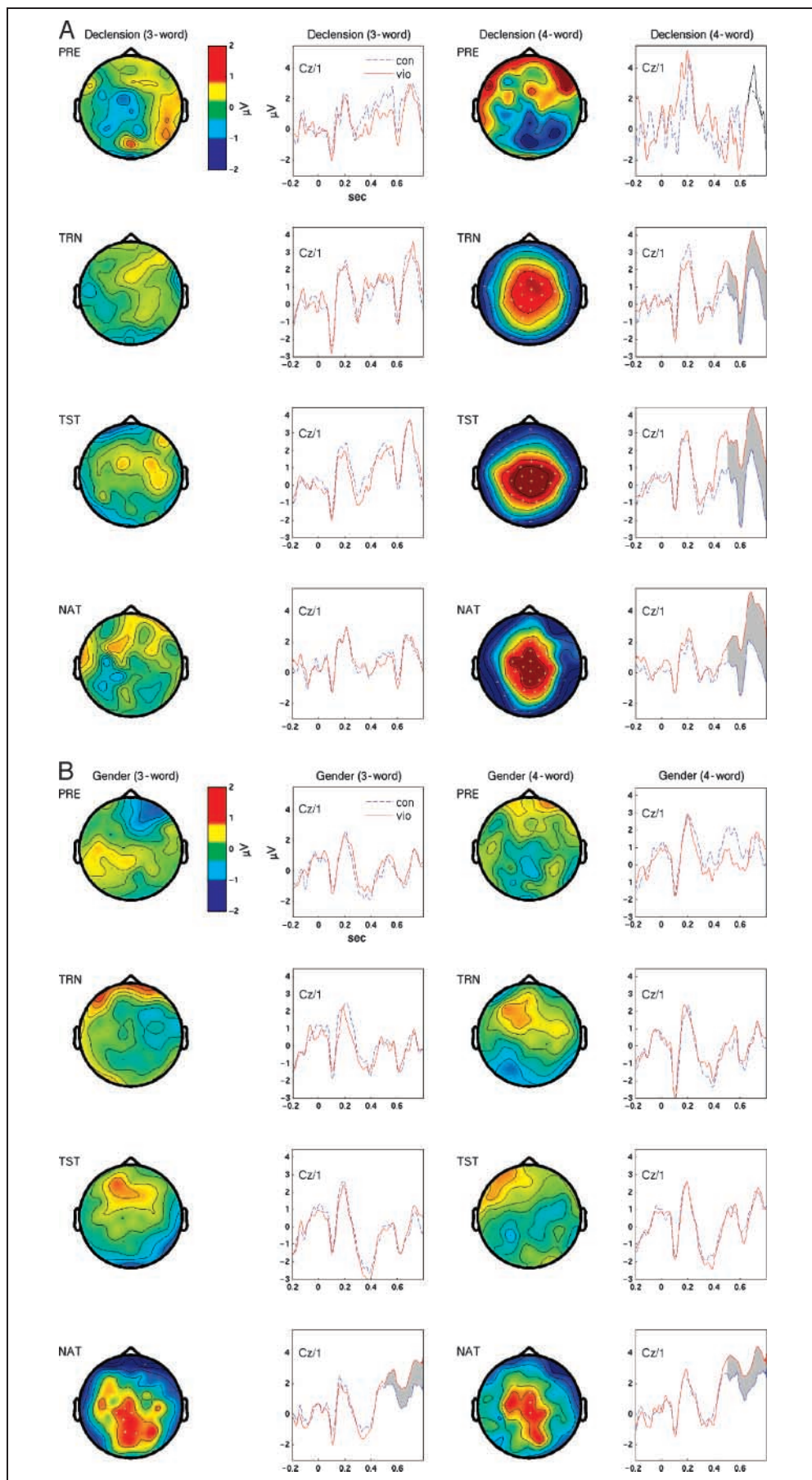
### Event-related Potentials

The cluster analysis identified significant P600-like effects in the training and posttest phases for the four-word declension violation contrast, similar to native German speakers, as shown in Figure 3A. There were no significant violation responses in the pretest phase for the learners. Also, there were no effects for the gender contrast in either the pre-, training, or posttest phases in the learners (Figure 3B).

In the training phase, for the four-word declension contrast (Figure 3A), the P600 effect ( $1.342 \mu\text{V}$ ) was present in one cluster ( $sum-T = 74.19, p < .001$ ) on 21 central and posterior sensors (1–19, 25, 28), with a corresponding negative effect ( $1.264 \mu\text{V}$ ) occurring in one cluster ( $sum-T = 62.41, p < .001$ ) over 20 peripheral sensors (35–40, 46–57, 60–61). In the posttest phase, there was a significant P600 effect for the four-word declension contrast ( $1.79 \mu\text{V}$ ) in one cluster ( $sum-T = 92.9, p < .001$ ) on central and posterior sensors (1–7, 10–17, 24–31), and corresponding to this, a negative effect ( $-1.70 \mu\text{V}$ ) in a single cluster ( $sum-T = -86.5, p < .001$ ) on frontal sensors (20, 22, 33–40, 47–57, 60–61). Similarly, in native German speakers, there was a P600 effect for the four-word declension contrast ( $1.65 \mu\text{V}$ ) in one cluster ( $sum-T = 97.28, p < .001$ ) on 23 central and posterior sensors (1–18, 26–30), with corresponding negative effects occurring in three clusters ( $-1.68 \mu\text{V}$ ,  $sum-T = -28.60, p = .01$ ;  $-1.77 \mu\text{V}$ ,  $sum-T = -10.66, p = .04$ ;  $-1.53 \mu\text{V}$ ,  $sum-T = -9.79, p = .05$ , respectively) on various peripheral posterior sensors (Cluster 1: 45, 53–59; Cluster 2: 48, 49, 61; Cluster 3: 39, 51, 52). In the three-word declension contrast, there were no significant effects in either the learners or the native German speakers.

The native German speakers also showed significant P600 effects for both the three-word and four-word gender contrasts (Figure 3B). For the three-word gender contrast, the P600 effect ( $1.20 \mu\text{V}$ ) occurred in one cluster ( $sum-T = 10.54, p = .05$ ) on four central sensors

**Figure 3.** ERP traces and topographic plots for the violation – response (violation – control) in the P600 time window by number of words and phase (pretest [PRE], training [TRN], and posttest [TST]), for the (A) declension contrasts, and (B) the gender contrasts. Native speaker controls (NAT) are plotted along the bottom. The ERP traces show channel Cz/1 with significant P600 amplitude differences shaded in gray. The control conditions are plotted with a dashed blue line and the violation condition as a solid red line. In the topographic plots, electrodes within significant clusters are plotted for both positive (“+”) and negative (“-”) average potentials.





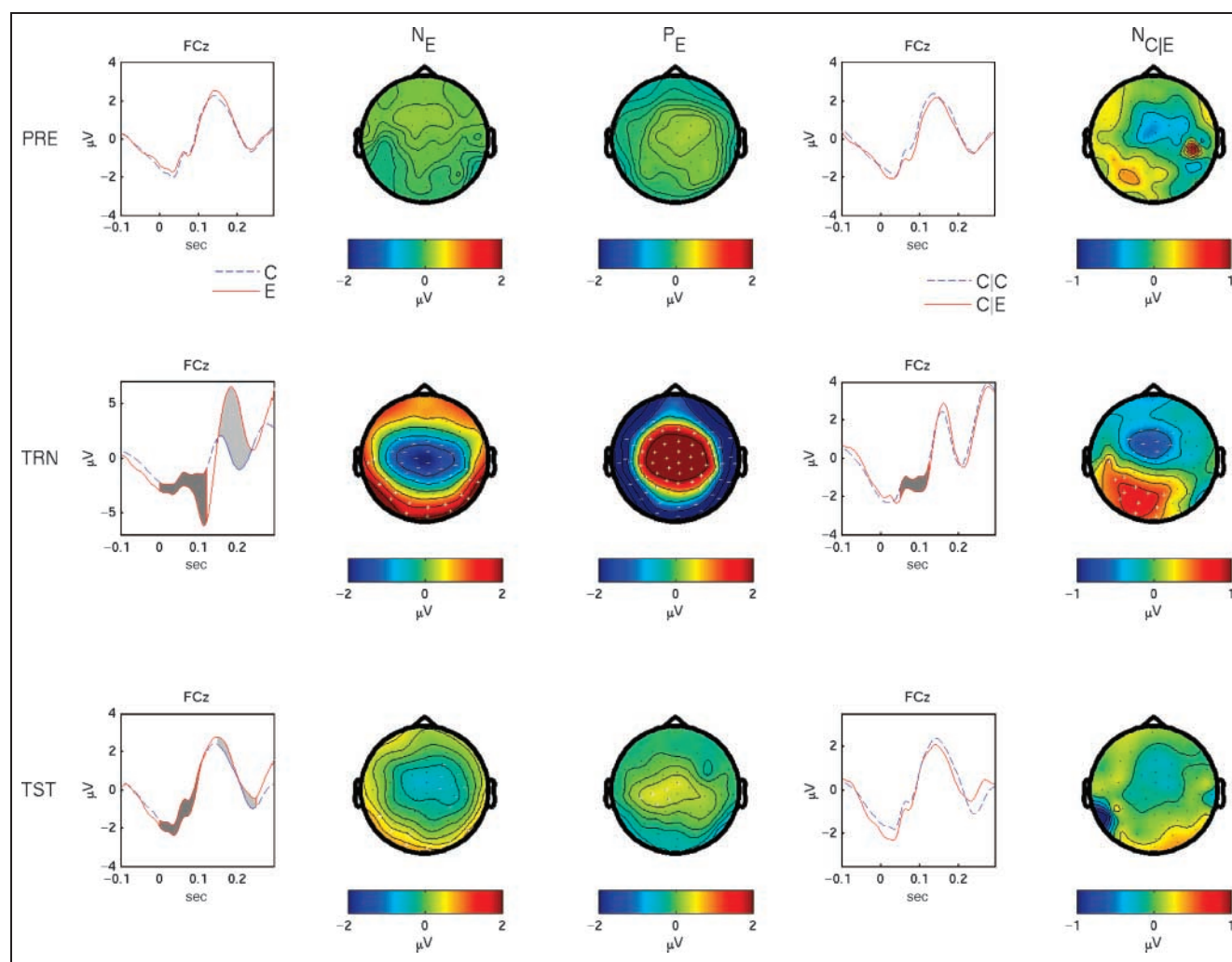
(1–3, 5), with no corresponding negative potential effect. For the four-word gender contrast, the effect ( $1.15 \mu\text{V}$ ) was present in one cluster ( $\text{sum-}T = 16.91, p = .012$ ) on seven central–posterior sensors (5, 6, 15, 16, 27, 28, 31), also with no corresponding negative effect.

### Error-related Potentials

In addition to the evoked responses to words within the phrases, additional evoked responses could be observed to the classification decisions made by participants. Figure 4 shows the topography of the average evoked potential for incorrect responses minus correct responses in the pretest, training, and posttest phases for the time window of 0–120 msec after the onset of the classification response (baselined with respect to the average

potential in the window  $-100$  to  $0$  msec), collapsed over all of the phrase types. During the training phase a medial-frontal ( $N_E$ ) effect was present (average:  $-1.20 \mu\text{V}$ ) in one cluster ( $\text{sum-}T = -116.7, p < .001$ ) on 16 fronto-central sensors (1–3, 10–19, 22–24), as well as a corresponding positive potential ( $1.38 \mu\text{V}$ ) in one cluster ( $\text{sum-}T = 95.2, p < .001$ ) on 15 posterior sensors (40–46, 53–59, 60). During the posttest phase, a negativity was present ( $-0.36 \mu\text{V}$ ) in one cluster ( $\text{sum-}T = 18.9, p = .037$ ) on eight right fronto-central sensors (3, 9, 10–12, 21–23), as well as a corresponding positive potential ( $0.55 \mu\text{V}$ ) in one cluster ( $\text{sum-}T = 21.2, p = .030$ ) on six posterior sensors (44, 54–59). The  $N_E$  pattern was not present on the pretest.

Figure 4 shows that a greater positive error potential ( $P_E$ ) was observed following errors (relative to correct trials) in the time window of 150–250 msec following the



**Figure 4.** ERP traces and topographic plots for the error-related responses. Column 1 shows ERP traces at FCz for the pretest (PRE), training (TRN), and posttest (TST) phases averaged over type of violation. Correct responses are plotted as a dashed blue line, incorrect responses as a solid red line, the significant  $N_E$  is indicated with dark gray shading, the  $P_E$  in light gray. Columns 2 and 3 show the topography of the  $N_E$  and  $P_E$  (error – correct) for same phases. Column 4 shows the ERP traces for the conditional negative responses. Correct|correct (C|C) is plotted as a dashed blue line and correct|error (C|E) is plotted as a solid red line, with the significant effect during training shaded in dark gray. Column 5 shows the topography of the difference potential  $C|E - C|C$ . In all topographies, the electrodes within significant clusters are plotted for both positive (“+”) and negative (“-”) average difference potentials.

response (baselined to  $-100$  to  $0$  msec). This error positivity was not present in the pretest phase, but was present on the training phase ( $1.98 \mu\text{V}$ ) in one cluster (sum- $T = 194.6$ ,  $p < .001$ ) over fronto-central sensors (1–26, 29–34), with a corresponding negative potential effect ( $-2.09 \mu\text{V}$ ) in one cluster (sum- $T = -178.7$ ,  $p < .001$ ) on ventral sensors (36–51). The error positivity was also observed in the posttest ( $0.35 \mu\text{V}$ ) in one cluster (sum- $T = 14.8$ ,  $p = .039$ ) on left central sensors (2, 6, 7, 16, 17, 31).

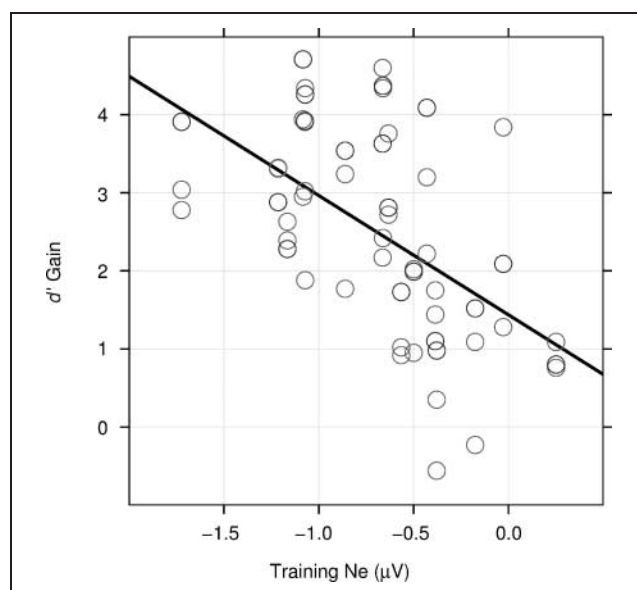
Learners progressed from making relatively more errors early in the task to making relatively fewer errors later in the task, as they acquired grammatical knowledge. The response to a correct trial when the previous trial was an error might therefore show some electrophysiological indication of learning. To investigate this, a conditional negative ERP difference waveform ( $N_{C|E}$ ) was calculated for the training phase trials by comparing the amplitude of the waveforms on correct response trials occurring in two different contexts: Correct response trials that occurred after an *error* on the previous trial, and correct response trials that occurred after a *correct* response on the previous trial. We calculated the average difference amplitudes (correct|error – correct|correct) for the time window of 50–125 msec (based on inspection of the waveforms, note that this is 50 msec later than in the  $N_E$  analysis), baselined to  $-100$  to  $0$  msec before the behavioral response. This analysis revealed a significant negative potential ( $-0.62 \mu\text{V}$ ; see Figure 4) in one cluster (sum- $T = 26.9$ ,  $p = .014$ ) on medial frontal sensors (2, 7–9, 18–21, 33–34), and a corresponding significant positive potential ( $0.54 \mu\text{V}$ ) in one cluster (sum- $T = -34.1$ ,  $p = .004$ ) on left posterior sensors (27–30, 43–47, 56–58). There was no  $N_{C|E}$  effect in the pretest or posttest phases, and a comparable analysis examining a possible positive potential counterpart to the  $N_{C|E}$  did not reveal any effects. The  $N_{C|E}$  effect in training represents a greater negative potential with an anterior distribution like that of the  $N_E$  effect, but it can be distinguished from the  $N_E$  effect because the  $N_{C|E}$  occurs on correct response trials, whereas the  $N_E$  effect is the result of comparing error- and correct-response trials.

### Regression Analyses: Discrimination and Error Responses

Regression analyses were conducted to characterize how behavioral performance was related to the  $N_E$  measures. The  $N_E$  analysis is intended to show how the  $N_E$  (calculated over all types of violations) is related to behavioral discrimination ability. Additional behavioral measures (see Methods) were included as predictors. The regression weights ( $b$ ) for the best-fitting model are listed in each section to indicate the direction and relative magnitude of the reported effects. Plots of fitted parameters versus standardized residuals indicated no evidence of

dependence, and the predictors were not highly correlated (no  $r > .5$ ). Analyses using the P600 effects seen in the violation contrasts, as well as the  $N_{C|E}$  in place of the  $N_E$ , were also conducted, but no significant relations were observed.

The change in discrimination ( $d'$ ) from pretest to posttest was predicted by the amplitude of the  $N_E$  during training, the type of violation, Goethe Institute Test performance, gender test performance, and the error RT. Figure 5 shows the relation between  $d'$  change and training  $N_E$  amplitude. The likelihood ratio tests for each of these effects were as follows:  $N_E$  amplitude (LRT = 5.903,  $p = .015$ ), violation type (LRT = 20.0,  $p < .001$ ), Goethe Institute Test performance (LRT = 4.148,  $p = .04$ ), gender test performance (LRT = 7.143,  $p = .008$ ), and error RT (LRT = 10.537,  $p = .001$ ). The regression parameter for the  $N_E$  effect was negative ( $b = -0.3750$ ), indicating that the larger the (negative) amplitude of the  $N_E$ , the larger the (positive) improvement in  $d'$  from pretest to posttest. For the violation type effect, a post hoc test showed that gender violations were associated with a smaller change in  $d'$  from pretest to posttest than the other two conditions [ $t(35) = -4.1600$ ,  $p < .001$ ]; the other two conditions improved the same amount. The fitted parameter for Goethe Institute Test performance was negative ( $b = -0.3224$ ), indicating that learners who began the experiment with higher proficiency score improved less from pretest to posttest. The parameters for the gender test ( $b = 0.4303$ ) and error RT ( $b = 0.2779$ ) were both positive, indicating that learners who had higher scores on the gender test also improved more in  $d'$ , and learners with a larger difference between error and correct response times improved more as well. This analysis shows that the  $N_E$  amplitude during the



**Figure 5.** The amplitude of the  $N_E$  during training as a function of the amount of improvement in discrimination from pretest to posttest.

training phase predicted improvement in grammatical discrimination, adjusting for the influence of other significant predictors.

## DISCUSSION

P600 responses to grammatical violations emerged in Dutch participants for declension violations but not gender violations after less than a day of training with rules of gender and declension in German NPs. These electrophysiological responses remained after 1 week. The response to the declension violation in the learners was similar to the P600 response seen in native German speakers. This response was observed on a much shorter time scale than in the previous work of Mueller et al. (2005) or Osterhout et al. (2005). The rapidly emerging response may reflect a processing difference between grammatical declension violations and other types of grammatical violations eliciting late positive potentials in L2 learners. To our knowledge, the P600 effect for the type of declension violation used here has not been previously investigated. These results suggest that at the macroscopic scale of brain function observed with EEG, some degree of reorganization of the grammatical processing system is possible after relatively short periods of training in adult language learners. The emergence of the P600 violation responses indicate that populations of neurons react to violations of declension constraints after training, even though they did not do so before training.

Note, however, that our choice of participants favored the learning of a context-dependent morphosyntactic rule in this first study and the results cannot be generalized to other groups of L2 German learners for at least two reasons. First, although Dutch inflectional morphology is poor compared to German, the type of context-dependent morphosyntactic rule we were interested in is also found in Dutch, albeit in a simpler form. It is conceivable that Dutch learners, being familiar with this type of rule, acquired the contingency of preceding context and adjective inflection more easily than learners from other languages would have. Second, it should be noted that our participants had some previous experience with German, so it is likely that the changes in violation responses that we observed reflect not only the acquisition of new grammatical knowledge but also reactivation of previous knowledge. The degree of reactivation of previous knowledge can be estimated by the improvement during the pretest phase, where no instruction or feedback was given. This was relatively low compared to the performance improvement following explicit instruction on the relevant rules. Therefore, it seems that the bulk of the change that we observed was related to learning of the German paradigm that we presented. Therefore, although the P600 and behavioral responses observed in the present experiment cannot be taken as representative of naïve learners, the responses

are likely to be representative of second-language learners who have some previous classroom experience, but are no longer actively using their second language. Finally, in future investigations of this type of morphosyntactic learning, it would be beneficial to include greater variability of items to avoid repetition. In the present design, a relatively small number of adjectives and nouns were used to reduce the lexical learning requirement, leading to sequence repetition. An interesting extension of the present design would be to compare rates of learning for items that repeat versus those that do not, with or without feedback.

Our second research question concerned possible evidence supporting the linguistic assumption of hierarchical feature specifications on adjective suffixes. Although this assumption allows for an elegant description of highly syncretic inflectional paradigms, there is, to date, only some psycholinguistic evidence for processing differences between adjective forms with different feature specifications. Behavioral data by Clahsen et al. (2001) have shown longer lexical decision times for adjective forms specified for more features and a dependency of the size of cross-modal priming effects on feature overlap between prime and target adjective forms. Penke et al. (2004) found behavioral violation effects for incorrectly specified determiners and adjectives only when the incorrect inflectional affixes signaled positive, nonmatching feature values (e.g., adjective plus *-m* or *-r* in a context specifying either accusative masculine singular, dative feminine singular, or genitive plural).

In our study, hierarchical feature specification predicted differential violation responses for incorrect strong forms (e.g., mit dem \*kleinem Fenster) and incorrect weak forms (e.g., mit \*kleinen Fenster). We indeed observed a difference between the two types of declension violations. In contrast to incorrect strong adjective forms, which resulted in a P600 response in both the learners and native German speakers, incorrect weak forms did not result in such a response in any of the two groups. Our findings, therefore, provide evidence for the notion that the processing of adjective forms in normal phrasal comprehension does not simply involve a checking of correct or incorrect form, but rather a checking of syntactic feature specifications. Mismatches in the form of the presence of redundant features evoke a different response from mismatches in the form of an absence of required syntactic features.

Note, however, that for the present experiment this interpretation relies mainly on the learner data. For the native German control subjects, the strong declension violations could have a valid continuation as a plural (“mit kleinen Fenstern” [with small windows]), so that despite the fact that no plural forms were presented in our paradigm, their knowledge of this possibility continuation may have led them to interpret the phrase as grammatical at the adjective.

Whereas behavioral discrimination of the gender contrast improved with training, no ERP gender violation effect was observed in the learners as opposed to German native speakers, who showed a P600 response. Of course the absence of evidence cannot necessarily be taken as support for the lack of a neurophysiological violation response. It could be the case that an effect was present, but was present in brain regions which did not offer favorable recording sensitivity. Previous EEG studies examining similar gender contrasts have shown mixed results. Sabourin (2003) found some evidence for a P600 effect to Dutch grammatical gender violations in proficient German (L1)–Dutch (L2) bilinguals who, as in the present study, were able to discriminate between grammatical and ungrammatical phrases (Sabourin, Stowe, & de Haan, 2006). For definite NP gender violations (e.g., “het/\*de kleine kind. . .”; [the<sub>neuter/\*common</sub> small child]), a P600 effect was observed, whereas for indefinite NP gender violations (e.g., “. . . een gekke/\*gek manier”; [a funny<sub>common/\*neuter</sub> manner]), no P600 was seen. A P600 in English learners of Spanish has been observed (Tokowicz & MacWhinney, 2005), possibly due to the greater regularity of gender marking. It is possible that grammatical gender is more difficult to acquire than other grammatical distinctions, even when there is a great deal of overlap between the source (L1) and target (L2) languages, or alternatively, that the relationship between behavioral discrimination and the L2 electrophysiological response is more variable for grammatical gender. One important factor is that for grammatical gender, a number of associations pairing each noun with its appropriate gender category must be learned, whereas with the other grammatical distinctions, grammatical rules can be inferred (also see discussion in Williams & Lovatt, 2005).

Past research on error evaluation has shown two prominent ERP components observable when participants make errors in a wide variety of performance tasks (Falkenstein, Hohnsbein, Hoormann, & Blanke, 1991), the error-related negativity (ERN, or  $N_E$ ), and the error positivity ( $P_E$ ). The  $N_E$  occurs approximately 0–120 msec after an error response, and it has a frontal and central scalp distribution. The  $P_E$  occurs between 200 and 400 msec after the error response, with a broad scalp distribution. These ERP components are related to how participants evaluate the outcomes of their responses, and thus, may be related to how participants learn from feedback. To our knowledge, the error-related components have not yet been related to language learning, however.

Here we observed a greater magnitude response for errors ( $N_E$ ) compared to correct trials during training, and in the posttest. The medial frontal scalp distribution of the negative  $N_E$  potential is consistent with either an error-related negativity or a feedback-related negativity. In the training trials, it is likely that this response reflects both an error- and feedback-related negative potential, because the feedback signal followed shortly after the response. In the case of the posttest trials however,

there was no feedback signal, and the response is likely to be an error-related negative potential alone. The positive error component is also likely to be a reflection of an error-processing mechanism, as it appeared in concert with the  $N_E$ .

In addition to the  $N_E$  response, a conditional response, termed the  $N_{C|E}$ , was observed for correct trials such that the ERP response to correct trials was more negative following incorrect responses on the previous trial as compared to correct responses on the previous trial. Unlike the  $N_E$  effect, however, this  $N_{C|E}$  effect was not correlated with changes in the behavioral discrimination measures. Rather than learning, the  $N_{C|E}$  might reflect the response kinetics of the  $N_E$  response itself, indicating the manner in which the  $N_E$  changes from trial to trial, analogous to a priming effect. This may be related to the change of the response over pairs of trials, and thus may not be directly coupled to the final outcome of learning. Note also that Hajcak, Nieuwenhuis, Ridderinkhof, and Simons (2005) did not find an  $N_{C|E}$  effect in their study of error-response measures. More research on the  $N_{C|E}$  will be necessary to draw any firm conclusions.

The work reviewed in the Introduction characterized the responses of language learners to grammatical violations at relatively stable points during second-language acquisition, and also some of the factors that determine whether learners reach those points. This previous work mainly characterizes how levels of linguistic knowledge (or past experience) are related to violation effects. The  $N_E$  effects observed in the present experiment could also be interpreted as a response related to the knowledge level of the learners, as they acquired knowledge of German morphosyntax. In this account, learners who acquire more German grammatical knowledge would be more likely to become aware of whether their responses are correct or incorrect, and would therefore be more likely to show an  $N_E$  effect. This could occur whether or not the  $N_E$  effect reflects part of the grammatical learning mechanism itself. Some of the results of the regression analyses, however, make this interpretation of the  $N_E$  effect unlikely. First, the  $N_E$  was a significant predictor of performance in the training phase but not the posttest phase, where discrimination ability, and thus, grammatical knowledge, was highest. Given that in the posttest no feedback was provided, this pattern suggests, furthermore, that the response to the external feedback rather than participants’ evaluation of their own responses underlies the predictive value of the  $N_E$  during training. Second, the  $N_E$  response, but not the P600 response, during training predicted the improvement in performance from the pretest to the posttest phase. This pattern shows that the correlation with knowledge level that can be plausibly assumed for the P600 response was not sufficient for predicting performance in the subsequent posttest phase, despite the fact that performances during training and posttest were themselves highly correlated.

The  $N_E$  response might therefore be interpreted as being related to the acquisition of knowledge rather than the use of knowledge as such. Learning from feedback involves the evaluation of errors, or the evaluation of the difference between desired and actual outcomes. Holroyd and Coles (2002) proposed a reinforcement learning account of the  $N_E$  effect in which a negative reinforcement learning signal is propagated from the mesencephalic dopamine system to the anterior cingulate (ACC), which uses the signal to adapt performance on choice response time tasks. In their account, the  $N_E$  reflects the operation of a general error monitoring system, occurring in cases where executive control is involved in evaluating choices between alternative responses. The error monitoring system is involved in the detection of errors in responses, and in the use of error information to improve accuracy in responses. This error monitoring system might be involved in grammatical learning when the paradigm involves a choice response task, such as the classification of linguistic strings as acceptable or not acceptable according to some grammatical rule. By hypothesis, ACC–dopaminergic system would become linked to a network of language-related cortical regions that would represent the relevant features and rules for classification of word strings. These regions would form a network of areas that apply this information during parsing and comprehension of the linguistic strings presented to the learner, and in turn, evaluate whether the presented string is consistent or not consistent with the grammatical rules that the learner is hypothesizing. In a choice response task, the basis for a response would be the learner's representation of a grammatical rule, or the learner's representation of grammatical features that distinguish acceptable and unacceptable classifications.

The present results suggest that (rapidly) changing P600 and  $N_E$  responses can be observed to grammatical violations during the course of grammatical learning in adult language learners. In addition, a component previously associated with error-driven learning was observed to correlate with behavioral discrimination improvement. However, there were differences in the ERP responses between the different violation types, while behavioral discrimination was remarkably similar. Although only the declension violations led to a violation effect in the learners, the behavioral discrimination measures showed improvement for all violation types. Finally, the magnitude of the  $N_E$  effect was correlated with behavioral discrimination, and the relationship was stronger during the training and posttest phases, as would be expected if learners were acquiring discrimination ability.

This pattern of results suggests that the mechanisms that are responsible for feedback-driven learning should be modeled as a compartment of a system separate from, but coupled to, the mechanisms that detect grammatical violations. The system responsible for violation detection must, in some way, be informed and modulated by

feedback (e.g., Müller, Möller, Rodriguez-Fornells, & Münte, 2005), because the feedback negativity during training predicted performance improvement. Nonetheless, our results suggest that the two components are not coupled directly on each trial because the magnitude of the grammatical violation response during training did not predict performance improvement. They might still be loosely coupled, however, if detection ability for a repeated pattern of violations increases over time, based on the information provided by feedback. Stronger feedback responses might drive an adjustment of the violation detection system more strongly, and this effect would be reflected in stronger violation responses on later trials.

### Acknowledgments

The authors would like to thank the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO, Netherlands Organisation for Scientific Research) for their support of this research. Daniel von Rhein and Esther Meeuwissen assisted in the collection of the EEG data.

Reprint requests should be sent to Doug Davidson, F. C. Donders Centre for Cognitive Neuroimaging, P. O. Box 9101, 6500 HB Nijmegen, The Netherlands, or via e-mail: doug.davidson@fcdonders.ru.nl.

### REFERENCES

- Bierwisch, M. (1967). Syntactic features in morphology: General problems of so-called pronominal inflection in German. In *To honor Roman Jakobson* (pp. 239–270). The Hague: Mouton.
- Blevins, J. P. (1995). Syncretism and paradigmatic opposition. *Linguistics and Philosophy*, 18, 113–152.
- Cahill, L., & Gazdar, G. (1997). The inflectional phonology of German adjectives, determiners, and pronouns. *Linguistics*, 35, 211–245.
- Clahsen, H., Sonnenstuhl, I., Hadler, M., & Eisenbeiss, S. (2001). Morphological paradigms in language processing and language disorders. *Transactions of the Philological Society*, 99, 247–277.
- DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, 52, 43–94.
- Falkenstein, M., Hohnsbein, J., Hoormann, J., & Blanke, L. (1991). Effects of crossmodal divided attention on late ERP components 2. Error processing in choice reaction tasks. *Electroencephalography and Clinical Neurophysiology*, 78, 447–455.
- Ferree, T., Luu, P., Russell, G., & Tucker, D. (2001). Scalp electrode impedance, infection risk and EEG data quality. *Clinical Neurophysiology*, 112, 536–544.
- Friederici, A. D., Steinhauer, K., & Pfeifer, E. (2002). Brain signatures of artificial language processing: Evidence challenging the critical period hypothesis. *Proceedings of the National Academy of Sciences, U.S.A.*, 99, 529–534.
- Goethe-Institut. (2005). *Placement test German*. Retrieved May 3, 2005, from [www.goethe.de](http://www.goethe.de).
- Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP-measure of syntactic processing. *Language and Cognitive Processes*, 8, 439–483.

- Hahne, A. (2001). What's different in second-language processing? Evidence from event-related brain potentials. *Journal of Psycholinguistic Research*, 30, 251–266.
- Hahne, A., & Friederici, A. D. (2001). Processing a second language: Late learners' comprehension mechanisms as revealed by event-related brain potentials. *Bilingualism: Language and Cognition*, 4, 123–141.
- Hajcak, G., Nieuwenhuis, S., Ridderinkhof, K. R., & Simons, R. F. (2005). Error-preceding brain activity: Robustness, temporal dynamics, and boundary conditions. *Biological Psychology*, 70, 67–78.
- Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, 109, 679–709.
- Johnson, J. S., & 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. *Cognitive Psychology*, 20, 60–99.
- Johnson, J. S., & Newport, E. L. (1991). Critical period effects on universal properties of language: The status of subadjacency in the acquisition of a second language. *Cognition*, 39, 215–258.
- Kutas, M., Van Petten, C., & Kluender, R. (2006). Psycholinguistics electrified II: 1995–2005. In M. Traxler & M. Gernsbacher (Eds.), *Handbook of psycholinguistics* (pp. 659–724). Amsterdam: Academic Press.
- Lenneberg, E. H. (1967). *Biological foundations of language*. New York: Wiley.
- Maris, E. (2004). Randomization tests for ERP-topographies and whole spatiotemporal data matrices. *Psychophysiology*, 41, 142–151.
- Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164, 177–190.
- Mueller, J. L., Hahne, A., Fujii, Y., & Friederici, A. D. (2005). Native and nonnative speakers' processing of a miniature version of Japanese as revealed by ERPs. *Journal of Cognitive Neuroscience*, 17, 1229–1244.
- Mueller, J. L., Hirotsani, M., & Friederici, A. (2007). ERP evidence for different strategies in the processing of case markers in native speakers and non-native learners. *BMC Neuroscience*, 8, 1471–2202.
- Müller, S. V., Möller, J., Rodriguez-Fornells, A., & Münte, T. F. (2005). Brain potentials related to self-generated and external information used for performance monitoring. *Clinical Neurophysiology*, 116, 63–74.
- Ojima, S. A., Nakata, H. A., & Kakigi, R. (2005). An ERP study of second language learning after childhood: Effects of proficiency. *Journal of Cognitive Neuroscience*, 17, 1212–1228.
- Osterhout, L., & Holcomb, P. J. (1995). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31, 785–806.
- Osterhout, L., McLaughlin, J., Kim, A., Greenwald, R., & Inoue, K. (2005). Sentences in the brain: Real-time reflections of sentence comprehension and language learning. In M. Carreiras & J. C. Clifton (Eds.), *The on-line study of sentence comprehension: Eyetracking, ERP, and beyond* (pp. 271–308). London: Psychology Press.
- Osterhout, L., McLaughlin, J., Pitkanen, I., Frenck-Mestre, C., & Molinaro, N. (2006). Novice learners, longitudinal designs, and event-related potentials: A paradigm for exploring the neurocognition of second-language processing. *Language Learning*, 56, 199–203.
- Osterhout, L., Poliakov, A., Inoue, K., McLaughlin, J., Valentine, G., Pitkanen, I., et al. (in press). Second-language learning and changes in the brain. *Journal of Neurolinguistics*. doi: 10.1016/j.jneuroling.2008.01.001.
- Penke, M., Janssen, U., & Eisenbeiss, S. (2004). Psycholinguistic evidence for the underspecification of morphosyntactic features. *Brain and Language*, 90, 423–433.
- Pettersson, K. M., Forkstam, C., & Ingvar, M. (2004). Artificial syntactic violations activate Broca's region. *Cognitive Science*, 28, 383–407.
- Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer-Verlag.
- Raven, J. (1962). *Advanced progressive matrices, set II*. London: H. K. Lewis.
- Rossi, S., Gugler, M. F., Friederici, A., & Hahne, A. (2006). The impact of proficiency on syntactic second-language processing of German and Italian: Evidence from event-related potentials. *Journal of Cognitive Neuroscience*, 18, 2030–2048.
- Sabourin, L., Stowe, L. A., & de Haan, G. J. (2006). Transfer effects in learning a second language grammatical gender system. *Second Language Research*, 22, 1–19.
- Sabourin, L. L. (2003). *Grammatical gender and second language processing: An ERP study*. Unpublished PhD, Rijksuniversiteit Groningen, Groningen.
- Sakai, K. L. (2005). Language acquisition and brain development. *Science*, 310, 815–819.
- Schlenker, P. (1999). La flexion de l'adjectif en allemand: La morphologie de haut en bas. *Recherches linguistiques de Vincennes*, 28, 115–132.
- Tokowicz, N., & MacWhinney, B. (2005). Implicit and explicit measures of sensitivity to violations in second language grammar—An event-related potential investigation. *Studies in Second Language Acquisition*, 27, 173–204.
- Weber-Fox, C., & Neville, H. J. (2001). Sensitive periods differentiate processing of open- and closed-class words: An ERP study of bilinguals. *Journal of Speech, Language, and Hearing Research*, 44, 1338–1353.
- Williams, J. N., & Lovatt, P. (2005). Phonological memory and rule learning. *Language Learning*, 55, 177–233.
- Wunderlich, D. (1997). Der unterspezifizierte Artikel. In M. S. u. K.-H. R. Christa Dürscheid (Ed.), *Sprache im Fokus* (pp. 47–55). Tübingen: Niemeyer.
- Zwicky, A. M. (1986). German adjective agreement in GPSG. *Linguistics*, 24, 957–990.