LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

EMPIRICAL STUDY

The Role of Conflicting Representations and Uncertainty in Internal Error Detection During L2 Learning

Sybrine Bultena,[a,b] Claudia Danielmeier,[c] Harold Bekkering,[a,b] and Kristin Lemhöfer[a,b]

[a]Radboud University Nijmegen, [b]Donders Institute, and [c]University of Nottingham

Abstract: Internal error monitoring as reflected by the error-related negativity (ERN) component can give insight into the process of learning a second language (L2). Yet, early stages of learning are characterized by high levels of uncertainty, which obscures the process of error detection. We examine how uncertainty about L2 syntactic representations, induced by different levels of language conflict, is reflected in ERN patterns during learning. German learners of Dutch performed a feedback-guided gender decision task in their L2 and provided subjective certainty ratings for their responses. Initially, high-conflict items yielded more uncertainty and ERN modulations were reversed (i.e., correct responses elicited larger amplitudes than errors). Two rounds of feedback resulted in an increase of accuracy, reduced uncertainty, and normalization of the ERN effect, signaling effective error monitoring. These outcomes demonstrate how subjective intuitions about response accuracy affect performance monitoring during L2 learning.

Keywords feedback-guided learning; performance monitoring; grammatical gender; language conflict; uncertainty; cognates

We would like to thank Iris Verpaalen for her help in the data collection. CD was supported by the British Academy grant SG152607

Correspondence concerning this article should be addressed to Sybrine Bultena, Centre for Language Studies, Radboud University Nijmegen, Netherlands. Email: s.bultena@let.ru.nl

DOI: 10.1111/lang.12401

Introduction

According to popular belief, we learn from our mistakes, which implies that the brain monitors performance, and such assumption is generally also made in models of second language (L2) learning (e.g., the Noticing Hypothesis; Schmidt, 1990). Yet, within the field of L2 learning, little neuroscientific data is available to support this notion, even though a relevant domain-general event-related potential (ERP) component has long been recognized as a valuable tool to study performance monitoring. Monitoring of decisional performance modulates the error-related negativity (ERN), a sharp frontal negative deflection peaking within 100 milliseconds of committing an error. This component is commonly observed for domain-general action execution errors (for a review see Gehring, Liu, Orr, & Carp, 2011), as well as language selection errors (Zheng, Roelofs, Farquhar, & Lemhöfer, 2018), and is considered to index internal error detection.

The amplitude of the ERN has been shown to depend on the certainty with which error detection takes place (Pailing & Segalowitz, 2004). The process of L2 learning is typically characterized by a large degree of uncertainty, for example, regarding the syntactic correctness of an utterance, be it one's own or that of someone else (Johnson, Shenkman, Newport, & Medin, 1996). Learners first need to acquire knowledge or stabilize correct representations before being able to evaluate their own response accuracy. Before such knowledge is in place, learners are thus unlikely to optimally engage in internal error detection evidenced by the ERN. An absence of the ERN effect has, for example, been observed in nonlinguistic situations when rule learning was impossible due to invalid feedback (Eppinger, Kray, Mock, & Mecklinger, 2008) or when bilinguals could not perceive the difference between a correct and erroneous response in their L2 (Sebastian-Gallés, Rodríguez-Fornells, de Diego-Balaguer, & Díaz, 2006). We hypothesize that successful learning could be seen as a reduction in uncertainty and should therefore be accompanied by a progression toward the occurrence of an ERN. This study will focus on the issue of L2 grammar learning and investigate which behavioral and neural changes accompany the learning of a difficult grammatical feature, here L2 gender. The difficulty of learning this feature for our population of interest, German learners of Dutch, is mainly due to cross-language gender incompatibility for some nouns, especially words that are cognates between the two languages (see also Lemhöfer, Schriefers, & Hanique, 2010; Lemhöfer, Spalek, & Schriefers, 2008).

Background Literature

The ERN is typically observed in speeded choice reaction time (RT) tasks where errors are due to premature responding on the level of perceptual awareness or action execution, such as in Flanker tasks. The difference between the large response-locked negativity for errors (ERN) and the smaller negativity for correct responses (CRN) is known as the *ERN effect* and is thought to reflect internal error detection (Gehring, Goss, Coles, Meyer, & Donchin, 1993) or prediction error detection (Alexander & Brown, 2011; Holroyd & Coles, 2002). The size of the ERN effect can be modulated. For example, it is larger for more easily detected errors (Falkenstein, Hoormann, Christ, & Hohnsbein, 2000), in case of greater response conflict (Danielmeier, Wessel, Steinhauser, & Ullsperger, 2009), when more attention is devoted to errors (Maier & Steinhauser, 2016), and for perceived errors as compared to unperceived errors (Wessel, Danielmeier, & Ullsperger, 2011). Of particular interest to learning situations, these findings imply that variation in the size of the ERN goes hand-in-hand with changes in subjective certainty about the accuracy of the response (Scheffers & Coles, 2000). For instance, Pailing and Segalowitz (2004) observed an effect of uncertainty on the ERN effect, such that uncertainty about performance in a perceptual task was reflected by a larger CRN component, resulting in similar-sized negativities for both errors and correct responses (i.e., a cancellation of the ERN effect). In the same vein, Scheffers and Coles (2000) asked participants to rate their confidence regarding a just-given response and showed that ERP amplitude increased with participants' confidence of having made an error. Consistent with this view, work by Boldt and Yeung (2015) points to a shared mechanism for error detection and confidence judgments. After every response in a visual perception task, the authors asked participants to rate the certainty of their response on a 6-point scale, ranging from *certainly wrong* to *certainly correct*. Both the amplitude of the ERN and the subsequent error positivity (Pe; a component associated with error awareness) correlated with subjective certainty, such that the ERN was most negative for items rated as having elicited a certainly wrong response and least negative for items rated as having induced a certainly correct response. These findings indicate that error-related ERP components are subjective, reflecting a certainty-dependent continuum, rather than a binary error detection mechanism. Although the studies discussed above concern decisions based on sensory information that do not explicitly involve learning, they suggest that high levels of uncertainty, as present in the early stages of L2 learning, may be characterized by reduced ERN effects.

Beginning L2 learners are often faced with uncertainty due to a lack of knowledge and unstable representations, as indicated by inconsistent behavioral responses on grammaticality judgments in L2 learners of English (e.g., Johnson et al., 1996). Although studies on neurocognitive performance monitoring in the domain of L2 learning are scarce, the few available studies suggest that uncertainty plays a role. A feedback-based L2 training study on the acquisition of a complex and difficult-to-learn morpho-syntactic feature by Davidson and Indefrey (2011) looked at response-locked ERP components. Prior to training, behavioral accuracy was low and response-locked negativities for errors and correct responses did not differ. In the course of training, during which participants received feedback, behavioral performance improved and simultaneously a difference between the ERN and CRN waveforms emerged. In comparison to the classic ERN effect, however, the observed effect was small: The similar-sized ERN and CRN components resembled the pattern observed for uncertainty (Pailing & Segalowitz, 2004) and presumably reflect the difficulty to detect errors in the case of a newly learned feature.

Apart from the usual uncertainty involved in learning something new, L2 learners sometimes face an additional challenge. It is commonly accepted that L1 influences the processing and acquisition of a L2, especially so in the domain of syntax (Caffarra, Molinaro, Davidson, & Carreiras, 2015). Coactivation of competing L1 representations may thus further decrease confidence in performance or could lead to false intuitions about correct L2 representations, when these are incongruent between the L1 and the L2. A case in point is that of cross-language differences in grammatical gender of orthographically similar translation equivalents (i.e., cognates). German and Dutch both use gendered articles and share many cognates, but the gender for these cognates is not always congruent across languages, resulting in persistent gender errors when German learners of Dutch use their L2. When investigating the effects of cognate status and gender congruence for German learners of Dutch, Lemhöfer et al. (2010) observed that gender incongruent cognates, in particular, yield many errors regarding gender assignment, both before and after training, pointing to robust L1 transfer for this category. Lemhöfer, Schriefers, and Indefrey (2014) furthermore showed that when presented with nouns preceded by either correct or incorrect gendered articles in a sentence context, these learners' ERPs reflected the detection of a syntactic violation only when determiners violated participants' intuitions about a noun's grammatical gender, even though it was objectively correct in their L2. Subjective accuracy may thus affect ERP components more than objective accuracy. In this respect, it is interesting to note that response-locked components in a nonlinguistic action execution task similarly lead to an ERN for objectively correct responses when these were misclassified as errors (Scheffers & Coles, 2000).

The persistent errors for gender incongruent cognates formed the starting point of a previous study by our group (Bultena, Danielmeier, Bekkering, & Lemhöfer, 2017). By means of a feedback-guided gender assignment task, we examined whether advanced German learners of Dutch show signs of error detection on gender incongruent cognates in Dutch (Dutch *het*$_{neuter}$ *strand*/*German der*$_{masculine}$ *Strand*) as reflected by the ERN effect. The task involved three consecutive rounds, with participants receiving corrective feedback after each trial, enabling them to learn over the course of the experiment. The critical items were cognates of Dutch and German with incompatible genders across languages (high language conflict). In the first round, learners made many errors on target trials, and their ERPs showed no clear difference between ERN and CRN components. Following feedback, behavioral results indicated a rapid improvement in accuracy, accompanied by a small but significant ERN effect in the final round. Interestingly, a closer inspection of the results in the first round suggested that the ERN effect was reversed, with marginally higher negativities for correct than erroneous responses. This result was reminiscent of that obtained by Lemhöfer et al. (2014), and suggested that correct responses violating L1 intuitions were in fact considered as "errors" by L2 learners. Yet, the stimuli used predominantly included gender incongruent cognates that could be considered high-conflict items, and only a limited number of filler items that involved low levels of language conflict. The total number of errors on target and filler items therefore included very few errors on low conflict items, which prevented us from properly studying the effect of language conflict. Furthermore, overall certainty ratings obtained in a posttest were positively correlated with the size of individual ERN effects, suggesting that more certainty led to better error monitoring. In the current study, we aimed to look more closely at the effect of language conflict on the size of ERN and CRN components during learning, and how subjective certainty about response accuracy develops in the course of learning.

The Current Study

Starting from the premise that successful learning should lead to a reduction in uncertainty, that is, a gradual increase in certainty, we investigated how subjective certainty induced by cases of high and low language conflict relates to differences between correct and incorrect responses during learning.

Table 1 Predictions for the experimental design

	Before feedback (Round 1)	After feedback (Rounds 2 + 3)
Low conflict	ERN = CRN	ERN > CRN
High conflict	ERN < CRN	ERN > CRN

We asked a group of German L2 learners, similar to those tested in Bultena et al. (2017) to select the correct determiner for Dutch nouns, and, this time, we obtained a direct metacognitive measure of response certainty. In a departure from the previous experiment, learners were asked to give certainty ratings for their responses before receiving corrective feedback. Additionally, in order to manipulate levels of uncertainty throughout the experiment, we varied the degree of L1–L2 conflict by including nouns that are gender-compatible and nouns that are gender-incompatible between the two languages, and that are either cognates (e.g., *auto*/*Auto*; "car") or noncognates (e.g., *fiets*/*Fahrrad*; "bicycle"), allowing for a comparison between high (gender incongruent cognates) and low (gender congruent cognates, gender congruent noncognates, and gender incongruent noncognates) language-conflict items.

Based on available evidence from perceptual decision tasks that do not involve learning, we hypothesized that a reduction in uncertainty as a result of learning, as measured by ratings, would be accompanied by an increase in the ERN effect. More specifically, we expected that the learning process would show different stages reflected by distinct ERN modulation patterns, depending on the degree of cross-language conflict. Prior to receiving feedback, errors and correct responses for low-conflict items should initially yield similar-sized response-locked negativities in line with subjective certainty accounts, while for items that present a high L1–L2 conflict, the ERN effect may be reversed. That is, the negativity associated with correct responses (*de auto*) could be larger than for errors (**het auto*), because what is objectively correct is subjectively perceived as incorrect according to intuitions based on L1 knowledge, and vice versa. After having received feedback, when participants develop more stable representations about correct and incorrect responses and thus become more certain, response-locked negativities should gradually differentiate between ERN and CRN components for both high and low conflict items, reflecting improvements in internal error monitoring. Expected effects have been summarized in Table 1.

Method

Participants

A total of 30 students at Radboud University responded to an online recruitment announcement made through a participant panel and took part in the experiment after signing the informed consent form. Two participants had to be excluded due either to technical problems or health issues during recording. This left data of 28 participants for analysis (four male, 24 female; $M_{\text{age}} =$ 22 years, $SD = 2$, range = 18–25 years), who had no history of neurological or psychiatric disease, had normal or corrected-to-normal vision, and were right-handed according to an abridged version of the Oldfield handedness questionnaire (Oldfield, 1971). All participants were native speakers of German who spoke Dutch as a second language, in addition to English and, in most cases, at least one other foreign language. Most of them had started to learn Dutch with the purpose of studying in the Netherlands, at least one year before taking part in the study, and a large majority of them lived in the Netherlands at the time of testing ($N = 23$). Participants filled out a questionnaire in which they quantified their motivation to learn Dutch in general (general learning motivation, perfectionism, perseverance, confidence) and their motivation to learn during the experiment (task motivation). This questionnaire was based on the Attitude/Motivation Test Battery (Gardner, 1985) complemented by questions on task performance inspired by Luu, Collins, and Tucker (2000; for the full list of questions see Appendix S1). Behavioral measures of L2 proficiency, use, and motivation to learn the language are summarized in Table 2. Participants received course credit or were paid (€10 per hour) for their participation.

Materials

A total of 132 Dutch nouns were used for the feedback-guided gender decision task. Cross-language noun similarity (cognate/noncognate) and gender congruence between Dutch and German (congruent/incongruent) were manipulated to create high- and low-conflict conditions; cognate status and gender congruence were not used as factors in the design. Cognates were defined as translation equivalents that scored low in terms of orthographic Levenshtein distance (number of character changes/average word length; Van Orden, 1987) between the German and Dutch forms ($M_{cognate} = .18$, $SD = .22$ vs. $M_{non\text{-}cognate} = .96$, $SD = .19$). German nouns with masculine (*der*) and feminine (*die*) gender were considered to be congruent with common (*de*) gender in Dutch. We selected 44 cross-language gender incongruent cognates, 22 gender congruent cognates, 44 gender congruent noncognates, and 22 gender incongruent noncognates (see Appendix S2 for a full list of the stimuli). Gender

Table 2 Means and standard deviations regarding L2 Dutch use and proficiency and scores reflecting motivation to learn the language ($N = 28$)

		Mean	*SD*	range
Years of experience learning Dutch		3.4	2.3	1–10
Dutch age of acquisition		19	1.8	14–23
LexTALE score (vocabulary size) in Dutch		69	11	50–90
Self-rated frequency	Speaking	6.3	0.9	4–7
	Listening	6.1	1.2	3–7
	Reading	5.3	1.7	1–7
Self-rated proficiency	Speaking	4.8	1.0	3–7
	Listening	5.6	0.9	3–7
	Writing	4.4	1.2	2–6
	Reading	5.7	0.8	3–7
	Overall	4.9	0.9	3–6
Self-rated motivation	General learning motivation	17	1.7	13–20
	Perfectionism	16	2.8	8–20
	Perseverance	17	2.1	13–20
	Confidence	14	3.0	9–19
	Task motivation	13	2.5	9–19

Note. The LexTALE score represents Dutch vocabulary size based on an averaged percentage correct over word and non-word items on a lexical decision task. Frequency and proficiency ratings were based on a 7-point scale ranging from 1 (low) to 7 (high). The overall score reflects participants' average estimation. Motivation scores are summated scores across four questions per dimension based on 5-point scales (max 20 points per dimension). An overview of the motivation questions can be found in Supporting Information I.

incongruent cognates were considered high conflict, while items of the other three categories were classified as low conflict. The low-conflict condition was a combined set of gender congruent cognates, and congruent as well as incongruent noncognates by necessity, because previous studies have shown that these learners make relatively few errors on these three word categories and that the numbers of errors made are comparable across items (Bultena et al., 2017; Lemhöfer et al., 2010). A minimum number of six to eight error trials being required to compute a grand average ERN waveform (Olvet & Hajcak, 2009), we decided to include a larger number of low- (88) than high- (44) conflict trials.

All nouns were used in their singular nondiminutive form. Occurrences of *de* (common gender; a combination of masculine and feminine gender) and *het* (neuter gender) words were equiprobable across the four word categories (apart from a minor difference in the case of incongruent noncognates due to limited availability of neuter items; see Appendix S2). Analyzes were always performed after collapsing across *de* and *het* items. Independent samples *t* tests showed that high and low conflict conditions were matched on word length in letters (high conflict: $M = 5.6$, $SD = 1.4$ vs. low conflict: $M = 5.4$, $SD = 1.5$, $p = .709$) and SUBTLEX word form log frequency (high conflict $M = 2.8$, $SD = 0.6$ vs. low conflict $M = 3.0$, $SD = 0.6$, $p = .262$) in Dutch (Brysbaert & New, 2009). To ensure correct identification of each noun, a color picture of an object against a white background was selected from free-access internet databases for each stimulus. Pictures were resized to fit a template of 180 × 180 pixels (96 dpi, screen resolution 1,280 × 1,024). An additional set of 18 words and matching pictures was used for practice, including items from all word categories.

Procedure

Participants were told that they were taking part in a learning study. In a feedback-guided gender decision task, they were asked to decide on the correct gendered article (*de* or *het*) for a Dutch noun by means of a button press on a response box, and rate the certainty of the correctness of their response, before they were presented with feedback on each trial. All 132 nouns were presented in three consecutive rounds, allowing participants to learn the correct representations over the course of the experiment. Item presentation within round was pseudorandomized using Mix (Van Casteren & Davis, 2006), based on Dutch gender, gender congruence with German, and cognate status, with a maximum of four items from the same category in a row. The experiment started with 18 practice trials, which were not presented in the subsequent three rounds.

Upon arrival, participants signed informed consent and filled out a language background questionnaire (see Table 2) and the Oldfield handedness questionnaire, after which they were prepared for the electrophysiological experiment. Subsequently, participants were asked to name all the pictures used as experimental and practice stimuli to check for noun familiarity, using bare nouns only. Pictures that could not be named were marked as unfamiliar and excluded from the analyses. Prior to the gender decision task, participants were verbally instructed to avoid movements and excessive blinking as much as possible.

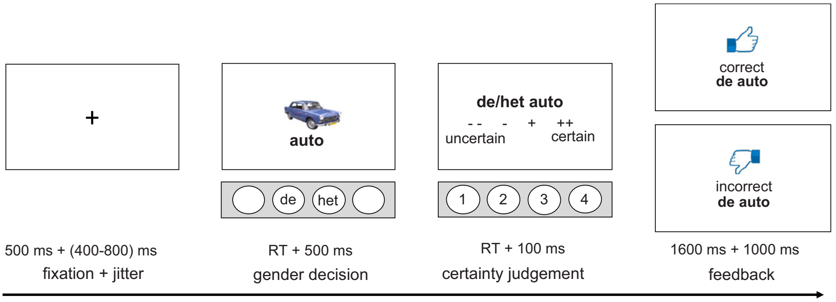

Figure 1 Graphical display of the trial sequence. Added times (+) are intervals in between screens, during which participants saw a blank screen (or the response screen in case of the gender response). On the feedback screen, the correct determiner noun combination was presented together with accuracy feedback for the participant's last gender response. Car picture taken from http://freeimage.com (credits: Michal Zacharzewski, SXC). [Color figure can be viewed at wileyonlinelibrary.com]

Every trial (see Figure 1 for a schematic representation of events) started with a fixation cross for 500 milliseconds followed by a jittered blank screen (a random interval between 400 and 800 milliseconds). Then, a picture (48 × 48 mm on screen, viewed from approximately 50 cm at a viewing angle of 4.5 degrees) and its accompanying noun (Arial 16 pts, black) printed underneath were displayed in the center of a white screen until 500 milliseconds after a response had been recorded. Subsequently, a rating screen was presented depicting the actual response (e.g., *het auto*) and a 4-point Likert scale ranging from uncertain (*onzeker*) to certain (*zeker*). All responses were recorded with an in-house designed four-button box. Participants were instructed to rest their left and right index fingers on the middle two buttons for a fast gender response and move their fingers back to this position after making a certainty response. Following the rating response, participants were presented with corrective feedback including information on response accuracy in the form of a thumbs up or down symbol and the word *goed* ("correct") or *fout* ("incorrect") as well as the correct determiner-noun combination, for 1,600 milliseconds. After this, participants saw a blank screen for 1,000 milliseconds, during which they were encouraged to blink gently. Although participants were encouraged to respond quickly, response accuracy was emphasized over response speed, and there was no time limit for either button press.

The practice trials and three experimental rounds of the gender decision task lasted approximately 40 minutes, including self-paced breaks in the middle and at the end of every round. After each round, participants received information about their accuracy in the preceding round as indicated by a percentage and were encouraged to try and improve this score in the subsequent round.

Following testing and a short hair washing break, participants completed a pen and paper posttest in which they were asked to fill in the correct determiner for all 132 listed nouns and tick one of four boxes to indicate their certainty for each response. Afterwards, they performed the Dutch version of the LexTALE task (Lemhöfer & Broersma, 2012), which measures vocabulary size as an indication of proficiency, and they filled out a digital version of the motivation questionnaire.

Electrophysiological Recording and Preprocessing

The electroencephalogram (EEG) was recorded using active electrodes from 60 scalp sites, arranged according to the extended international 10–20 system (ActiCAP, Brain Products, GmbH, Gilching, Germany), online referenced to the left mastoid (ground electrode placed at AF7). This number of electrodes is beneficial when using ICA decomposition to de-noise the data as sources of noise can be identified better with more electrodes. We measured the horizontal and vertical electro-oculogram (EOG) from the electrodes positioned at the outer canthi of the left and right eye, and above and below the right eye. Electrode impedance was kept below 10 kΩ. EEG signals were recorded continuously using two BrainAmp DC amplifiers in combination with BrainVision Recorder software (Brain Products, GmbH, Gilching, Germany), converted with a 16-bit resolution and sampled at 500 Hz. Recording filters were set to a low cut-off of 0.016 Hz and a high cutoff of 125 Hz. Triggers were sent out to the recording computer at stimulus onset, gender response onset, and feedback onset.

EEG data were preprocessed and analyzed using EEGlab (Delorme & Makeig, 2004) and MATLAB (Mathworks, Natick, MA, USA). For a few participants, bad channels caused by cable breakage (maximally 3) were removed from individual datasets before any preprocessing. EEG data were rereferenced offline to a common average reference based on all electrodes except EOG channels, and then subsequently high-pass filtered with a 0.1-Hz cutoff and low-pass filtered with a 30-Hz cutoff to correct slow drifts and reduce high-frequency noise, respectively. Subsequently, four-second-long stimulus-locked epochs that included both the responses and feedback presentation were extracted from the continuous data to reduce the file size for ICA decomposition. All items that were unfamiliar to participants ($M = 12$, $SD = 9$) were removed from individual datasets at this point. Baseline correction was performed relative to the 200 milliseconds preceding stimulus presentation. Prior to ICA transformation, bad epochs were rejected based on visual inspection using the joint probability tool in EEGlab (5 *SD*s), which led to an average loss of six trials ($SD = 3$) per participant. Independent component analysis (Infomax algorithm) was performed on the segmented data of each participant. A total of 60 components (or fewer for datasets with bad channels) were computed, and screened for eye, muscle, and heartbeat artefacts based on topography, power spectrum, and trial activity as shown for each component in EEGlab. An average of five components ($SD = 2$) was removed before ICA back-transformation. The artefact corrected datasets were re-epoched to create response-locked and feedback-locked segments. Based on these segments, individual averages sorted by round, accuracy, and conflict were created per participant, which formed the basis of subsequently created grand averages. A minimum of six trials per condition were used as a criterion to include a participant in the analyses (Olvet & Hajcak, 2009; cf. Fischer, Klein, & Ullsperger, 2017).

Data Analysis

Dependent variables in the behavioral data consisted of error rates, response times (RTs), and certainty ratings. The EEG data were analyzed as response-locked waveforms, but baseline corrections were performed in the 200 milliseconds prior to stimulus onset. To analyze the response-locked ERN and CRN, trough-to-peak amplitudes were computed at electrode FCz, because initial comparisons for Fz, FCz, and Cz had shown that overall effects were maximal at FCz (cf. Bultena et al., 2017; waveforms for other electrodes are shown in Supplementary Information VII). The peak was defined as the maximal negative amplitude within 100 milliseconds after response onset, and the trough as the maximal positive amplitude between 100 milliseconds before response onset and the negative peak, in agreement with previous studies (e.g., Danielmeier et al., 2009; Endrass, Klawohn, Schuster, & Kathmann, 2008; Wessel & Ullsperger, 2011). The first response-locked component was followed by a second negative peak, similarly quantified as a trough-to-peak difference at FCz (see Bultena et al., 2017). Its amplitude difference was measured between the maximal negative peak in the time window between 200 and 300 milliseconds postresponse onset and the maximal positive trough in the 100 milliseconds preceding the negative peak. Because this way of quantifying the second peak is strongly dependent on the effect in the first peak, the later effect was additionally quantified as the mean amplitude between 150 and 400 milliseconds, based on visual inspection of the difference wave.

Behavioral and ERP responses were analyzed for effects of three factors: response accuracy (correct/error), language conflict (high/low), and round. The number of rounds differed for each dependent variable. For response times, three rounds were included, while for error rate analyses and certainty ratings, the posttest on paper was regarded as an additional (fourth) round. For the ERP analyses, the three rounds of the main experiment were divided post hoc into *before feedback* (Round 1) and *after feedback* (Rounds 2 and 3) to ensure that a minimum number of six error trials per cell was available (per accuracy and condition in each participant), as recommended by Olvet and Hajcak (2009). Before ERPs were averaged over the last two rounds, we verified that waveform patterns looked similar between rounds. Data of three participants who made fewer than six errors in the last two rounds had to be discarded. Data of two more participants were discarded because ICA correction failed.

All dependent variables were analyzed using repeated measures ANOVAs. Interactions reaching significance were followed up by planned comparisons involving paired samples *t* tests or planned contrasts, depending on the type of comparison. When both two- and three-way interactions were present, only the latter are reported. In all cases, alpha was set at .05 and Greenhouse-Geisser corrections are reported when the assumption of sphericity was violated.

In addition to response-locked analyses, we also looked at feedback-locked components to examine how learners responded to feedback. These findings are reported in Supporting Information VII.

Results

Behavioral Performance

Analyses were performed on familiar items only. Nouns marked as unfamiliar during familiarization (9.0% in total) were excluded from analyses. Noun familiarity was high for both high- ($M = 89\%$, $SD = 8$, range 72%–100%) and low-conflict items ($M = 94\%$, $SD = 5$, range 82%–100%). Overall, participants made a total of 21% errors on familiar items over the three rounds of the gender decision task.

A two-way ANOVA on error rates with language conflict and round (four levels) as factors showed significant main effects of conflict, $F(1, 24) = 162.33$, $p < .001$, $\eta_p^2 = .871$, and round, $F(2.02, 48.36) = 141.77, p < .001$, $\eta_p^2 = .855$, as well as an interaction between these factors, $F(3,72) = 69.73$, $p < .001$, $\eta_p^2 = .744$. High-conflict items yielded more errors ($M = 35\%$, $SE = 2$) than low-conflict items ($M = 11\%$, $SE = 1$). Follow-up planned contrasts for the low-conflict condition indicated a significant decrease in errors between every round and the next (ps $< .004$; $M_1 = 17\%$, $SE = 1$; $M_2 = 13\%$, $SE = 1$; $M_3 = 8\%$, $SE = 1$; $M_{\text{post}} = 6\%$, $SE = 1$), and an even stronger decrease for each round (ps $< .001$) for the high-conflict condition ($M_1 = 63\%$, $SE = 3$; $M_2 = 39\%$, $SE = 3$; $M_3 = 24\%$, $SE = 3$; $M_{\text{post}} = 15\%$, $SE = 2$), as can be seen in Figure 2 (top panel).

A three-way repeated measures ANOVA on RTs with accuracy, conflict, and round (three levels) as factors reveal significant effects of accuracy, $F(1, 23) = 36.20, p < .001,\ \eta_p^2 = .611$, conflict, $F(1, 23) = 5.35, p = .030,\ \eta_p^2 = .189$, and two-way interactions between accuracy and conflict, $F(1, 23) = 43.32, p < .001,\ \eta_p^2 = .653$, and between accuracy and round, $F(2, 46) = 16.89, p < .001,\ \eta_p^2 = .423$. Paired samples t tests were run to compare the RTs for errors and correct responses per round and conflict condition. For the low conflict conditions, these comparisons indicated faster response times for correct responses compared to errors across all three rounds (ps $< .001$), indicating that errors were not due to response speed. For the high conflict condition, however, a different pattern emerged. In round one, erroneous responses ($M = 1641$, $SE = 51$) were faster than correct responses ($M = 1{,}817$, $SE = 75$; $t(24) = -3.53, p = .002$), while round two showed no difference ($t < 1$) between errors ($M = 1{,}704$, $SE = 81$) and correct responses ($M = 1{,}691$, $SE = 72$), and round three indicated that error responses ($M = 1{,}801$, $SE = 90$) were slower than correct responses ($M = 1{,}484$, $SE = 79$; $t(24) = 4.83$, $p < .001$), similar to the low conflict condition (see Figure 2, middle panel).

Certainty ratings were analyzed to check whether language conflict affected participants' confidence about their performance. A three-way repeated measures ANOVA on certainty ratings with accuracy, language conflict, and round (four levels) yielded main effects of accuracy, $F(1, 22) = 99.89$, $p < .001$, $\eta_p^2 = .820$, language conflict, $F(1, 22) = 13.80$, $p = .001$, $\eta_p^2 = .386$, and round, $F(3, 66) = 6.02, p = .001,\ \eta_p^2 = .215$, in combination with a three-way interaction, $F(3, 66) = 5.17, p = .003,\ \eta_p^2 = .190$. Paired samples t tests revealed significantly higher ratings for correct responses as compared to errors (round 1: $M_{\text{c}} = 3.1, SE = .08, M_{\text{e}} = 2.5, SE = .08$; round 2: $M_{\text{c}} = 3.2, SE = .09$, $M_{\text{e}} = 2.6, SE = .12$; round 3: $M_{\text{c}} = 3.4, SE = .09, M_{\text{e}} = 2.6, SE = .12$; post-test: $M_{\text{c}} = 3.6, SE = .07, M_{\text{e}} = 2.4, SE = .18$) in all four rounds for the low-conflict items (ps $< .001$), whereas errors and correct responses in the high-conflict condition (Round 1: $M_{\text{c}} = 2.5$, $SE = .08$, $M_{\text{e}} = 2.6$, $SE = .08$; Round 2: $M_{\text{c}} = 2.9$, $SE = .11$, $M_{\text{e}} = 2.5$, $SE = .12$; Round 3: $M_{\text{c}} = 3.1$, $SE = .11$, $M_{\text{e}} = 2.5, SE = .13$; post-test: $M_{\text{c}} = 3.4, SE = .10, M_{\text{e}} = 2.6, SE = .13$) showed such a difference only after the first round (ps $< .001$). For high-conflict items

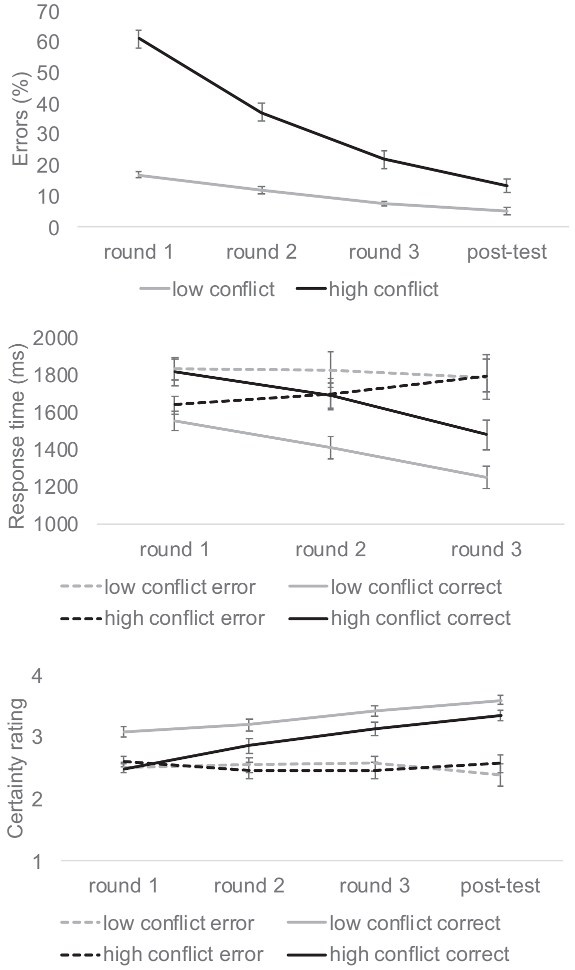

Figure 2 Behavioral data. Top panel shows error rates for high- and low-conflict conditions over rounds. The middle panel shows RTs for high- and low-conflict conditions by accuracy and round. The bottom panel shows certainty ratings for high- and low-conflict conditions by accuracy and round. Certainty ratings were given on a 4-point scale ranging from uncertain (1) to certain (4). Error bars in all graphs reflect standard errors.

in round 1, no difference was present between the certainty ratings for correct and error responses, $t(23) = 1.29, p = .211$. As can be seen in Figure 2 (bottom panel), correct responses on high-conflict items received lower certainty ratings than correct responses on low-conflict items.

In sum, the behavioral data indicated that, high conflict items yielded more errors than low conflict items, but learning rates significantly increased with every round of feedback hand-in-hand with certainty ratings. Certainty ratings for high-conflict items were initially low for incorrect and correct responses alike. Following behavioral improvement, correct responses for high-conflict items started to receive higher certainty scores than incorrect responses, but the ratings remained lower for high- than low-conflict items. Response times similarly point to differences between the conflict conditions: Whereas correct responses on low-conflict items were consistently faster than incorrect responses across rounds, correct responses on high-conflict items were, in fact, slower than incorrect responses before feedback, but this pattern reversed after feedback. Note that additional representations of the behavioral data in terms of proportions of responses by certainty rating, and accuracy rates and certainty ratings by word category are presented in Supporting Information III and IV, respectively.

Response-Locked ERPs

The response-locked waveforms displayed two peaks, the first of which is referred to as the response-locked negativity, and the other as second negativity (see Figure 3). Trough-to-peak differences have additionally been visualized in bar graphs (Figure 3, middle panel).

To examine how the degree of language conflict affected error detection, we considered how the factors' accuracy (two levels), language conflict (two levels), and round (two levels) affected response-locked negativities. A three-way repeated measures ANOVA indicated main effects of conflict, $F(1, 22) = 47.69, p <.001,\ \eta_p^2 = .684$, and round, $F(1, 22) = 6.65, p = .017,\ \eta_p^2 = .232$, in combination with two-way interactions between conflict and accuracy, $F(1, 22) = 20.75, p < .001,\ \eta_p^2 = .485$, conflict and round, $F(1, 22) = 15.17, p = .001,\ \eta_p^2 = .408$, accuracy and round, $F(1, 22) = 17.80, p <.001,\ \eta_p^2 = .447$, and a three-way interaction, $F(1, 22) = 13.46, p = .001,\ \eta_p^2 = .380$. Paired samples *t* tests were performed to compare errors and correct responses per condition and round. These showed that low-conflict items yielded similar amplitudes for errors and correct responses before feedback, $t(22) = -1.69, p = .105$, but significantly larger amplitudes for errors compared to correct responses after feedback had been received, $t(22) = -4.44,\ p < .001$. High-conflict items, however, showed a reverse effect with larger response-locked negativities for correct compared to erroneous responses before feedback, $t(22) = 2.93$, $p = .008$, but larger ERN than CRN amplitudes after feedback, $t(22) = -3.10$, $p = .005$.

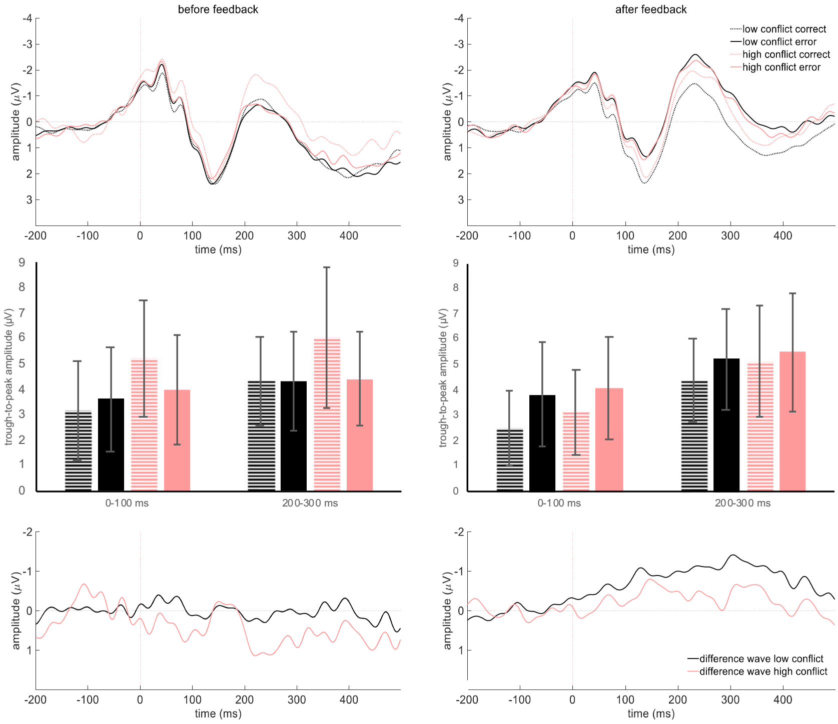

Figure 3 Behavioral and response-locked ERP data by conflict. Top panel shows response-locked waveforms at electrode FCz. Time point 0 on the *x*-axis indicates response onset. An additional baseline correction was conducted for visualization purposes only on the 200 milliseconds prior to the response. The middle panel shows bar graphs that represent trough-to-peak amplitudes before feedback (round 1) and after feedback (Rounds 2 and 3) for response-locked ERN/CRN (0–100 milliseconds) and subsequent second negativities (200–300 milliseconds) by accuracy and conflict condition. Error bars show standard deviations. The bottom panel shows difference waves (error - correct) for high- and low-conflict conditions before and after feedback. [Color figure can be viewed at wileyonlinelibrary.com]

When quantified as trough-to-peak differences, the second negativities following the ERN and CRN waveforms by and large mirrored the effects on the first negativities. A three-way repeated measures ANOVA on the second negativity showed a main effect of conflict, $F(1, 22) = 28.59, p < .001,\ \eta_p^2 = .565$, as well as two-way interactions between accuracy and conflict, $F(1, 22) = 14.29, p = .001,\ \eta_p^2 = .394$, accuracy and round, $F(1, 22) = 7.29, p = .013,\ \eta_p^2 = .249$, and a three-way interaction among accuracy, conflict, and round,

$F(1, 22) = 10.02, p = .004, \ \eta_p^2 = .313$. Paired samples *t* tests for low conflict items showed no difference before feedback, $t < 1$, but indicated significantly larger amplitudes for errors compared to correct responses, after feedback, $t(22) = -3.28$, $p = .003$, while high-conflict items showed a reverse effect with larger negativities for correct responses before feedback, $t(22) = 2.86$, $p = .009$, and a nonsignificant difference after feedback, $t < 1$.

Second negativities were additionally analyzed as mean amplitudes between 150 and 400 milliseconds; but a similar three-way repeated measures ANOVA showed no significant main effects or interactions (most *F*s < 1, *p*s $> .110$, $\eta_p^2 > .112$).

Correlation analyses were performed to examine the relation between individual difference measures (years of experience, AoA, LexTALE scores, and self-rated proficiency) and the four ERN effects (the average difference between ERN and CRN measures for the high- and low-conflict conditions in the before feedback and after feedback rounds for each individual). A Bonferroni correction (adjusted alpha $.05/(4 \times 4) = .003$) was applied to correct for multiple comparisons. These analyses revealed no significant effects.

Discussion

This study set out to examine how subjective certainty for a difficult-to-learn grammatical feature induced by conflicting language representations affects performance monitoring during learning. We aimed to test if a reduction in subjective certainty on response accuracy regarding gender assignment during L2 learning would be accompanied by an increase in the size of the ERN effect as an index of successful error monitoring. In addition, we wanted to see if the previously observed reverse ERN effect (Bultena et al., 2017) for items with a high degree of language conflict would stand the comparison with low conflict items. The findings are in agreement with the patterns predicted (see Table 1) and replicate effects observed in our previous study (Bultena et al., 2017).

Improved Performance and the Occurrence of Internal Monitoring

The behavioral results point to clear differences between performance on high- and low-conflict items, especially before learners were provided with feedback. Prior to feedback, the gender incongruent cognates yielded more errors than other items, replicating previous studies (Bultena et al., 2017; Lemhöfer et al., 2010). Interestingly, in contrast to the case of low-conflict items, response times in the high-conflict condition were slower for errors than for correct responses in this round, and certainty ratings for high-conflict items were generally low, regardless of accuracy. The lower certainty ratings for high-conflict items suggest that the L2 learners, most of whom were immersed in a Dutch environment and had thus probably been exposed to correct target language output, were to some extent aware of their incorrect intuitions. When these learners did give a correct response that violated their L1 intuitions, their response times slowed down, which may reflect response uncertainty, induced by experienced language conflict, as part of a learning process in progress. In comparison, erroneous responses on high-conflict items were relatively fast, suggesting learners trusted their L1 intuitions to be correct here. The low-conflict items indicated a different but very robust pattern that pointed to error-related uncertainty, in that incorrect responses were slower and received lower certainty ratings. In the course of learning, the response patterns for high- and low-conflict items became more similar, as indicated by improved accuracy rates, accompanied by higher certainty ratings and faster response times for correct responses in the high conflict condition. Learners were thus sensitive to feedback and learned fast, as indicated by an increase in response accuracy and certainty ratings.

The interaction effects observed in response times were consistent with response-locked ERP components. We observed a significant ERN effect for low conflict items after participants had been presented with feedback, with larger negativities for errors compared to correct responses, indicating that the correct grammatical gender for those items was being learned. In line with the predictions listed in Table 1, learners were thus able to internally detect errors on gender assignment, but only after a round of feedback. Cross-language gender incongruent cognates, on the other hand, did not elicit the typical ERN effect. In Round 1, that is, before participants had received any feedback, responses for these high conflict items showed an inverted ERN effect, with larger negativities for correct responses as compared to errors. Because the effects were quantified as trough-to-peak measures, the difference for the correct high conflict condition before feedback could in part have arisen from a difference in the trough (at around −100 milliseconds; see Figure 3, top panel). In order to check that differences before response onset were not the main reason for the effect, we additionally plotted the stimulus-locked data (see Supporting Information VI). In the stimulus-locked data, a difference appears to be present for high conflict items with more negative waveforms for high conflict correct responses, yet, this pattern does not readily account for the lower trough (i.e., more positive waveform) before response onset. Furthermore, an additional analysis of the trough-to-peak measure on the response-locked component based on a smaller search window for the trough (50 milliseconds before the negative peak), not reported here, still pointed to the same pattern with larger values for the correct responses compared to errors in the high conflict condition.

The feedback-based behavioral improvements for high-conflict items led to more typical ERN effects in Rounds 2 and 3, suggesting the development of effective error monitoring as the experiment progressed, in agreement with what we hypothesized (see Table 1). The patterns observed in the response-locked components were furthermore mirrored in the second negativities that followed them, but only when these were quantified as trough-to-peak measures.

It must be noted that the pattern observed in our data differs from the classic ERP response to errors in speeded response tasks (the so-called "oops-responses"). The response-locked component observed in our study peaked relatively early and the data do not show evidence for a typical biphasic ERN-Pe pattern, as the negative deflection observed in the present data could be said to be less sharp. Furthermore, the positivity observed at around 100 milliseconds is incongruent with the error positivity, which is commonly found between 200 and 500 milliseconds postresponse (Falkenstein et al., 2000; Nieuwenhuis, Ridderinkhof, Blom, Band, & Kok, 2001; Steinhauser & Yeung, 2012). Moreover, the waveforms following the ERN go in the opposite direction: The after feedback data point to a larger positivity for correct responses than errors, contrary to what would be expected of the Pe. Instead of a biphasic pattern, our data seem to show multiple negative peaks, which have previously been associated with theta oscillations in the ERN literature (cf. Gehring et al., 2011; Ullsperger, Danielmeier, & Jocham, 2014). The shorter peak latency of the effect replicates the effect observed by Bultena et al. (2017) and could be explained by the relatively long response times in the present task. The long response times could in turn have led to a preresponse rather than a postresponse conflict, that usually occurs for too fast error responses in speeded response tasks.

In spite of the differences between the typical error index and the ERP effect observed here, we consider this effect to come under the ERN umbrella. The differences in the shape of the effects can be accounted for by differences between the learning task used in the current study and the speeded response task typically used to elicit an ERN. The pattern seen in the present data is more consistent with learning paradigms or tasks that require memory retrieval, as was also observed in our previous learning study (Bultena et al., 2017). The similarity between the effects in the present and previous study support a robust role for internal monitoring mechanisms during learning. Comparable ERN studies that have investigated memory, language processing, and learning have found mixed evidence for the occurrence of a classic ERN. Three previous studies show a pattern similar to ours, reporting an ERN, but no Pe

(Rodriguez-Fornells, Kofidis, & Münte, 2004; Sebastian-Gallés et al., 2006) or a sustained negativity instead of a Pe (Davidson & Indefrey, 2011). Three other studies do show biphasic ERN-Pe effects, in errorless learning paradigms (Hammer, Heldmann, & Münte, 2013; Heldmann, Markgraf, Rodríguez-Fornells, & Münte, 2008) or in a language learning context (Davidson & Indefrey, 2009). The inconsistency likely reflects the fact that errors can have multiple possible causes (see Hoffmann & Beste, 2015), resulting in differences in the morphology of the ERN/Pe complex. In addition, the wide distribution of RTs in the decision task could have increased ERP variability: Peaks may have been present at slightly different response latencies within and across participants, with uncertainty arising either before, during, or slightly after button presses, which could have altered the ensuing negative deflection (see Falkenstein et al., 2000).

Alternatively, the pattern of multiple negative peaks, giving rise to a sustained negativity in the difference wave, could be thought of as a slow wave reflecting additional processing load for the high-conflict items in the decision task. Relatedly, the ERN has previously been interpreted to reflect an ongoing process of response checking (Falkenstein et al., 2000; Vidal, Hasbroucq, Grapperon, & Bonnet, 2000). Future studies could look at time–frequency analyses, in order to further evaluate such interpretation.

For now, we believe that our ERN interpretation is best suited to the data we have collected, since it can explain both the increase of an effect in response to feedback, as well as the difference between high and low conflict condition.

The Role of Conflicting Representations and Uncertainty

In terms of uncertainty, we note that the manipulation of language conflict led to more subjective uncertainty for high-conflict items, as intended, and that behavioral learning led to a reduction in uncertainty across high- and low-conflict items. Although a direct modulation of responses certainty in terms of the response-locked negativities could not be shown given the low number of trials for some of the certainty responses (see Appendix S4), the behavioral data did show that an increase in performance goes hand-in-hand with a reduction of uncertainty. Moreover, the increase in certainty ratings over rounds was accompanied by a discrepancy between ERN and CRN that increased as participants received more feedback. As soon as participants had learned from their mistakes and gave correct responses, they also managed to accurately detect their own errors, pointing to rapid updating of representations during learning.

The previously observed reversed ordering of ERN and CRN components for high-conflict items in round one of the learning task reported in Bultena et al. (2017) was confirmed more strongly by the current data. This implies that incorrect, L1-driven intuitions for cognates regarding gender assignment are very persistent. Errors elicited small ERNs prior to participants receiving feedback suggesting that they were then not detected as such. The large CRN component for correct responses is in line with previous findings that subjective certainty modulates the response monitoring process (Pailing & Segalowitz, 2004; Scheffers & Coles, 2000). The reversal of the ERN and CRN components prior to feedback could also be interpreted as incorrect error monitoring: Correct responses yielded an error signal, such that German learners of Dutch, when deciding on the correct determiner for cross-language gender incongruent cognates, based their first responses on L1 and perceived a subjective error when the answer was actually correct (cf. Lemhöfer et al., 2014). This idea is supported by the behavioral data, which show slower RTs for correct responses, in combination with relatively low certainty ratings. An alternative account of the inverted ERN effect could involve conflict monitoring: Response conflict, when present before participants give a response, is known to slow down RTs and increase error rates (Danielmeier et al., 2009) and has been associated with larger response-locked negativities in language production (Acheson, Ganushchak, Christoffels, & Hagoort, 2012). We prefer to interpret our findings in terms of uncertainty, however, because such interpretation offers a more general explanation of the mechanism underlying response monitoring, in line with a unifying account of the neural generator of the ERN effect (Alexander & Brown, 2011).

These results thus speak in favor of the idea that error monitoring depends on a subjective representation of correctness, in line with previous accounts of the ERN outside the domain of learning (Boldt & Yeung, 2015; Pailing & Segalowitz, 2004; Scheffers & Coles, 2000). The subjectivity of error monitoring is consistent with reverse effects in RTs and response-locked ERPs observed in round one for items with incongruent cross-language gender representations. Strong intuitions about what is a correct response caused by interfering L1 representations can lead to high levels of uncertainty. We furthermore note that ERN effects reported here were generally stronger than in our previous experiment (Bultena et al., 2017), which may be due to the collection of subjective certainty ratings, which have shown to increase performance monitoring elsewhere (Grützmann, Endrass, Klawohn, & Kathmann, 2014).

Conclusion

All in all, the present findings demonstrate that intuitions based on established knowledge, especially those calling upon incongruent representations for coactivated items in a bilingual's mind, as well as uncertainty about behavioral performance, play an important role in internal performance monitoring in a language learning context. In addition, our results highlight the use of ERP components relating to internal error monitoring, in the form of response-locked negativities, as useful tools to track the L2 learning process over time.

Final revised version accepted 29 December 2019

Open Research Badges

This article has earned Open Data and Open Materials badges for making publicly available the digitally-shareable data and the components of the research methods needed to reproduce the reported procedure and results. All data and materials that the authors have used and have the right to share are available at https://osf.io/ugx35. All proprietary materials have been precisely identified in the manuscript.

References

Acheson, D. J., Ganushchak, L. Y., Christoffels, I. K., & Hagoort, P. (2012). Conflict monitoring in speech production: Physiological evidence from bilingual picture naming. *Brain and Language*, *123*, 131–136. https://doi.org/10.1016/j.bandl.2012.08.008

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, *14*, 1338–1344. https://doi.org/10.1038/nn.2921

Boldt, A., & Yeung, N. (2015). Shared neural markers of decision confidence and error detection. *Journal of Neuroscience*, *35*, 3478–3484. https://doi.org/10.1523/JNEUROSCI.0797-14.2015

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977–990. https://doi.org/10.3758/BRM.41.4.977

Bultena, S., Danielmeier, C., Bekkering, H., & Lemhöfer, K. (2017). Electrophysiological correlates of error monitoring and feedback processing in second language learning. *Frontiers in Human Neuroscience*, *11*, 29. https://doi.org/10.3389/fnhum.2017.00029

Caffarra, S., Molinaro, N., Davidson, D., & Carreiras, M. (2015). Second language syntactic processing revealed through event-related potentials: An empirical review. *Neuroscience and Biobehavioral Reviews*, *51*, 31–47. https://doi.org/10.1016/j.neubiorev.2015.01.010

Danielmeier, C., Wessel, J. R., Steinhauser, M., & Ullsperger, M. (2009). Modulation of the error-related negativity by response conflict. *Psychophysiology*, *46*, 1288–1298. https://doi.org/10.1111/j.1469-8986.2009.00860.x

Davidson, D. J., & Indefrey, P. (2009). An event-related potential study on changes of violation and error responses during morphosyntactic learning. *Journal of Cognitive Neuroscience*, *21*, 433–446. https://doi.org/10.1162/jocn.2008.21031

Davidson, D. J., & Indefrey, P. (2011). Error-related activity and correlates of grammatical plasticity. *Frontiers in Psychology*, *2*, 219. https://doi.org/10.3389/fpsyg.2011.00219

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Endrass, T., Klawohn, J., Schuster, F., & Kathmann, N. (2008). Overactive performance monitoring in obsessive-compulsive disorder: ERP evidence from correct and erroneous reactions. *Neuropsychologia*, *46*, 1877–1887. https://doi.org/10.1016/j.neuropsychologia.2007.12.001

Eppinger, B., Kray, J., Mock, B., & Mecklinger, A. (2008). Better or worse than expected? Aging, learning, and the ERN. *Neuropsychologia*, *46*, 521–539. https://doi.org/10.1016/j.neuropsychologia.2007.09.001

Falkenstein, M., Hoormann, J., Christ, S., & Hohnsbein, J. (2000). ERP components on reaction errors and their functional significance: A tutorial. *Biological Psychology*, *51*, 87–107. https://doi.org/10.1016/S0301-0511(99)00031-9

Fischer, A. G., Klein, T. A., & Ullsperger, M. (2017). Comparing the error-related negativity across groups: The impact of error- and trial-number differences. *Psychophysiology*, *54*, 998–1009. https://doi.org/10.1111/psyp.12863

Gardner, R. C. (1985). The Attitude/Motivation Test Battery: Technical Report [Manual]. Retrieved http://publish.uwo.ca/˜gardner/docs/AMTBmanual.pdf

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, *4*, 385–390. https://doi.org/10.1111/j.1467-9280.1993.tb00586.x

Gehring, W. J., Liu, Y., Orr, J. M., & Carp, J. (2011). The error-related negativity (ERN/Ne). In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford Handbook of Event-Related Potential Components* (pp. 231–291). Oxford: Oxford University Press.

Grützmann, R., Endrass, T., Klawohn, J., & Kathmann, N. (2014). Response accuracy rating modulates ERN and Pe amplitudes. *Biological Psychology*, *96*, 1–7. https://doi.org/10.1016/j.biopsycho.2013.10.007

Hammer, A., Heldmann, M., & Münte, T. F. (2013). Errorless and errorful learning of face-name associations: An electrophysiological study. *Biological Psychology*, *92*, 169–178. https://doi.org/10.1016/j.biopsycho.2012.11.003

Heldmann, M., Markgraf, U., Rodríguez-Fornells, A., & Münte, T. F. (2008). Brain potentials reveal the role of conflict in human errorful and errorless learning. *Neuroscience Letters*, *444*(1), 64–68. https://doi.org/10.1016/j.neulet.2008.07.042

Hoffmann, S., & Beste, C. (2015). A perspective on neural and cognitive mechanisms of error commission. *Frontiers in Behavioral Neuroscience*, *9*, 50. https://doi.org/10.3389/fnbeh.2015.00050

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*, 679–709. https://doi.org/10.1037//0033-295X.109.4.679

Johnson, J. S., Shenkman, K. D., Newport, E. L., & Medin, D. L. (1996). Indeterminacy in the grammar of adult language learners. *Journal of Memory and Language*, *35*, 335–352. https://doi.org/10.1006/jmla.1996.0019

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*, 325–343. https://doi.org/10.3758/s13428-011-0146-0

Lemhöfer, K., Schriefers, H., & Hanique, I. (2010). Native language effects in learning second-language grammatical gender: A training study. *Acta Psychologica*, *135*, 150–158. https://doi.org/10.1016/j.actpsy.2010.06.001

Lemhöfer, K., Schriefers, H., & Indefrey, P. (2014). Idiosyncratic grammars: Syntactic processing in second language comprehension uses subjective feature representations. *Journal of Cognitive Neuroscience*, *26*, 1428–1444. https://doi.org/10.1162/jocn_a_00609

Lemhöfer, K., Spalek, K., & Schriefers, H. (2008). Cross-language effects of grammatical gender in bilingual word recognition and production. *Journal of Memory and Language*, *59*, 312–330. https://doi.org/10.1016/j.jml.2008.06.005

Luu, P., Collins, P., & Tucker, D. M. (2000). Mood, personality, and self-monitoring: Negative affect and emotionality in relation to frontal lobe mechanisms of error monitoring. *Journal of Experimental Psychology: General*, *129*, 43–60. https://doi.org/10.1037/0096-3445.129.1.43

Maier, M. E., & Steinhauser, M. (2016). Error significance but not error expectancy predicts error-related negativities for different error types. *Behavioural Brain Research*, *297*, 259–267. https://doi.org/10.1016/j.bbr.2015.10.031

Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J., Band, G. P., & Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology*, *38*, 752–760. https://doi.org/10.1111/1469-8986.3850752

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*, 97–113.

Olvet, D. M., & Hajcak, G. (2009). The stability of error-related brain activity with increasing trials. *Psychophysiology*, *46*, 957–961. https://doi.org/10.1111/j.1469-8986.2009.00848.x

Pailing, P. E., & Segalowitz, S. J. (2004). The effects of uncertainty in error monitoring on associated ERPs. *Brain and Cognition*, *56*, 215–233. https://doi.org/10.1016/j.bandc.2004.06.005

Rodriguez-Fornells, A., Kofidis, C., & Münte, T. F. (2004). An electrophysiological study of errorless learning. *Cognitive Brain Research*, *19*, 160–173. https://doi.org/10.1016/j.cogbrainres.2003.11.009

Scheffers, M. K., & Coles, M. G. H. (2000). Performance monitoring in a confusing world: Error-related brain activity, judgments of response accuracy, and types of errors. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 141–151. https://doi.org/10.1037//0096-1523.26.1.141

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*, 129–158. https://doi.org/10.1093/applin/11.2.129

Sebastian-Gallés, N., Rodríguez-Fornells, A., de Diego-Balaguer, R., & Díaz, B. (2006). First- and second-language phonological representations in the mental lexicon. *Journal of Cognitive Neuroscience*, *18*, 1277–1291. https://doi.org/10.1162/jocn.2006.18.8.1277

Steinhauser, M., & Yeung, N. (2012). Error awareness as evidence accumulation: Effects of speed-accuracy trade-off on error signaling. *Frontiers in Human Neuroscience*, *6*, 240. https://doi.org/10.3389/fnhum.2012.00240

Ullsperger, M., Danielmeier, C., & Jocham, G. (2014). Neurophysiology of performance monitoring and adaptive behavior. *Physiological Reviews*, *94*, 35–79. https://doi.org/10.1152/physrev.00041.2012

Van Casteren, M., & Davis, M. H. (2006). Mix, a program for pseudorandomization. *Behavior Research Methods*, *38*, 584–589. https://doi.org/10.3758/BF03193889

Van Orden, G. C. (1987). A ROWS is a ROSE: Spelling, sound, and reading. *Memory & Cognition*, *15*, 181–198. https://doi.org/10.3758/BF03197716

Vidal, F., Hasbroucq, T., Grapperon, J., & Bonnet, M. (2000). Is the “error negativity” specific to errors? *Biological Psychology*, *51*, 109–128. https://doi.org/10.1016/S0301-0511(99)00032-0

Wessel, J. R., Danielmeier, C., & Ullsperger, M. (2011). Error awareness revisited: Accumulation of multimodal evidence from central and autonomic nervous systems. *Journal of Cognitive Neuroscience*, *23*, 3021–3036. https://doi.org/10.1162/jocn.2011.21635

Wessel, J. R., & Ullsperger, M. (2011). Selection of independent components representing event-related brain potentials: A data-driven approach for greater objectivity. *NeuroImage*, *54*, 2105–2115. https://doi.org/10.1016/j.neuroimage.2010.10.033

Zheng, X., Roelofs, A., Farquhar, J., & Lemhöfer, K. (2018). Monitoring of language selection errors in switching: Not all about conflict. *PLoS ONE*, *13*(11), e0200397. https://doi.org/10.1371/journal.pone.0200397

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. Motivation Questionnaire
Appendix S2. Stimulus Materials
Table S2. Matching of stimuli per word category
Figure S3-1. Error rates over rounds per word category.
Figure S3-2. Certainty ratings (*SD*) over rounds per word category.
Appendix S4. Proportions of Responses by Certainty Rating
Figure S4-1. Proportion of error and correct responses for low conflict items, sorted by certainty rating.
Figure S4-2. Proportion of error and correct responses for high conflict items, sorted by certainty rating.
Appendix S5. Additional Information on Response-Locked Analyses
Figure S5. Response-locked waveforms for electrodes Fz (top panel), FCz (middle panel), and Cz (bottom panel).
Figure S6. Waveforms time-locked to stimulus onset of the noun presentation in round 1 (before feedback) and rounds 2 and 3 (after feedback).
Appendix S7. Results Regarding Feedback-Locked Components
Figure S7. Panel a shows feedback-locked waveforms at electrode FCz for positive and negative feedback on high and low conflict conditions per round.

Appendix: Accessible Summary (also publicly available at https://oasis-database.org)

How the Errors We Make Help Us Learn a Second Language

What This Research Was About and Why It Is Important

Acquiring a second language (L2) includes mastering the use of articles. Successful L2 learners know which articles goes with which noun and they are able to recognize their own errors. This is reflected by a wave of brain activity that peaks shortly after participants give a response and that signals whether the response was internally evaluated as correct or incorrect (a kind of automatic "oops-response"). Before such knowledge is in place, however, the learning process is characterized by uncertainty regarding the correct use of grammatical rules, especially when first language intuitions contradict the rules of the L2. In this study, we aimed to test how uncertainty and conflicting cross-language intuitions affect the brain's ability to monitor one's own errors. We found that L1 knowledge sometimes wrongly guides our intuitions, but that learners learn quickly from these errors.

What the Researchers Did

- We asked advanced German learners of Dutch to select the correct article to use with Dutch nouns according to grammatical gender (common *de* or neuter *het*) and to rate how certain they felt about their answer.
- To make the task more difficult, we included words similar across languages (or cognates, e.g., Dutch *auto* and German *Auto* both mean "car" in English), whose gender is different in the two languages (*auto* is assigned the common gender in Dutch, but the neuter gender in German, such that German learners think *het auto* rather than *de auto* is correct in Dutch).
- Participants received feedback after every trial and all words were presented in three subsequent rounds, such that they had a chance to learn from errors they made in a previous round.
- Throughout the task, we measured participants' brain activity using an electroencephalographic system.

What the Researchers Found

- As expected, L2 learners initially made many errors when they made grammatical gender decisions on cognate items whose gender differs between Dutch and German, such as *auto*. More interestingly, they showed error-like brain signatures for correct responses, and correct-like signature for errors, suggesting they were guided by L1 knowledge. They were also uncertain about their responses.
- After receiving feedback in the first round, the number of errors decreased markedly and brain signatures became more aligned with classic patterns: Brain responses to errors were increasingly more error-like, and participants indicated that they felt increasingly more certain about their responses.

Things to Consider

- Uncertainty and conflicting cross-language intuitions influence how learners perceive their own performance.

- Although learners quickly learned to give correct responses, their brain activity did not indicate immediate confident error monitoring. It is thus likely that new grammatical knowledge requires consolidation over time.

Materials and data: Materials are publicly available in the Supporting Information in the online version of this article at the publisher's website. See also https://osf.io/ugx35

How to cite this summary: Bultena, S., Danielmeier, C., Bekkering, H., & Lemhöfer, K. (2020). How the errors we make help us learn a second language. *OASIS Summary* of Bultena, Danielmeier, Bekkering et al. (2020) in *Language Learning*. https://oasis-database.org