

Syntactic Processing in L2 Depends on Perceived Reliability of the Input: Evidence From P600 Responses to Correct Input

Kristin Lemhöfer and Herbert Schriefers
Radboud University

Peter Indefrey
Radboud University and Heinrich-Heine University

In 3 ERP experiments, we investigated how experienced L2 speakers process natural and correct syntactic input that deviates from their own, sometimes incorrect, syntactic representations. Our previous study (Lemhöfer, Schriefers, & Indefrey, 2014) had shown that L2 speakers do engage in native-like syntactic processing of gender agreement but base this processing on their own idiosyncratic (and sometimes incorrect) grammars. However, as in other standard ERP studies, but different from realistic L2 input, the materials in that study contained a large proportion of incorrect sentences. In the present study, German speakers of Dutch read exclusively objectively correct Dutch sentences that did or did not contain subjective determiner “errors” (e.g., *de boot* “the boat,” which conflicts with the intuition of many German speakers that the correct phrase should be *het boot*). During reading for comprehension (Experiment 1), no syntax-related ERP responses for subjectively incorrect compared to correct phrases were observed. The same was true even when participants explicitly attended to and learned from the determiners in the sentences (Experiment 2). Only when participants judged the correctness of determiners in each sentence (Experiment 3) did a clear P600 appear. These results suggest that the full and native-like use of subjective grammars, as reflected in the P600 to subjective violations, occurs only when speakers have reason to mistrust the grammaticality of the input, either because of the nature of the task (grammaticality judgments) or because of the salient presence of incorrect sentences.

Keywords: second language, syntactic processing, gender agreement, correct input, P600


Supplemental materials: <http://dx.doi.org/10.1037/xlm0000895.supp>

Speaking several languages fluently and without errors seems to be one of the major requirements of modern life. Due to work, study, or family circumstances, many people frequently switch between languages or live in a second language (L2) environment entirely, where they are often faced with the unspoken expectation that their L2 use be functionally equivalent to that of a native speaker. However, reality tells another story: Even after many years of exposure to a second language (L2), the level of proficiency that learners reach usually falls short of native-like competence, certainly with respect to grammar (DeKeyser, Alfi-Shabtay, & Ravid, 2010; Hartshorne, Tenenbaum, & Pinker, 2018; Patkowski, 1980).

Consequently, a large number of studies have looked into the differences between first (L1) and second language (L2) syntactic processing and, in particular, whether and under which circumstances L2 processing becomes native-like. One of the approaches that cognitive scientists most frequently adopt to answer this question is the use of event-related potentials (ERPs) in the electroencephalogram (EEG; see Caffarra, Molinaro, Davidson, & Carreiras, 2015, for a review). This method has the advantage that it can be used during natural language comprehension without an additional, possibly unnatural and strategy-inducing task like grammaticality judgments.

In this field of research, L2 speakers are typically presented with (spoken or, more frequently, written) sentences that do or do not contain violations of the syntactic structure under investigation. The systematic deflections of the EEG that are elicited by such violations—the event-related potentials—are then compared to native speakers’ responses that are either collected in the same study or that are standardly reported by others. The most robust and commonly observed ERP component for syntactic violations is the P600, a positive, mostly parietally distributed deflection in the EEG starting around 500–600 ms after onset of the word at which the syntactic violation becomes evident (Hagoort, Brown, & Groothusen, 1993; Kaan, Harris, Gibson, & Holcomb, 2000; Molinaro, Barber, & Carreiras, 2011; Osterhout & Mobley, 1995; see also Brouwer, Fitz, & Hoeks, 2012; Sassenhagen, Schlesewsky, & Bornkessel-Schlesewsky, 2014; van de Meerendonk, Kolk, Viss-

This article was published Online First July 13, 2020.

 Kristin Lemhöfer and Herbert Schriefers, Donders Institute for Brain, Cognition and Behavior, Radboud University; Peter Indefrey, Donders Institute for Brain, Cognition and Behavior, Radboud University, and Department of General Linguistics, Heinrich-Heine University.

The raw data of the experiments are available from the Radboud University data repository at <http://hdl.handle.net/11633/aab2qbttd>. This work was supported by a *veni* grant from the Netherlands Organization for Scientific Research (NWO) to Kristin Lemhöfer (Grant 016.084.015). We thank Julia Lennertz for her assistance in collecting the data.

Correspondence concerning this article should be addressed to Kristin Lemhöfer, Donders Institute for Brain, Cognition and Behavior, Radboud University, P.O. Box 9102, 6500 HC Nijmegen, the Netherlands. E-mail: k.lemhofer@donders.ru.nl

ers, & Chwilla, 2010, for alternative accounts of the P600 than purely syntactic processing).

In that field of research, similar ERP patterns in L1 and L2 speakers are taken to indicate similar neural processes, such that L2 speakers' processes are considered "native-like" when their ERP signals become statistically indistinguishable from those of native (L1) speakers (e.g., they show a P600 of similar amplitude, scalp distribution, and latency). Conversely, ERP signals in L2 speakers that are qualitatively different from those in native speakers (e.g., displaying an N400 rather than a P600; Proverbio, Cok, & Zani, 2002; Tanner, Inoue, & Osterhout, 2014; Xue et al., 2013) are taken to reflect non-native-like processes. Finally, and importantly, a complete lack of an ERP difference between conditions with correct and incorrect sentences, as sometimes observed in L2 speakers (Hahne & Friederici, 2001; Ojima, Nakata, & Kakigi, 2005), is mostly interpreted as an absence of (online) sensitivity for the violated syntactic feature, for example, gender agreement (e.g., Tokowicz & MacWhinney, 2005). Note that the investigation of similarities and differences between L1 and L2 processing or between monolingual compared to bilingual L1 processing should not be seen as presupposing that monolingual L1 processing is the standard of comparison and that a deviation from that standard is in any way deficient (see, e.g., Fricke, Zirnstein, Navarro-Torres, & Kroll, 2019; Grosjean, 1989; Vanhove, 2019). Rather, both differences and similarities are relevant for a better understanding of L1 and L2 processing. To correctly interpret L1/L2 differences in terms of processing, however, it is crucial that the relevant experimental conditions should not have a "L1-bias."

In our previous study (Lemhöfer et al., 2014), for example, we showed that null-effects for (objective) violations of gender agreement do not necessarily indicate insensitivity of L2 speakers to a grammatical feature (like, in our case, grammatical gender), but can be due to the use of different, namely *subjective*, syntactic representations that can be objectively correct or incorrect. In that study, German L2 speakers of Dutch as well as a native Dutch control group read Dutch sentences for comprehension which contained gender-marked determiner-noun phrases that did or did not violate gender agreement between the determiner and the noun. While the Dutch control group showed the expected P600 for incorrect sentences, this ERP effect was missing in the German speakers of Dutch. However, we then conducted a second analysis of the same data of the L2 group in which the trials were not categorized according to objective correctness but according to subjective correctness. Subjective correctness had been assessed for each individual participant in a separate offline gender assignment task on the Dutch target nouns. These nouns included "gender-incompatible" cognate nouns for which gender assignment is notoriously difficult for German speakers of Dutch due to incorrect L1-to-L2 transfer (Lemhöfer, Schriefers, & Hanique, 2010; Lemhöfer, Spalek, & Schriefers, 2008). In this analysis, P600 effects (as well as the less standardly seen LAN effects) did arise in this group of L2 speakers in response to phrases that violated a given participant's idiosyncratic subjective gender representation.

Thus, in that study, only an analysis that took participants' "subjective grammars" into account was able to reveal that L2 speakers did actually process gender agreement similarly to native speakers, rather than being insensitive to it, as the null-effects in the original analysis for objective correctness initially seemed to

suggest (see Davidson & Indefrey, 2009; Foucart & Frenck-Mestre, 2011; Meulman, Stowe, Sprenger, Bresser, & Schmid, 2014; Sabourin, Stowe, & de Haan, 2006 for other examples of null-effects of gender agreement violations in L2 speakers). These null-effects in other studies adopting the traditional "objective correctness" manipulation might have been the result of a lack of control of subjective correctness: Many trials in the objectively correct condition might have been subjectively incorrect and have given rise to an unanticipated P600, while the opposite might have happened in the objectively incorrect condition (no P600 due to subjective correctness), hence blurring and possibly eliminating the difference between the two (objectively correct vs. incorrect) conditions. At least for the German L2 Dutch speakers tested by Lemhöfer et al. (2014), it was not the processing of gender agreement that was non-native-like but the nature of their L2 representations on which this processing was based.

To our knowledge, the Lemhöfer et al. (2014) study is the first to demonstrate the use of idiosyncratic, subjective grammars by L2 speakers in an online measure of syntactic processing like ERPs. The current study aims to further investigate the scope of this subjective grammar use, especially under conditions that are more realistic than was the case in the previous study. In that study, even though we aimed at mimicking a natural reading situation (e.g., by not using any additional task apart from comprehension questions), the experiment might not have been completely natural after all: As is inherent to the violation paradigm in syntactic processing research, our materials contained a substantial proportion of obviously incorrect sentences. The presence of these incorrect sentences was presumably very salient to the participants because nouns were presented twice, once with the correct and once with the incorrect determiner. Second, the materials contained not only instances of gender violations, but also of number agreement violations in a control condition, which are easy to detect for German speakers of Dutch. The large proportion of obviously incorrect sentences in Lemhöfer et al. is clearly at variance with natural L2 input, which, especially in the written modality, is usually correct.

Thus, it is not clear whether the use of subjective syntactic representations by L2 speakers is conditional upon the presence of obvious syntactic violations. Possibly, these violations can cause a certain degree of "syntactic awareness," or of "mistrust" in the input, that directs the reader's attention toward the syntactic violations and that causes syntactic processing to be more thorough—and, in this case, more native-like—than it would be under more realistic circumstances. Indeed, researchers have suggested that L2 speakers normally tend to make less use of syntactic cues during (natural) sentence comprehension ("shallow" processing), even if they master the relevant syntactic features in offline tests (Clahsen & Felser, 2006; Prévost & White, 2000; Sagarra & Herschensohn, 2010). This may be especially true for grammatical gender that is lexicalized (i.e. gender values have to be stored together with lexical items; cf. Williams & Lovatt, 2005) and therefore is hard to memorize. Furthermore, gender agreement is very rarely crucial for comprehension and disambiguation. Thus, possibly, L2 speakers do not normally process gender as thoroughly as L1 speakers, unless task or input characteristics (like a high incidence of agreement errors) lead them to do so. To test this possibility, we conducted similar experiments as in the previous study (Lemhöfer et al., 2014) but now avoiding the presence of objective agreement

errors by using materials similar to written L2 sentence input in real life.

Our investigation of the use of subjective grammars in L2 is closely linked to the so far essentially unexplored question of how L2 speakers process corrective input in real-time, that is, correct input that mismatches their own erroneous syntactic representations. Because our sentences will all be objectively correct, any mismatch between the input and an L2 speaker's subjective representation should, in principle, serve as a signal for learning and memory updating. Such learning would, of course, require the prior detection of the conflict.

As the syntactic feature under investigation, we use grammatical gender expressed on singular definite determiners in Dutch. Dutch definite determiners take the form *de* for so-called common gender nouns (collapsing masculine and feminine gender) and *het* for neuter nouns. Noun gender, especially in the case of the large number of cognate nouns (form-similar translations), is often, but not always "compatible" between translations in German and Dutch, with Dutch common gender corresponding to German masculine and feminine and Dutch neuter gender corresponding to German neuter gender.¹ We exploit this fact by using cognates that either have compatible gender in Dutch and German (e.g., neuter in both languages) or differ in gender between Dutch and German (e.g., neuter in Dutch but masculine in German). We know from previous studies that the L1-to-L2 transfer of gender for these types of words and in this population of speakers is pervasive, leading to mostly correct representations for compatible but incorrect representations for incompatible nouns (Lemhöfer et al., 2008, 2010). The stability of both correct and incorrect representations, a likely prerequisite of their use during syntactic processing, is high: Across two blocks of repeated item presentation, our earlier gender decision data show that 94% of compatible cognates that were correctly responded to in the first round were also responded to correctly in the second repetition; conversely, 81% of incompatible cognates that were assigned the incorrect determiner in the first block received the same response in the second block (Lemhöfer et al., 2010).

For our approach based on subjective correctness, we deviate from the standard methodology that is usually employed in ERP studies of syntactic processing (just as in Lemhöfer et al., 2014). Usually, the conditions (normally "syntactically correct" vs. "syntactically incorrect") are determined beforehand and administered to every participant alike. Here, we explicitly chose to keep all sentences objectively correct and compare sentences differing in *subjective correctness*. By definition, this subjective correctness is likely to vary from participant to participant. While one L2 learner might already have learned that the Dutch gender of *boot* ("boat") is not neuter like in German, but common gender (e.g., through having received explicit training on it), another might not. Therefore, we assessed each participant's idiosyncratic gender representations for our target words in a separate behavioral test session, for which we concealed as far as possible its relation to the nouns in the ERP experiment. These individual gender assignments then served as the basis for participant-specific sets of subjectively correct and incorrect experimental items.

Altogether, we will report three ERP experiments. In all three experiments, German speakers of Dutch read the same set of exclusively correct Dutch sentences, but we varied the degree of attention directed toward the syntactic feature under investigation,

that is, gender agreement, between determiners and nouns. This should allow us to detect the conditions under which L2 speakers do or do not engage in syntactic processing that is based on their subjective syntactic representations. Note that in contrast to the Lemhöfer et al. (2014) study, the present series of experiments will not involve a control group of Dutch native speakers, as these native speakers obviously do not have any incorrect syntactic representations.

As in the previous study, we opted for conditions that approximate real-life reading as closely as possible: First, at least in Experiment 1 that served as point of departure, participants read the sentences with no additional task than to comprehend the sentence content. Second, there were no objective syntactic errors in the materials to avoid an artificial focus on grammar. Third, we opted for high naturalness of the materials by constructing "real life" sentences similar to sentences taken from newspapers. As a result, these sentences thus varied considerably in terms of content and syntactic structure. While this choice necessarily implied some loosening of strict experimental control (like, e.g., having the critical words always in the same syntactic position), it allowed for participants to adopt a more natural reading mode. Furthermore, a large number (in our case, 272) of syntactically uniform sentences would likely have induced suspicions concerning the syntactic property of interest, as well as fatigue and boredom. While there was some random variation within the set of sentences, the materials were identical across the three experiments, which differed only in their task demands. Below, we will document details on this random variation.

Experiment 1: Sentence Reading for Comprehension

In Experiment 1, experienced German speakers of Dutch read correct Dutch sentences for comprehension. Critically, some of these sentences contained determiner-noun phrases that have a high chance of being subjectively incorrect to German readers. That is, they contained gender-incompatible cognates preceded by their gender-marked determiner, like *de boot* ("the boat," German: *das Boot*). The aim was to compare subjectively incorrect phrases with probably unproblematic ones containing gender-compatible cognates, like *de muis* ("the mouse," German: *die Maus*). To identify the correct and incorrect representations for each participant individually, participants completed an offline gender assignment task for the target nouns about one week before the EEG experiment. This task was followed by other language tests, to conceal the relation between the two experimental parts. Responses in the offline gender assignment task were then used to classify the critical sentences of the main ERP experiment into "subjectively correct" (i.e. the critical noun had been assigned the correct gender by this participant and matched the determiner occurring in the sentence) and "subjectively incorrect" (the critical noun had been assigned the incorrect gender. Thus it mismatched the correct determiner contained in the sentence).

¹ Official statistics for the occurrence of gender compatibility between Dutch and German are lacking. According to our own count of the 300 most frequent Dutch nouns in CELEX, 237 of them are cognates with German, and 88% of those cognates are gender-compatible with German. Among the (73) non-cognates, only 59% are gender-compatible.

Method

Participants. Given the lack of other previous ERP studies with a similar manipulation of subjective syntactic correctness in L2 speakers, we based the size of our participant sample on Lemhöfer et al. (2014). In that study, a P600 to the same kind of subjective violation of gender agreement in a participant sample from the same population used in the present study was reliably observed for a sample size of 20 participants. Using the effect size obtained in that study (in the 700–1000 ms window at posterior electrodes: Cohen's $d_z = 0.648$), we conducted a power analysis using G*Power, Version 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009), which showed that in order to obtain a power of .80, we would need a sample size of $N = 17$ to detect a similarly sized effect with a one-tailed test (which would be justified given the strongly predicted direction of the effect) and $N = 21$ with a two-tailed test (which we did use in the end). Thus, we decided to aim at a final sample size around 21. Note that this is also within the upper range of the sample size used in studies that investigate the processing of (objective) gender agreement violations between determiners and nouns in L2 speakers (Sabourin & Stowe, 2008: $N = 14$ [German native speakers] and $N = 8$ [native speakers of romance languages]; Dowens, Vergara, Barber, & Carreiras, 2010: $N = 22$; Dowens, Guo, Guo, Barber, & Carreiras, 2011: $N = 24$; Foucart & French-Mestre, 2011: $N = 16$).

Participants were 27 native speakers of German who were enrolled in a study program taught in Dutch at Radboud University Nijmegen (NL), that is, who were immersed in a Dutch environment. They all indicated to be right-handed and nondyslexic and had been raised with German as the only mother tongue. Four of them did not complete the second experimental session with the EEG measurements either because they did not show up for that session (two cases) or because their data from the offline gender assignment task in Session 1 showed too few errors to provide sufficient numbers of subjectively incorrect trials in the EEG analysis (<20 items; two cases). Furthermore, one participant was rejected for excessive artifacts and blinks (>30). Thus, 22 speakers remained for the final analyses.

Of these 22 speakers, all but one were female. They all completed a language background questionnaire, the main results of

which are summarized in Table 1, along with those of the participants of the other two experiments that will be reported further on. 91% of the participants indicated they had used Dutch recently (today or the day before). All participants also spoke English as a foreign language, with a self-rated frequency of use of 4.7 ($SD 1.3$) on a scale from 1 to 7.

Materials.

Target Nouns. We selected 68 pairs of Dutch nouns (gender-compatible and incompatible with German) from the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995) as target nouns. As we were striving for a substantial number of subjectively incorrect trials, and as we did not know in advance how many such trials each of our participants would contribute, we opted for a relatively high number of trials as our point of departure. Forty-nine out of the 136 words (36%) had also been part of the Lemhöfer et al. (2014) study (which contained only 80 words in the gender conditions). We did not include all words from that study because we had a different control condition (sentences with indefinite rather than definite determiners; see below), and not all words were suitable to be paired with an indefinite determiner (e.g., mass nouns).

All selected words were Dutch-German cognates, that is, words that are similar in form and have obvious common etymological roots with their German translation, with varying degrees of orthographic overlap between the translations (e.g., *helm*, German translation: Helm, "helmet"; *paleis*, German: Palast, "palace"). Critically, one member of each pair was compatible with the gender of its German translation (e.g., *muis* "mouse," which is common gender in Dutch and feminine in German), and the other member was incompatible (e.g., *boot* "boot," which is common gender in Dutch but neuter in German). With "compatible," we refer to nouns that are either neuter in both languages or common gender in Dutch and feminine or masculine in German (see also Lemhöfer et al., 2008). For half of the pairs (34), both nouns were of Dutch common gender taking the determiner *de*. For the other half, both nouns were of Dutch neuter gender (determiner *het*). The nouns in each pair were matched for Dutch and German lemma word frequency (absolute and Dijkstra logarithmic) according to CELEX, length in letters, and degree of orthographic overlap

Table 1
Characteristics of Participants in the Three Experiments

| Characteristic | Exp. 1 | Exp. 2 | Exp. 3 |
|-----------------------------------|-------------|-------------|-------------|
| <i>N</i> | 22 | 21 | 21 |
| Age | 23.1 (2.2) | 23.2 (2.8) | 22.9 (1.9) |
| (average) year of study | 2.9 (1.0) | 2.7 (1.3) | 2.8 (1.1) |
| Age of first Dutch acquisition | 20.1 (1.5) | 19.7 (1.3) | 20.2 (1.5) |
| Years of experience with Dutch | 3.2 (1.3) | 3.5 (2.0) | 2.8 (1.7) |
| LexTALE score (%) | 74.4 (3.7) | 74.3 (7.9) | 72.1 (10.0) |
| Self-ratings (1–7) | | | |
| Freq. of reading Dutch literature | 5.41 (1.65) | 4.95 (1.40) | 4.90 (1.48) |
| Freq. of speaking Dutch | 5.77 (0.97) | 5.62 (1.02) | 5.00 (1.55) |
| Freq. of Dutch TV/radio | 3.95 (1.68) | 3.67 (1.88) | 3.19 (2.02) |
| Overall reading experience | 5.09 (1.19) | 4.86 (1.15) | 4.86 (1.01) |
| Overall writing experience | 4.86 (0.99) | 4.52 (1.32) | 4.62 (1.28) |
| Overall speaking experience | 5.23 (0.97) | 5.00 (1.52) | 5.00 (1.27) |

Note. Standard deviations are given in parentheses. There were no significant ($p < .05$) differences between experiments for any of the measures in one-way ANOVA's.

between the translations, as measured by the percentage of overlapping letters (Dijkstra, Miwa, Brummelhuis, Sappelli, & Baayen, 2010).²

The characteristics of the target words are summarized in Table 2. All target nouns are listed in [online supplementary materials A](#).

Sentences. All target nouns, directly preceded by either their correct definite determiner or by the (non-gender-marked) indefinite determiner, were embedded in two sentences (one with the definite determiner *de* or *het* and one with the indefinite determiner *een*), like in the following examples for the target “boat”:

Na tien jaar moest de boot opnieuw geverfd worden.

(After 10 years, the boat had to be repainted.)

and

Het meisje wil met een boot de wereld rondvaren.

(The girl wants to sail around the world in a boat.)

The indefinite determiner condition was intended as a non-gender-marked control condition where potential a priori differences between gender-compatible and gender-incompatible cognate nouns could be controlled for.³

As mentioned above, it was important to us that participants would engage in as natural reading as possible. To this end, we used a wide variety of naturalistic sentences inspired by newspaper and magazine sentences like those in the German newspaper corpus Leipziger Wortschatz (<http://wortschatz.informatik.uni-leipzig.de/de>). Care was taken that no other incompatible cognates occurred in the sentences in a position preceding the critical one. Targets were never the first or last word in the sentence. The length of the sentences in words and the position of the target within the sentences are summarized for the four conditions in Table 3. All sentences are listed in [online supplementary materials A](#).

Cloze probability of the targets in the sentences was assessed in a sample of 15 native speakers of Dutch. Mean cloze probability for the critical determiner-noun phrase was 1.2% ($SD = 5.1\%$), with no significant differences between the four conditions ($p = .21$ in a one-way ANOVA).

Fifty-six of the sentences (= 20%) were followed by yes/no comprehension questions, which participants answered by pressing a button on a button box (right hand for yes, left for no). An example for such a question that, in this case, followed the sentence in the first example above, was: “Zag de boot er na tien jaar nog goed uit?” (“Was the boat still looking good after 10 years?”).

Procedure.

Behavioral session. About one week (five to nine days) before the ERP experiment, participants came in for a behavioral testing session to, as they were told, assess their language proficiency in Dutch. First, they received a paper questionnaire containing all 136 target items and were asked to write the correct definite determiner (*de* or *het*) in front of each noun. They were also asked to give a confidence rating of each response on a four-point scale (very uncertain/fairly uncertain/fairly certain/very certain).⁴ The order of items was identical for all participants, with no more than three compatible or incompatible items in a row and no more than three *de-* or *het-*items in a row.

After the gender assignment task, participants completed the Dutch version of the LexTALE (www.lextale.com; see Lemhöfer

& Broersma, 2012), a yes/no lexical-decision task to infrequent Dutch words and highly “plausible” nonwords, on paper, as well as a language background questionnaire already summarized in Table 1.

ERP session. In the second session, participants were asked to read Dutch sentences for comprehension while their EEG was being recorded. There were 272 sentences containing target nouns (136 targets, each occurring with the definite and indefinite determiner) distributed across six blocks with 44 or 46 sentences each. To make block length equal (47 sentences), the first one or three sentences of each block were filler sentences. 56 of the sentences (= 20%) were followed by yes/no comprehension questions. Fifteen of these questions followed sentences containing compatible targets in the definite condition, 10 for incompatible targets in the definite condition, 13 for compatible targets in the indefinite condition, and 18 for incompatible targets in the indefinite condition. Sentences with and without questions were randomly distributed over blocks but with the restriction that there were never more than two successive sentences with questions. The number of questions in a block of 47 sentences ranged between six and 12 ($M = 9.3$).

There were four experimental lists, counterbalancing item order and whether a given target occurred first with the definite or indefinite determiner. All targets occurred once in the first and once in the second half of the experiment. Target pairs (one compatible and one incompatible target, see Materials section) were kept together for maximal comparability, such that they always occurred together with the same determiner type in the same block. Sentence order had been pseudorandomized such that there were no more than three sentences containing compatible/incompatible targets or *de/het/een* determiners before the target in a row.

Sentences were presented word-by-word in the middle of the computer screen. The screen background was set to light gray, the letters were black 24 pt Arial letters. Before a sentence started, a fixation cross was presented for 500 ms in the center of the screen. The first word appeared after a blank screen had been shown for 250 ms. Each word was shown for 500 ms. In between words, there was a blank screen for 300 ms. This protocol followed the one used by Lemhöfer et al. (2014) that had proven appropriate for the reading speed of L2 speakers in Dutch (given that Dutch words can be very long due to compounding; see materials in [online supplementary materials A](#)).

After the last word of a sentence that was presented together with the period, there was an interval of 1500 ms (blank screen) before either the comprehension question or the next fixation cross occurred. Questions were presented as complete sentences and remained on the screen until the participant responded.

² The exact calculation of this measure was as follows: the number of overlapping letters (order-sensitive) in the two translations divided by the average of Dutch and German word length, multiplied by 100. Thus, e.g., the calculated overlap between *paleis* (Dutch) and *palast* (German) was 67% (four shared letters P, A, L, and S of six letters).

³ As explained below (EEG Data Analysis), this condition was dropped from the analyses later and regarded as a filler condition.

⁴ We assessed certainty in order to be able to split the data into certain vs. uncertain responses; however, cell sizes did not allow for this in the end.

Table 2
Characteristics of Dutch Target Nouns

| Characteristic | Compatible | | Incompatible | |
|---|------------------------|--------|------------------------|--------|
| | <i>M</i> (<i>SD</i>) | Range | <i>M</i> (<i>SD</i>) | Range |
| Length in letters | 5.72 (1.66) | 3–11 | 5.74 (1.58) | 3–11 |
| Dutch word frequency (per one million occurrences) | 41.1 (61.3) | 1–340 | 42.4 (57.6) | 1–298 |
| Log frequency Dutch ^a | 1.31 (0.54) | 0–3 | 1.33 (0.55) | 0–2 |
| German word frequency (per one million occurrences) | 37.4 (61.9) | 0–322 | 37.2 (59.5) | 0–341 |
| Log frequency German ^a | 1.13 (0.70) | 0–3 | 1.20 (0.60) | 0–3 |
| Orthographic overlap D-G (%) | 86.2 (14.6) | 50–100 | 87.1 (17.0) | 33–100 |

^a The logarithmic frequency was calculated as $\log_{10}(\text{word frequency per one million occurrences} + 1)$.

For the EEG recording, an elastic cap was used that contained 27 passive tin electrodes (Electro-Cap International, Eaton, OH). The positions of electrodes are shown in Figure 1. We also placed electrodes on the mastoids for referencing and on the forehead (between the eyes) as the ground. The data were first referenced to the left mastoid electrode and later rereferenced to the average of both mastoids. Impedances for the EEG electrodes were kept below 3 k Ω . To measure the EOG, two horizontal (at the outer side of both eyes) and two vertical electrodes (above and below the right eye) were placed. Impedances for EOG electrodes were below 5 k Ω . The EEG and EOG was sampled with a frequency of 500 Hz and amplified (time constant = 8 s, online bandpass = 0.02–30 Hz).

For the EEG analysis, the EEG and EOG signals were segmented into epochs from 100 ms before until 1000 ms after the onset of each target. The interval from 100 ms before until the onset of the target was used for baseline correction. Blink detection (using a threshold algorithm operating on the vertical eye channel) and ocular correction were applied using the Gratton and Coles algorithm as implemented in Brain Vision Analyzer Version 1.05 (Gratton, Coles, & Donchin, 1983). Artifacts were inspected after semiautomatic selection, that is, they were inspected if one or more EEG electrodes gave amplitudes below $-100 \mu\text{V}$ or above $+100 \mu\text{V}$ and subsequently rejected if a real artifact (rather than, e.g., a gradual drift) was present (0.7% of critical trials).

EEG data analysis. In our earlier study (Lemhöfer et al., 2014), we already observed similar effects as those that we were looking for here, that is, effects of violations of subjective correctness of word gender. This study showed, for a comparable population of Dutch L2 speakers, an anterior negativity (AN) at 200–600 ms postonset of the noun and a posterior positivity (P600) from 700 ms to the end of the epochs (1000 ms). Thus, we could have restricted our analysis to these time windows. However, in Lemhöfer et al. (2014), there were a number of important differ-

ences to the present study (most importantly, the presence of another violation type and a high proportion of obvious errors). Furthermore, Lemhöfer et al. (2014) was the only earlier study on subjective syntactic correctness we could rely on in this respect, and both the occurrence of an AN and the precise latency of a P600 is known to be highly variable between L2 studies (for reviews, see Caffarra et al., 2015; Kotz, 2009). Therefore, we decided to fully screen the data first in adjacent 100 ms windows for any potential effects and then to analyze the time windows that showed effects in that analysis.

For all analyses reported in the main text, we compared the two definite determiner conditions (subjectively correct vs. incorrect). To this end, we first selected only those target pairs (one compatible, one incompatible target, see above) for analysis for which the respective participant had given the “expected” responses in the offline gender task (i.e. the correct gender for the compatible target and the incorrect one for the incompatible target). These were, on average, 41 ($SD = 10$) of the 68 target pairs (= 60%). This way, we obtained equal cell sizes, as well as data from matched targets for the two conditions. We chose to discard the two indefinite “control” conditions (the same two sets of target nouns as for definite determiners, but now all paired with the non-gender-marked indefinite determiner *een*) for the main analysis. The reason for this is mainly that these two conditions gave rise to unexpected differential ERP patterns in some of the three experiments (see online supplementary materials B–D), while they should be equally correct subjectively. Our suspicion is that these effects may have arisen either because of some unfortunate instances of the use of indefinite determiners with respect to mass versus count nouns (e.g., “a rain,” which is possible, like in the Dutch phrase translated as “a rain of stars,” but not very common) or because of a hidden effect of the fact that the indefinite determiner is not gender-unmarked in German (*ein Mann* “a man,” *eine Frau* “a woman,” *ein Kind* “a child”). Hence, the use of this

Table 3
Characteristics of the Sentences in Experiments 1–3

| Characteristic | Definite determiners | | | | Indefinite determiners | | | |
|-------------------------------------|---------------------------|-------|-----------------------------|-------|---------------------------|-------|-----------------------------|-------|
| | Gender-compatible targets | | Gender-incompatible targets | | Gender-compatible targets | | Gender-incompatible targets | |
| | <i>M</i> (<i>SD</i>) | Range | <i>M</i> (<i>SD</i>) | Range | <i>M</i> (<i>SD</i>) | Range | <i>M</i> (<i>SD</i>) | Range |
| Sentence length (in words) | 9.0 (1.1) | 7–11 | 9.1 (1.2) | 6–11 | 9.0 (1.2) | 6–11 | 9.2 (1.0) | 7–11 |
| Target position in sentence (words) | 5.6 (1.4) | 3–9 | 5.9 (1.5) | 4–10 | 6.0 (1.6) | 3–9 | 6.1 (1.6) | 3–9 |

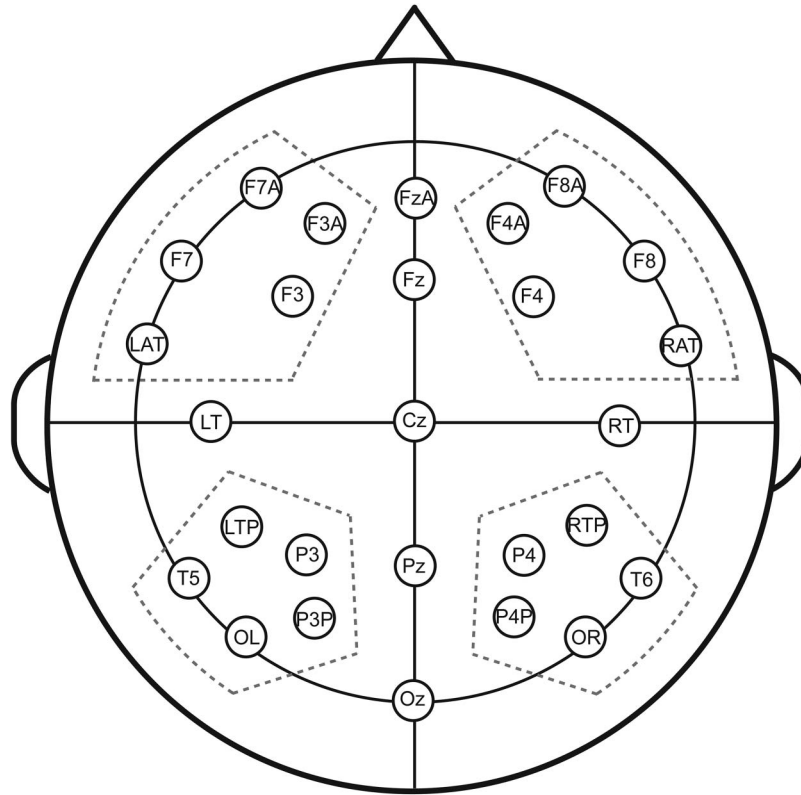


Figure 1. Positions of the electrodes on the EEG cap.

condition as a baseline did not turn out to be a very fortunate choice in the end, but the two definite conditions were well (pairwise) matched to function as their own controls. The graph showing all four conditions including the indefinite ones, as well as the full analysis of lateral sites in these four conditions, are both provided in [online supplementary materials B](#).

For the analysis of 100 ms windows (0–100 ms, 100–200 ms, etc.), we analyzed the midline as well as the four lateral quadrants of electrodes separately (the ones indicated by the dotted lines in [Figure 1](#)) by way of repeated-measures MANOVA's. These analyses were conducted with subjective correctness and electrode site as factors. As the effects we were looking for (P600 and possibly an earlier anterior negativity) should exceed a duration of 100 ms, our criterion for further analysis was an effect occurred in minimally two adjacent time windows exceeding a significance threshold of $p = .1$, corresponding to a combined significance threshold of ($p = 0.1 * 0.1 = 0.01$).

In the main analysis that followed in case of significant windows, we analyzed the midline and lateral electrodes separately. In the midline analysis, a repeated-measures ANOVA was conducted with the five electrodes as separate levels of the factor site and subjective correctness as the other factor. Possible interactions of site and subjective correctness were followed up with t tests (subjectively correct vs. subjectively incorrect) on each electrode.

In the overall analysis of lateral electrodes, data from the electrodes included in dashed lines in [Figure 1](#) were collapsed into quadrants. We conducted repeated-measures ANOVA's on these quadrant averages using the factors hemisphere (right vs. left),

region (anterior vs. posterior), and subjective correctness (correct vs. incorrect). Any possible interactions involving subjective correctness were followed up with planned simple effect ANOVAs to explore the nature of the interaction. In all analyses, we report only effects that involve the experimental factor subjective correctness.

Results

Behavioral results. In line with previous results (e.g., [Lemhöfer et al., 2010](#)), the error rate in the offline gender assignment task was 67.2% ($SD = 15.1%$) for gender-incompatible and 6.9% ($SD = 3.8%$) for gender-compatible nouns. Targets were included pairwise in the analysis if both items of a given pair elicited the “expected” response in the offline gender-assignment pretest, that is, when the gender-compatible noun received the correct gender and the gender-incompatible noun the incorrect gender. On average, 41 ($SD = 10$, range 22–59) of the 68 target pairs (= 61%) were included per participant. 0.68% of the critical trials were excluded as EEG artifacts.

The mean percentage of errors in the comprehension questions during the ERP sentence reading experiment was 6.3% ($SD = 3.1%$, range = 1.8–12.5%).

ERP results. [Figure 2](#) shows the ERP grand averages for subjectively correct and incorrect definite determiner-noun phrases in this experiment. Visual inspection suggests an unexpected sustained positive effect of subjectively incorrect compared to correct phrases at left-posterior sites starting from about 400 ms. Indeed,

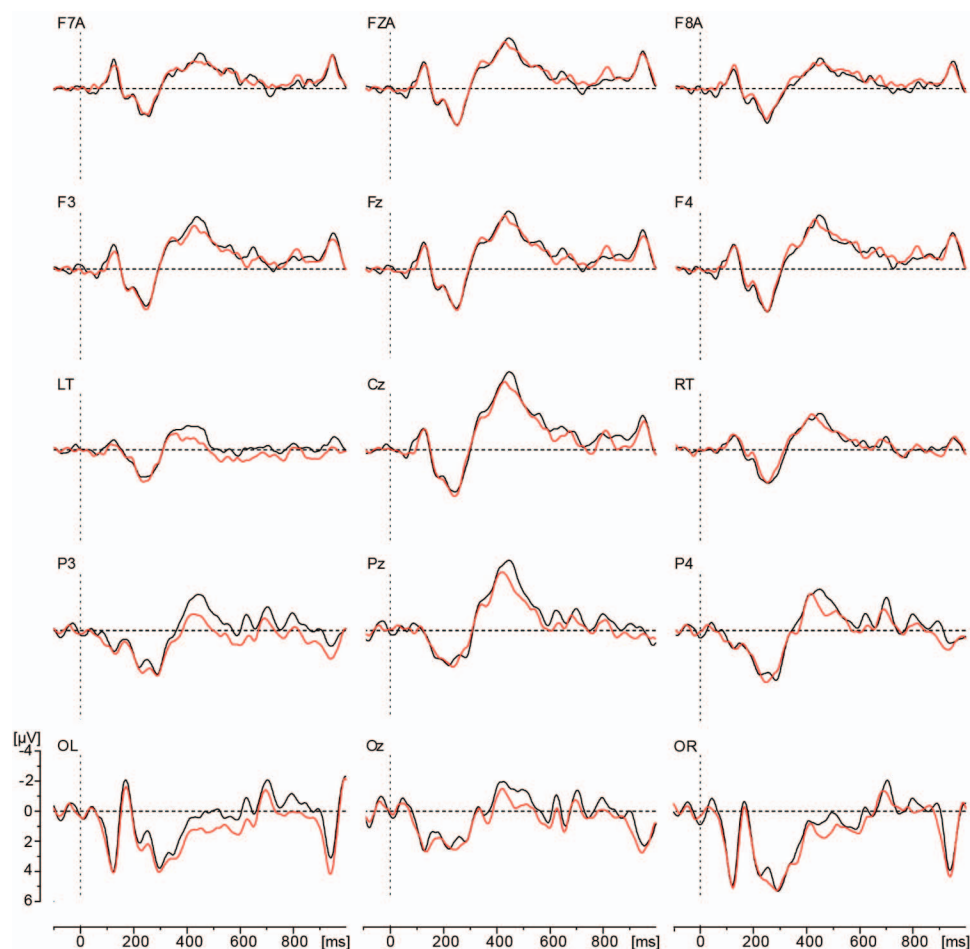


Figure 2. Grand-averaged ERP waveforms for the critical noun in subjectively correct (black) and incorrect (red or light gray) determiner-noun phrases in Experiment 1, for all midline and a subset of lateral electrodes. See the online article for the color version of this figure.

the analysis of 100 ms windows (see Table SB1 in online supplementary materials B) showed that this effect was present from 400 to 700 ms (though only marginally significant for 500–700 ms) after noun onset in the left posterior quadrant. We thus analyzed this time window further in an overall analysis (see “EEG analysis” in the methods section). Later time windows where a P600 would be expected (starting from 600 ms or later) did not show significant effects in this analysis.

The overall analysis of the 400 to 700 ms window showed no effect of subjective correctness in the midline analysis (both $F < 1$ for correctness and site \times correctness). On lateral electrodes, however, there was an interaction of hemisphere and subjective correctness, $F(1, 21) = 4.72$, $p = .041$, $\eta_p^2 = .184$. None of the other effects involving correctness were significant (all $p > .17$). Following up on the interaction, it became apparent that the effect was carried by the left hemisphere, which showed a (however nonsignificant) trend toward a positivity for subjectively incorrect phrases (mean difference $0.81 \mu\text{V}$; $F(1, 21) = 3.45$, $p = .077$, $\eta_p^2 = .141$). In contrast, the right hemisphere did not show an effect of correctness ($F < 1$).

Discussion

In Experiment 1, we looked for an electrophysiological correlate of the use of idiosyncratic gender representations and of the processing of subjective gender agreement violations. We expected the emergence of syntactic ERP effects in response to subjective incorrectness as compared to subjective correctness, similar to what we observed in Lemhöfer et al. (2014). In that study, L2 speakers showed a posterior positivity at 700–1000 ms (a P600) and an anterior negativity at 200–600 ms after onset of a noun that subjectively violated the gender of the preceding determiner. Here, we only observed a trend toward a different effect, namely a positivity at left posterior electrodes at 400–700 ms after noun onset. The time window of this effect is clearly too early to be interpretable as a P600, especially for L2 speakers (for which the P600 is typically delayed, if present at all; Kotz, 2009). Also, the left-lateralized distribution of this effect is not typical for a P600. Thus, the marginal significance, the early time window, the atypical left lateralization, and the divergence from our earlier results all speak against this effect being a genuine ERP correlate

of syntactic violation processing. In any case, the response we observed here is clearly less pronounced and less typical for a P600 in terms of latency and laterality than what we observed in the 2014 study.

Thus, in Experiment 1 with only objectively correct Dutch sentences, we did not observe the expected ERP effects for violations of subjective correctness when participants read sentences for comprehension. This seems to suggest that under the circumstances of the present study, experienced L2 speakers do not use their subjective grammars to engage in a similar syntactic processing routine as L1 speakers do. The findings are in contrast with the syntactic ERP effects for subjective correctness in our previous study on the same task, violation type, and population, but with the presence of objectively incorrect Dutch sentences (Lemhöfer et al., 2014). Possibly, our participants did not pay any attention to the determiners in the present situation where there was no reason to “mistrust” the correctness of the input. In contrast, the more “dubious” input in the Lemhöfer et al. (2014) study (where each target noun occurred, within one session, once with the correct and once with the incorrect definite determiner) might have induced an attentional focus on syntax (and more specifically, on determiner-noun phrases) in participants, giving rise to the observed effect.

To understand more precisely when L2 speakers do or do not use their subjective representations of word gender, we conducted a second experiment in which we introduced two new features compared to the previous one: First, we directed the participants’ attention toward the occurrence of subjective agreement errors, but this time (different to the Lemhöfer et al., 2014, study) while preserving the perceived reliability of the input. We did this by explicitly telling participants that the sentences they would read, and in particular the definite determiners, were all correct and that they should use them to correct their possibly incorrect intuitions about Dutch determiners. Second, this new instruction enabled us to perform an explicit, behavioral check whether the instances of mismatch between the input and own representations had been detected: We conducted announced gender assignment tests before and after the EEG session and used the performance difference between these tests as an index of learning. If participants improved on their gender assignments after compared to before reading the sentences, they must have detected the conflict between input and subjective representations during reading.

Hence, we repeated Experiment 1 with a new group of participants and with the same materials. Because we wanted to keep everything unchanged except for an additional attentional focus on determiners, we now gave participants a twofold instruction: First, to understand the sentences in meaning to be able to answer the occasional comprehension questions; and second, to attend to the used determiners and, if necessary, to use them to learn the correct gender in case of incorrect subjective gender representations. The offline gender assignment task was now administered immediately before the EEG experiment, as well as a second time after the experiment to assess the amount of learning. This posttest was announced to the participants in the beginning of the session, such that participants knew that their improvement on their use of Dutch determiners would be tested after the EEG reading experiment.

The question thus was whether the P600 that we observed in the Lemhöfer et al. (2014) study would reappear under these conditions of increased attention to subjective gender agreement. Note that a reoccurrence of the P600 under more syntax-focused in-

structions would be compatible with L1 studies showing larger P600’s in grammaticality judgment tasks than in tasks focusing on meaning (Hahne & Friederici, 2002; Kolk, Chwilla, van Herten, & Oor, 2003; Osterhout & Mobley, 1995).

Experiment 2: Sentence Reading for Comprehension and for Syntactic Learning

Method

General procedure. In contrast to the previous experiment, the experiment was administered in a single session. It started with the pre-experiment gender assignment task on paper (identical to the one in Experiment 1). After that, participants read the sentences with their EEG being measured, followed by the repetition of the offline gender assignment task. This posttest was identical to the pretest except for a new order of items that followed the same criteria (no more than three *de-* or *het-* words in a row, etc.). Finally, they completed the LexTALE vocabulary test and the language background questionnaire (see Table 1 for results).

Participants. 28 German speakers of Dutch, from the same population and selected with the same criteria as in Experiment 1, participated. Of these 28 participants, three made too few gender errors in the offline gender assignment task preceding the experiment, such that cell sizes for the subjectively incorrect condition would be too small (<20) to analyze. Two had to be excluded because of excessive artifacts. Finally, despite the instructions, two participants did not show any learning of determiners; that is, they did not show any increase in accuracy rates for the offline gender assignment tasks administered after compared to before the EEG experiment. All other participants showed an error reduction of at least 9.4% (mean 33% error reduction; see Table 4 for more details of learning). We decided to exclude the two participants who did not show learning because we could not be certain that they did indeed attend to the determiners in the sentences as instructed.

Thus, the final sample consisted of 21 participants (six males), 95.2% of them reported to have used Dutch the same day or the day before. They all also spoke English as a foreign language (self-rated frequency of use of 3.95 on a 7-point scale, $SD = 1.88$). The other results of the language background questionnaire are given in Table 1.

Materials and procedure. The sentences, trial lists, and procedural details were identical to those in Experiment 1, except for the new instruction of the EEG part and the offline posttest (see General Procedure above).

Table 4
Percentages of Errors in the Offline Gender Assignment Tasks in Exp. 2

| Target type | Before sentence reading with EEG | After sentence reading with EEG |
|----------------------|----------------------------------|---------------------------------|
| Compatible targets | 8.8 (7.1) | 8.0 (4.8) |
| Incompatible targets | 61.3 (16.1) | 39.3 (16.9) |
| Total errors | 35.1 (8.6) | 23.6 (9.1) |

Note. Standard deviations are given in parentheses.

Results

Behavioral results. All selected participants improved in gender assignment accuracy from before to after the sentence reading part. Table 4 summarizes the data of the two offline gender tasks.

As can be seen in Table 4, there was a clear learning effect, especially for incompatible targets as expected (error reduction of 36%). This learning effect was highly significant in a pairwise t test, $t(20) = 10.01$, $p < .001$, while the already low error rates for compatible targets did not significantly change, $t(20) = 0.59$, $p = .56$.

The mean error rate in the comprehension questions was 6.3% ($SD = 3.7$, range = 0–14.3%).

ERP results. Targets were included pairwise in the analysis depending on the offline gender decisions in the pretest, comparable to Experiment 1. On average, 37 ($SD = 11$, range 19–60) of the 68 target pairs (= 54%) were included per participant. 0.42% of the data were excluded as EEG artifacts.

The EEG grand averages of the two conditions involving definite determiners are plotted in Figure 3. The analysis of consecutive windows of 100 ms (see Table SC1 in the online supplementary materials C) showed only one effect (however only marginally significant for 200–300 ms), namely an unexpected positivity in the right anterior quadrant from 100 to 300 ms after onset of the noun. There were no effects in later time windows.

Again, we only report the analysis of the definite determiner condition here; see online supplementary materials C for a figure and the full analysis of lateral sites including the indefinite condition.

In the 100 to 300 ms window, there were no effects of subjective correctness in the midline electrodes (both $F < 1$). At lateral sites, there was a significant interaction of hemisphere and correctness, $F(1, 20) = 8.35$, $p = .009$, $\eta_p^2 = .295$. All other effects involving correctness were nonsignificant (all $p > .17$). Inspection of the descriptive data pattern showed that the effect of subjective correctness went into different directions descriptively, that is, a negativity in the left and a positivity in the right hemisphere. However, none of these effects was significant in isolation: Separate analyses of the two hemispheres showed no effects involving subjective correctness in any hemisphere (left: all $p > .19$; right: all $p > .14$).

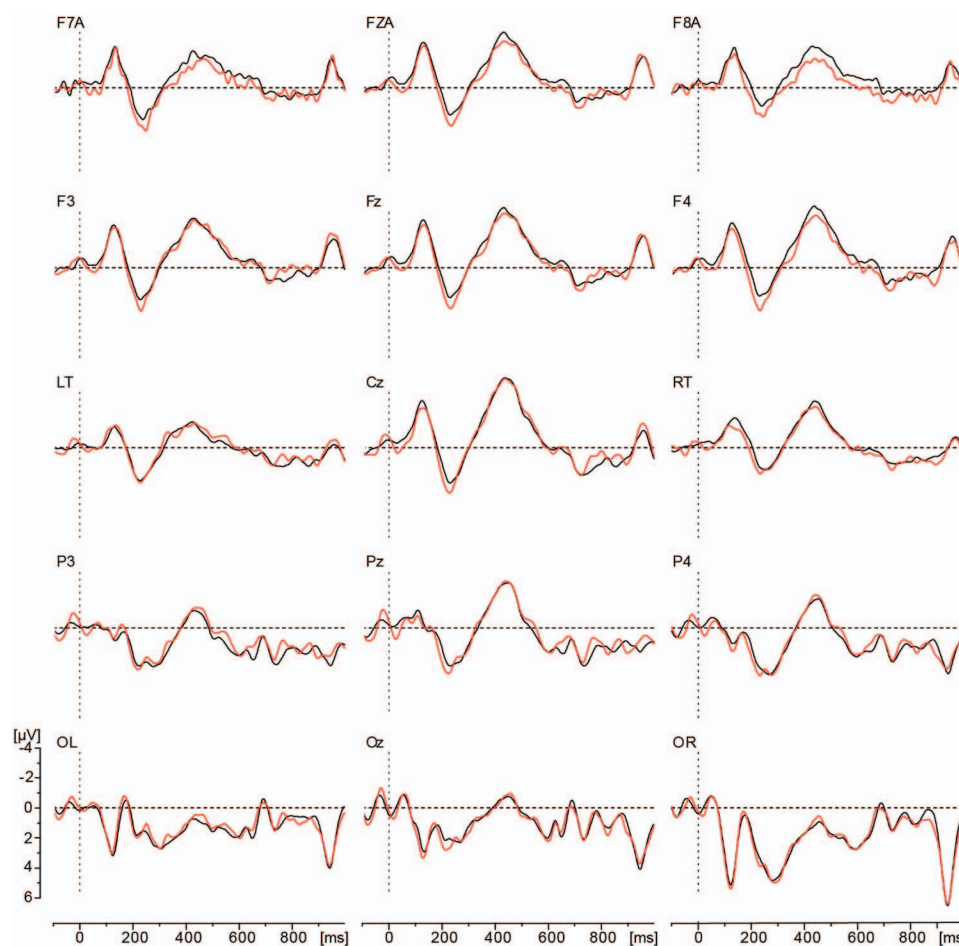


Figure 3. Grand-averaged ERP waveforms for the critical noun in subjectively correct (black) and incorrect (red or light gray) determiner-noun phrases in Experiment 2, for all midline and a subset of lateral electrodes. See the online article for the color version of this figure.

Discussion

Even though we had now asked the participants to pay special attention to the task-relevant determiners and to learn from them, we did not obtain any significant ERP differences between subjectively correct and incorrect determiner-noun phrases. The finding that we did not observe any standard syntactic ERP component like the P600 is especially surprising when considering that participants clearly learned from the input (see behavioral results, Table 4). Hence they must have detected the mismatch between the correct input and their incorrect gender representations for a considerable number of items. Apparently, mismatch detection per se does not elicit a P600 response.

One possible explanation is that the learning effect might have been carried by too few items to become observable in the EEG. On average, participants corrected their errors on 15 of the 68 incompatible targets, while their errors remained unchanged for 27 incompatible targets. We do not know whether they also detected the mismatch for these latter 27 targets. Maybe error detection, that is, the detection of a deviation of the input from a subjective gender representation, happened too infrequently to become visible in the ERP signal. However, an explorative analysis of the learned incompatible targets only (with only 16 participants who had at least learned 15 of these targets) did not show any trace of a P600 either.

A more likely implication of the results of Experiment 2 is that the P600 does not reflect the initial detection of a (here: subjective) grammatical violation but a later stage of further, more in-depth processing. We will return to this issue in more detail in the General Discussion. In short, the results of Experiment 2 suggest that mismatches between input and own syntactic representations were detected but not processed further in the way that L1 speakers do when they encounter (objective) syntactic errors. Thus, the allocation of attention to the to-be learned feature does not appear to be sufficient to elicit the P600 or any other visible ERP component in response to subjective syntactic violations.

But why, then, did a comparable group of L2 speakers show a P600 and anterior negativity to subjective violations of gender agreement between determiners and nouns in the Lemhöfer et al. (2014) study? There is one candidate characteristic left that differs between the Lemhöfer et al. study and the two experiments reported here, namely, the (perceived) syntactic reliability of the input. In our experiments, all sentences were correct, and, importantly, there was no reason for the participants to suspect that they might not be; in fact, in Experiment 2, they were explicitly told that all sentences were fully correct in Dutch. In contrast, half of the sentences in Lemhöfer et al. violated not only subjective but also objective correctness.

Our final experiment was designed to put this last candidate as a determinant for the occurrence of syntactic ERP effects (most notably the P600) in L2 speakers to a test. To this end, we tested whether violations of subjective correctness elicit the standard syntactic ERP effects when there are clear reasons to doubt the correctness of the input. Again, we kept the comparability between experiments maximal, using the same sentence materials and population, and changing only the task instructions. As a task that induces doubts on the reliability of the input, we used a grammatical (determiner) judgment task. That is, participants were asked to indicate, after each sentence, whether the determiners that had

occurred in the sentence were the correct ones for the noun they referred to. This may seem highly unusual at first, given that all sentences were actually correct; that is, the correct answer in terms of objective correctness would always be “yes.” However, with overall error rates of 35% or higher in the offline gender assignment task (collapsed across compatible and incompatible nouns), the task may not be perceived as unusual at all: on average, these 35% of the sentences would be judged “incorrect.” Importantly, giving this task would, of course, automatically elicit the implicit assumption in the participants that a substantial part of the sentences must be incorrect. If our idea holds true that perceived unreliability of the input drives P600 (and possibly AN) effects to subjective correctness violations, the P600 should reappear in this version of the sentence reading experiment.

Experiment 3: Sentence Reading With Grammaticality (Determiner) Judgments

Method

General procedure. In this experiment, the same sentences as in the two previous experiments were read by a new group of German speakers of Dutch, but with a different instruction: The task was a grammaticality judgment task. More specifically, participants were asked to attend to the definite determiner-noun phrases in the sentence and to press a button depending on whether these phrases had, according to the participant, all been correct, or whether at least one of them had been incorrect (i.e. the wrong gender-marked determiner preceded a given noun).

The experimental session started with the EEG sentence reading part. After that, we administered an offline gender assignment test as in the previous experiment, the LexTALE test, and the language background questionnaire (see Table 1 for the results of the latter two).

Participants. Another sample of 23 right-handed German speakers of Dutch took part, selected using the same criteria as before. One of these made too few gender errors, such that the number of target pairs that could be analyzed was too small (<20). One turned out to have been raised with a second native language and was excluded. Thus, the final sample for this experiment consisted of 21 participants.

Three of these participants were male. All except six (71%) reported to have used Dutch already on that day. All but two reported to speak English as a foreign language with a mean frequency of use of 4.27 ($SD = 1.61$) on a 7-point scale. All other results of the language background questionnaire are summarized in Table 1.

Materials and procedure. The sentences, trial lists, and procedural details were largely identical to those in Experiments 1 and 2. However, instead of occasional comprehension questions, the participants were now asked to indicate after each sentence whether they thought all the definite determiners in the sentences agreed in gender with the noun they were referring to. This was done by pressing a button on a button box (left = no, right = yes). If the answer had been “no,” they were asked to report to the experimenter which of the determiner-noun phrases they thought had been incorrect. The experimenter coded the answer and, in particular, whether the “no” answer referred to the target noun in the sentence.

For the sake of consistency with the other two experiments, we also administered an offline gender assignment test after the experiment, which was identical to the posttest in Experiment 2. However, the categorization of trials into subjectively correct versus incorrect ones was based on the most immediate judgments, that is, the grammaticality judgments.

In the end of the experimental session, the participants were fully debriefed about the design and hypotheses of the study.

Results

Behavioral results. Table 5 summarizes the data of the two tasks in which determiners were judged for correctness (grammaticality judgments), or produced on paper (offline task). As can be seen in Table 5, there were more errors on incompatible targets in the offline gender assignment task than errors (i.e. “no” answers) in the grammaticality judgments. This may be attributable to the hesitation to disrupt the flow of the experiment with a “no” answer during the EEG part because after every “no” answer, the participant had to tell the experimenter which noun he or she thought did not match its determiner. In fact, the mean consistency (i.e. the percentage of targets that received the same assignment in both tasks) was 80.5%.

ERP results. Trials were included in the ERP analysis when both members in a target pair received the “expected” button press in the immediate grammaticality judgments, that is, “yes” for the compatible and “no” for the incompatible targets. This was the case for, on average, 32 ($SD = 9$, range 19–56) of the 68 target pairs (= 46.7%). After blink correction, 0.12% of the data were excluded as EEG artifacts.

Figure 4 shows the ERP results for subjectively correct versus incorrect trials in the definite condition (see online supplementary materials D for the results including the indefinite condition). It can be seen from the figure that there was a large P600 effect at central-posterior electrodes. To explore the exact timing of this P600 (the latency of which is extremely variable in L2 populations), we again analyzed 100 ms time windows. This series of analyses (see Table SD1 in online supplementary materials D) showed a positivity at midline and posterior sites from 500 ms after onset of the noun until the end of the sampling period (1000 ms). We analyzed this time window again in a full analysis.⁵

The 500 to 1000 ms window showed a main effect of correctness at midline sites (more positive values for subjectively incorrect vs. subjectively correct trials; mean difference 0.97 μV ; $F(1, 20) = 6.16$, $p = .022$, $\eta_p^2 = .24$) and an interaction between correctness and site, $F(4, 17) = 3.48$, $p = .030$, $\eta_p^2 =$

.45. Pairwise t tests showed that the positivity was significant at central (Cz: mean difference 1.66 μV ; $t(20) = -3.10$, $p = .006$) and posterior sites (Pz: mean difference 1.58 μV ; $t(20) = -3.27$, $p = .004$; Oz: mean difference 0.86 μV ; $t(20) = -2.40$, $p = .026$).

The analysis of lateral sites showed no main effect of correctness, $F(1, 20) = 3.39$, $p = .08$. However, the effect of correctness interacted significantly with region, $F(1, 20) = 12.33$, $p = .002$, $\eta_p^2 = .38$. The other interactions were *ns* (p 's > .11).

We followed up the interaction of correctness and region by analyzing anterior electrodes separately from posterior ones. For anterior electrodes, there was no effect of correctness and no interaction of correctness with hemisphere (p 's > .50). In contrast, posterior sites showed a main effect of correctness (a positivity for incorrect trials; mean difference 1.09 μV ; $F(1, 20) = 9.49$, $p = .006$, $\eta_p^2 = .32$), replicating the P600 at posterior electrodes observed for subjectively incorrect determiner-noun phrases in the previous study (Lemhöfer et al., 2014). Note that the effect size ($\eta_p^2 = .32$) was similar to the one obtained in that study ($\eta_p^2 = .31$). Furthermore, the interaction between correctness and hemisphere for these posterior sites was not significant, $F(1, 20) = 3.07$, $p = .095$.

Discussion

In the last experiment of the series of three, we asked participants to judge the grammaticality of the input (and in particular, the correctness of the definite determiners), thereby implicitly suggesting that a substantial proportion of the sentences were grammatically incorrect. Remember, however, that objectively, this was not the case: All sentences were correct. With this modification of the instruction, we obtained a fairly large and early P600 for subjectively incorrect trials that started at 500 ms after onset of the critical noun and lasted until the end of the sampling interval (1000 ms). Even though we could not test a native control group due to the nature of the materials (only objectively correct sentences), this effect seems perfectly in line with P600 effects that native speakers show in response to objective determiner-noun agreement violations in terms of size, latency, and scalp distribution, at least in languages with nontransparent gender (see, e.g., Gunter, Friederici, & Schriefers, 2000; Hanulíková, van Alphen, van Goch, & Weber, 2012). It is also very similar to the one (550–1000 ms) we found for the same kind of gender agreement errors in native speakers in Lemhöfer et al. (2014), though in that study, readers did not give grammaticality judgments. Thus, under these circumstances of apparent input unreliability, the L2 speakers responded to “violations” in a native-like fashion.

General Discussion

In the present study, we conducted three experiments to investigate under which conditions L2 speakers do or do not use their

Table 5

Percentages of Errors in the Immediate Grammaticality Judgments and the Post-Experiment Offline Gender Assignment Task in Experiment 3

| Target type | Grammaticality judgments | Offline gender assignment task |
|----------------------|--------------------------|--------------------------------|
| Compatible targets | 6.7 (6.0) | 6.0 (5.5) |
| Incompatible targets | 52.1 (14.7) | 63.8 (17.4) |
| Total errors | 29.4 (8.7) | 34.9 (8.6) |

Note. Standard deviations are given in parentheses.

⁵ Inspection of Table SD1 also shows that additionally, in the left anterior quadrant, there was a (marginally) significant Correctness \times Site interaction from 700–900 ms due to a negativity at F7. However, because this time window would have been carried by only one electrode, and it was embedded in a larger analysis time window, we did not select it for a separate analysis.

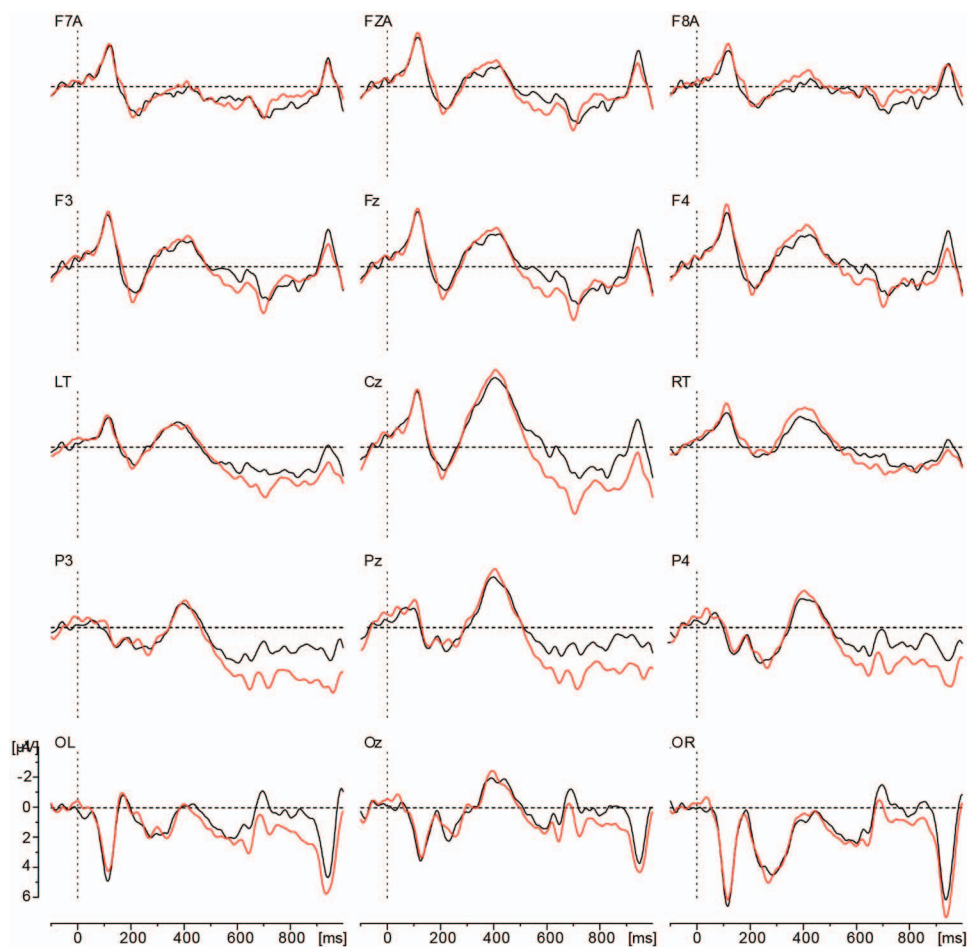


Figure 4. Grand-averaged ERP waveforms for the critical noun in subjectively correct (black) and incorrect (red or light gray) determiner-noun phrases in Experiment 3, for all midline and a subset of lateral electrodes. See the online article for the color version of this figure.

subjective and sometimes incorrect grammars during syntactic sentence processing. In particular, we wanted to shed more light on the neural correlates of L2 input situations in which the correct L2 input deviates from the speakers' incorrect subjective representations.

In our previous study (Lemhöfer et al., 2014), a reanalysis (in terms of subjective correctness) of what first looked like a null-effect (in terms of objective correctness) had shown that German advanced speakers of Dutch process grammatical gender very similarly to native speakers, except that they base that processing on their subjective and sometimes incorrect gender representations. However, like almost all other studies in the (L1 or L2) syntactic processing literature, this study included a high proportion of objectively incorrect sentences, possibly inducing a more syntax-focused processing strategy than would occur during natural sentence comprehension. Therefore, we conducted all experiments in the present study using only objectively correct sentences, an approach that is, to our knowledge, unprecedented in the relevant ERP literature.

In Experiment 1, German speakers of Dutch read correct Dutch sentences for comprehension while we measured their ERP re-

sponse on nouns that were preceded by the correct gender-marked definite determiner. Critically, this determiner could either be subjectively correct or incorrect, as assessed in an offline task prior to the experiment. In contrast to the findings by Lemhöfer et al. (2014), no significant ERP effects that are generally associated with the processing of a grammatical violation emerged. Thus, simply reading "unsuspicious" sentences for comprehension in the absence of obvious errors failed to evoke the thorough, native-like evaluation of grammatical correctness in L2 speakers that we had observed before. In Experiment 2, we again asked participants to read and understand the same sentences, but added the instruction to attend to the presented, correct gender-marked definite determiners and, where applicable, to learn from these. Gender assignment tests on the target nouns were administered before and after the EEG reading part. Participants showed considerable behavioral improvement from pre- to posttest, indicating that they had followed the instructions. However, surprisingly, again no ERP effects signaling the processing of subjective syntactic violations emerged. Finally, in Experiment 3, we again presented the same sentences, but had the participants judge the correctness of the (objectively correct) definite determiners in each sentence. With

this induced “suspicion” that the input was presumably not always grammatically correct, a large parietal P600 finally emerged.

At this point, a methodological note seems in order. We used natural sentences with variable syntactic structures, and we did so in order to mimic natural input within an experimental setting. Furthermore, our experimental design required to determine the exact composition of conditions individually per participant. Given these two points there is the theoretical possibility that the exact composition of conditions—that is, which individual sentences in a certain condition were included for which participants—differed between the three experiments, possibly contributing to the observed difference in ERP results (i.e. occurrence of a P600 in Experiment 3, but not in Experiments 1 and 2). Therefore, we compared the overall composition of conditions, that is, the set of all sentences that had actually entered the analyses for any participant, across experiments with respect to a number of syntactically relevant parameters (sentence length, position of target noun phrase [NP], syntactic class of the word preceding the target NP, syntactic role of target NP, and mean certainty ratings for offline gender assignments; see [online supplementary materials E](#) for tables, graphs, and analyses). We report only on the two definite conditions in this respect because the indefinite condition was ultimately treated as a filler condition (but note that in parallel to the definite conditions, no systematic confounds were revealed there either).

The graphs, tables and analyses presented in [online supplementary materials E](#) show that while there were a few subtle differences in the composition of conditions between experiments, none of these can be responsible for the pattern of results (P600 effects in Experiment 3 only) because those differences that exist show smaller differences between compatible and incompatible conditions in Experiment 3 relative to Experiments 1 and 2.

To summarize the earlier study ([Lemhöfer et al., 2014](#)) and the three experiments of the present study, we systematically varied the composition of the materials (the “traditional” mix of objectively correct and incorrect sentences in the earlier study vs. only objectively correct sentences in the present study) and the focus of attention (sentence content vs. determiner-noun agreement). Those two factors presumably resulted in systematic differences in the perceived grammatical reliability of the input. [Table 6](#) gives a schematic overview over all these experiments, the dimensions on which they differed, and whether we observed a P600 in the L2 learners or not.

This overview suggests that it is the factor of perceived reliability of the input (text in bold) that determined the occurrence of the P600 in our data: When the L2 speakers had no reason to mistrust the grammaticality of the input (Experiments 1 and 2 of the present study), they did not show any consistent ERP effect of subjective (in)correctness, and in particular no P600. In contrast, when they did have reason to doubt the correctness of the input (either because of the presence of obvious errors in the sentences, or because of a grammatical judgment task), they showed effects that are very similar to those in native speakers (see [Lemhöfer et al., 2014](#), for data of a native control group on the same type of sentences as used here). Thus, it seems to be a lack of trust in the correctness of the input that causes the emergence of full syntactic violation processing as we know it from monolinguals, mostly reflected in the P600.

Our data indicate that the P600 in L2 speakers does not reflect the *detection* of a subjective syntactic violation, that is, the detection of the mismatch between the input and subjective syntactic representations. Most critical to this conclusion are the results of Experiment 2: In this experiment, participants learned from those definite determiners which were in conflict with their own incorrect gender representations. This learning can hardly have happened without the prior detection of learning-relevant input (the mismatch between input and own representation). However, crucially, the behavioral improvement did not give rise to a P600 (or any other standard ERP component) for sentences that mismatched the speakers’ representations. This pattern of results implies that the P600 does not reflect processes of (subjective) violation detection but of *further processing*, which apparently does not normally take place when the input is perceived as reliable.

Indeed, the processing of syntactic violations is usually regarded to be a multiple-stage process, starting with the detection of the error and followed by a later, more strategy-driven stage of re-analysis, repair, monitoring, or context integration (e.g., [Brouwer et al., 2012](#); [Friederici, 1995](#); [Friederici, Steinhauer, & Pfeifer, 2002](#); [Kuperberg, 2007](#); [van de Meerendonk et al., 2010](#)). While the P600 is usually associated with the latter processing stage, the link between morphosyntactic error detection and ERP effects is less clear. This first stage has sometimes, but not always been claimed to be reflected in earlier negativities ([Bornkessel & Schlesewsky, 2006](#); [Münte, Matzke, & Johannes, 1997](#); see [Molinero et al., 2011](#), for an overview of the various accounts of ERP effects in L1 agreement processing). However, note that agreement violations do not always elicit early negativities even in native

Table 6
Overview of the Four Experiments of [Lemhöfer et al. \(2014\)](#) and the Present Study Concerning the Dimensions on Which They Differ

| Experiment | Input composition | Focus of attention | Perceived input reliability | Observed P600 |
|---|--------------------------------------|-----------------------|---|---------------|
| Lemhöfer, Schriefers, and Indefrey (2014) | correct & incorrect determiners | content | no (frequent occurrence of obvious errors) | yes |
| Present study exp. 1 | only objectively correct determiners | content | yes | no |
| Present study exp. 2 | only objectively correct determiners | content & determiners | yes | no |
| Present study exp. 3 | only objectively correct determiners | determiners | no (grammaticality judgment task) | yes |

Note. The dimension of interest, the occurrence of the P600, and the dimension that seems to determine it, perceived input reliability, are printed in bold to illustrate their tight relation.

speakers (e.g., Hagoort & Brown, 1999; Wicha, Moreno, & Kutas, 2004), let alone in L2 speakers (see Caffarra et al., 2015; Steinhauer, White, & Drury, 2009, for reviews), but such violations do reliably elicit P600's. Because processes of reanalysis and repair presumably require the prior detection of the error, it thus seems most likely that error detection alone is not (always) visible in EEG signatures. This is in line with the absence of ERP effects despite evidence of behavioral learning from subjective violations in Experiment 2. It also fits with other findings of a lack of ERP effects of gender violation in L2 speakers in the presence of (clearly above-chance) behavioral mastery of L2 grammatical gender, at least in case that uncontrolled differences in subjective correctness were not a major issue (Davidson & Indefrey, 2009; Meulman et al., 2014; Sabourin & Stowe, 2008).

We also conducted an analysis of the time-frequency data of Experiments 1 and 3 (Lewis, Lemhöfer, Schoffelen, & Schriefers, 2016), which would have gone beyond the scope of the present (already very extensive) report. The results suggested that the processing of gender violations is not only reflected in an event-related P600 response but also in an event-induced power increase in the beta band of the EEG frequencies, a band that has previously been associated with syntactic processing (Lewis, Wang, & Bastiaansen, 2015; Weiss & Mueller, 2012). Similar to the pattern observed in the present report, no significant differences between subjectively correct and incorrect phrases were observed for this band in Experiment 1, but differences did arise in Experiment 3.

Thus, the full pattern of our results across the four experiments summarized in Table 6 suggests that the mere detection of a subjective gender violation can take place without a corresponding trace in the EEG. This is remarkable, given that EEG effects to L2 input in beginning learners have occasionally been reported to precede behavioral effects (McLaughlin, Osterhout, & Kim, 2004). Still, violation detection without EEG effects is what might have happened in Experiment 1, although in that case, we have no (behavioral) means to tell whether participants detected the subjective violations in that experiment or not. But it is surely what happened in Experiment 2, where we do have such behavioral evidence for error detection. This finding has implications for the L2 syntactic processing literature in general: While null effects in the ERP are often interpreted as a complete lack of sensitivity to the type of syntactic violation (e.g., Meulman et al., 2014; Tokowicz & MacWhinney, 2005), we show that the mere detection of violations can take place without a visible correlate in the EEG. Our data also demonstrate that what is reflected in the P600 is the further processing of a previously detected violation, which, apparently, L2 speakers only engage in if they cannot trust the grammaticality of the input. Thus, a lack of trust in the input, and not the amount of attention to the critical syntactic feature, seems to be the decisive factor for the full use of subjective grammars in L2 speakers.

It should be noted however that the idea of perceived input reliability as a crucial factor for the occurrence of a P600 took shape mainly during the course of the present study (more specifically, after Experiment 1 and its conflicting results with those of Lemhöfer et al., 2014). We then went on to test this idea with what we thought would manipulate perceived input reliability in Experiments 2 and 3, but without explicitly measuring it (which seems hard to do). The results of those experiments were consistent with our idea. While this is reassuring, the post hoc nature of the

hypothesis requires additional and independent future empirical evidence before we can more safely rely on it.

Our findings also have implications for the nature of the P600 in general, even though we have to be cautious in generalizing them to native speakers. But also in native speakers, the P600 has been found to be eliminated or greatly reduced under certain circumstances, for example, when ungrammatical sentences are the rule rather than the exception in an experimental block (Coulson, King, & Kutas, 1998; Gunter, Stowe, & Mulder, 1997), when auditorily presented sentences are spoken with a foreign accent (Hanulíková et al., 2012), when the task is a superficial one and thus does not require syntactic analysis (Gunter & Friederici, 1999; Verhees, Chwilla, Tromp, & Vissers, 2015), or when the violation is not very salient or severe (Coulson et al., 1998; Mueller, Hahne, Fujii, & Friederici, 2005). Furthermore, the P600 seems to be generally smaller, or even absent, during mere reading for comprehension compared to reading for grammaticality judgments (Hahne & Friederici, 2002; Kolk et al., 2003; Osterhout & Mobley, 1995). Due to these modulations, the original view of the P600 as a pure reflection of linguistic syntactic processing has been questioned. Alternative proposals include the view of the P600 as a member of the domain-general family of P300 effects associated with the detection of improbable, task-relevant events (Coulson et al., 1998; Sassenhagen et al., 2014), and the claim that the P600 is a correlate of language monitoring, that is, a checking process for possible perceptual errors after encountering an unlikely linguistic event (Kolk et al., 2003; van de Meerendonk et al., 2010; van Herten, Chwilla, & Kolk, 2006).

Thus, it seems that even native speakers occasionally refrain from engaging in the (probably costly) reanalysis, repair, monitoring and/or integration processes reflected in the P600, depending on the degree of expectation of syntactic violations and on task characteristics. Our data add to this picture that perceived input reliability, or the anticipation of the occurrence of errors, also seems a crucial factor for the employment of these processes, at least in L2 speakers. However, its role cannot be assessed in traditionally designed ERP experiments, and especially not in native speakers, because these experiments necessarily comprise an objectively incorrect condition, which makes the input appear unreliable by definition. It is only with L2 speakers for whom objective and subjective correctness occasionally diverge that one gets a chance to investigate violation processing while at the same time preserving objective correctness of the input.

One final question concerns the generalizability of our results to other incorrectly represented aspects of L2 grammar. As already mentioned, correct gender agreement between determiner and noun is a grammatical feature that is usually not essential for comprehension. Thus, paying attention to this feature and trying to improve on it has presumably a relatively low priority for L2 speakers compared to other grammatical features that are more crucial for disambiguation during comprehension, like subject-verb or number agreement. The question is thus whether our results, pointing at a role of perceived input reliability for the deeper processing of mismatches between input and own representations, generalize to the processing of those other features. While we can only speculate on this point, certainly it should be mainly L2 errors on "nonessential" features that give rise to such a high incidence of robust errors. If L2 speakers failed to be

understood every time they erroneously said “het boot,” they would quickly learn to correct it. Thus, a comparable study on experienced L2 speakers displaying a high incidence of robust errors on a more essential grammatical feature is probably not possible. We certainly expect our findings to generalize to other “nonessential” features of L2 grammar (e.g., case marking) on which L2 speakers even with long experience are known to show persistent, “fossilized” errors. Whether or not subjective violations in more basic grammatical features, and in learners at earlier stages where even these basic features are still misrepresented, elicit a P600 is an open question that could be addressed by future research.

Summary and Conclusions

The present study is the first to look at the processing of subjective violations of gender agreement in L2 speakers while preserving the objective correctness of the experimental materials. The pattern of results suggests that L2 speakers use their idiosyncratic, subjective (and sometimes incorrect) gender representations to conduct a full, native-like syntactic sentence analysis only when they have reasons to mistrust the grammatical correctness of the input. Such mistrust can arise either due to the nature of the task (grammaticality judgments) or due to the presence of obvious, objective errors in the sentences as in Lemhöfer et al. (2014). We also found that the mere detection of subjective violations can occur without any associated ERP effects. This latter finding suggests that a lack of ERP effects in studies of L2 syntactic processing does not necessarily imply a complete insensitivity to the relevant syntactic feature.

References

- Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (Release 2) [CD-ROM]*. Philadelphia, PA: University of Pennsylvania, Linguistic Data Consortium.
- Bornkessel, I., & Schleuwsky, M. (2006). The extended argument dependency model: A neurocognitive approach to sentence comprehension across languages. *Psychological Review*, *113*, 787–821. <http://dx.doi.org/10.1037/0033-295X.113.4.787>
- Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about semantic illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, *1446*, 127–143. <http://dx.doi.org/10.1016/j.brainres.2012.01.055>
- 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. <http://dx.doi.org/10.1016/j.neubiorev.2015.01.010>
- Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics*, *27*, 3–42. <http://dx.doi.org/10.1017/S0142716406060024>
- Coulson, S., King, J. W., & Kutas, M. (1998). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and Cognitive Processes*, *13*, 21–58. <http://dx.doi.org/10.1080/016909698386582>
- 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. <http://dx.doi.org/10.1162/jocn.2008.21031>
- DeKeyser, R., Alfi-Shabtay, I., & Ravid, D. (2010). Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics*, *31*, 413–438. <http://dx.doi.org/10.1017/S0142716410000056>
- Dijkstra, T., Miwa, K., Brummelhuis, B., Sappelli, M., & Baayen, H. (2010). How cross-language similarity and task demands affect cognate recognition. *Journal of Memory and Language*, *62*, 284–301. <http://dx.doi.org/10.1016/j.jml.2009.12.003>
- Dowens, M. G., Guo, T., Guo, J., Barber, H., & Carreiras, M. (2011). Gender and number processing in Chinese learners of Spanish—evidence from event related potentials. *Neuropsychologia*, *49*, 1651–1659. <http://dx.doi.org/10.1016/j.neuropsychologia.2011.02.034>
- Dowens, M. G., Vergara, M., Barber, H. A., & Carreiras, M. (2010). Morphosyntactic processing in late second-language learners. *Journal of Cognitive Neuroscience*, *22*, 1870–1887. <http://dx.doi.org/10.1162/jocn.2009.21304>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. <http://dx.doi.org/10.3758/BRM.41.4.1149>
- Foucart, A., & Frenck-Mestre, C. (2011). Grammatical gender processing in L2: Electrophysiological evidence of the effect of L1–L2 syntactic similarity. *Bilingualism: Language and Cognition*, *14*, 379–399. <http://dx.doi.org/10.1017/S136672891000012X>
- Fricke, M., Zirnstein, M., Navarro-Torres, C., & Kroll, J. F. (2019). Bilingualism reveals fundamental variation in language processing. *Bilingualism: Language and Cognition*, *22*, 200–207. <http://dx.doi.org/10.1017/S1366728918000482>
- Friederici, A. D. (1995). The time course of syntactic activation during language processing: A model based on neuropsychological and neurophysiological data. *Brain and Language*, *50*, 259–281. <http://dx.doi.org/10.1006/brln.1995.1048>
- 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 of the United States of America*, *99*, 529–534. <http://dx.doi.org/10.1073/pnas.012611199>
- Gratton, G., Coles, M. G. H., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, *55*, 468–484. [http://dx.doi.org/10.1016/0013-4694\(83\)90135-9](http://dx.doi.org/10.1016/0013-4694(83)90135-9)
- Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language*, *36*, 3–15. [http://dx.doi.org/10.1016/0093-934X\(89\)90048-5](http://dx.doi.org/10.1016/0093-934X(89)90048-5)
- Gunter, T. C., & Friederici, A. D. (1999). Concerning the automaticity of syntactic processing. *Psychophysiology*, *36*, 126–137. <http://dx.doi.org/10.1017/S004857729997155X>
- Gunter, T. C., Friederici, A. D., & Schriefers, H. (2000). Syntactic gender and semantic expectancy: ERPs reveal early autonomy and late interaction. *Journal of Cognitive Neuroscience*, *12*, 556–568. <http://dx.doi.org/10.1162/089892900562336>
- Gunter, T. C., Stowe, L. A., & Mulder, G. (1997). When syntax meets semantics. *Psychophysiology*, *34*, 660–676. <http://dx.doi.org/10.1111/j.1469-8986.1997.tb02142.x>
- Hagoort, P., & Brown, C. M. (1999). Gender electrified: ERP evidence on the syntactic nature of gender processing. *Journal of Psycholinguistic Research*, *28*, 715–728. <http://dx.doi.org/10.1023/A:1023277213129>
- 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. <http://dx.doi.org/10.1080/01690969308407585>
- 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. <http://dx.doi.org/10.1017/S1366728901000232>

- Hahne, A., & Friederici, A. D. (2002). Differential task effects on semantic and syntactic processes as revealed by ERPs. *Cognitive Brain Research*, *13*, 339–356. [http://dx.doi.org/10.1016/S0926-6410\(01\)00127-6](http://dx.doi.org/10.1016/S0926-6410(01)00127-6)
- Hanulíková, A., van Alphen, P. M., van Goch, M. M., & Weber, A. (2012). When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience*, *24*, 878–887. http://dx.doi.org/10.1162/jocn_a_00103
- Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, *177*, 263–277. <http://dx.doi.org/10.1016/j.cognition.2018.04.007>
- Kaan, E., Harris, A., Gibson, E., & Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes*, *15*, 159–201. <http://dx.doi.org/10.1080/016909600386084>
- Kolk, H. H. J., Chwilla, D. J., van Herten, M., & Oor, P. J. W. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and Language*, *85*, 1–36. [http://dx.doi.org/10.1016/S0093-934X\(02\)00548-5](http://dx.doi.org/10.1016/S0093-934X(02)00548-5)
- Kotz, S. A. (2009). A critical review of ERP and fMRI evidence on L2 syntactic processing. *Brain and Language*, *109*, 68–74. <http://dx.doi.org/10.1016/j.bandl.2008.06.002>
- Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, *1146*, 23–49. <http://dx.doi.org/10.1016/j.brainres.2006.12.063>
- Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*, 325–343. <http://dx.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. <http://dx.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. http://dx.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. <http://dx.doi.org/10.1016/j.jml.2008.06.005>
- Lewis, A. G., Lemhöfer, K., Schoffelen, J.-M., & Schriefers, H. (2016). Gender agreement violations modulate beta oscillatory dynamics during sentence comprehension: A comparison of second language learners and native speakers. *Neuropsychologia*, *89*, 254–272. <http://dx.doi.org/10.1016/j.neuropsychologia.2016.06.031>
- Lewis, A. G., Wang, L., & Bastiaansen, M. (2015). Fast oscillatory dynamics during language comprehension: Unification versus maintenance and prediction? *Brain and Language*, *148*, 51–63. <http://dx.doi.org/10.1016/j.bandl.2015.01.003>
- McLaughlin, J., Osterhout, L., & Kim, A. (2004). Neural correlates of second-language word learning: Minimal instruction produces rapid change. *Nature Neuroscience*, *7*, 703–704. <http://dx.doi.org/10.1038/nn1264>
- Meulman, N., Stowe, L. A., Sprenger, S. A., Bresser, M., & Schmid, M. S. (2014). An ERP study on L2 syntax processing: When do learners fail? *Frontiers in Psychology*, *5*, 1072. <http://dx.doi.org/10.3389/fpsyg.2014.01072>
- Molinero, N., Barber, H. A., & Carreiras, M. (2011). Grammatical agreement processing in reading: ERP findings and future directions. *Cortex*, *47*, 908–930. <http://dx.doi.org/10.1016/j.cortex.2011.02.019>
- 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. <http://dx.doi.org/10.1162/0898929055002463>
- Münte, T. F., Matzke, M., & Johannes, S. (1997). Brain activity associated with syntactic incongruencies in words and pseudo-words. *Journal of Cognitive Neuroscience*, *9*, 318–329. <http://dx.doi.org/10.1162/jocn.1997.9.3.318>
- Ojima, S., Nakata, H., & Kakigi, R. (2005). An ERP study of second language learning after childhood: Effects of proficiency. *Journal of Cognitive Neuroscience*, *17*, 1212–1228. <http://dx.doi.org/10.1162/0898929055002436>
- Osterhout, L., & Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language*, *34*, 739–773. <http://dx.doi.org/10.1006/jmla.1995.1033>
- Patkowski, M. S. (1980). The sensitive period for the acquisition of syntax in a second language. *Language Learning*, *30*, 449–468. <http://dx.doi.org/10.1111/j.1467-1770.1980.tb00328.x>
- Prévost, P., & White, L. (2000). Missing surface inflection or impairment in second language acquisition? Evidence from tense and agreement. *Second Language Research*, *16*, 103–133. <http://dx.doi.org/10.1191/026765800677556046>
- Proverbio, A. M., Cok, B., & Zani, A. (2002). Electrophysiological measures of language processing in bilinguals. *Journal of Cognitive Neuroscience*, *14*, 994–1017. <http://dx.doi.org/10.1162/089892902320474463>
- Sabourin, L., & Stowe, L. A. (2008). Second language processing: When are first and second languages processed similarly? *Second Language Research*, *24*, 397–430. <http://dx.doi.org/10.1177/0267658308090186>
- Sabourin, L., Stowe, L. A., & de Haan, G. J. (2006). Transfer effects in learning a second language grammatical gender system. *Second Language Research*, *22*, 1–29. <http://dx.doi.org/10.1191/0267658306sr259oa>
- Sagarra, N., & Herschensohn, J. (2010). The role of proficiency and working memory in gender and number agreement processing in L1 and L2 Spanish. *Lingua*, *120*, 2022–2039. <http://dx.doi.org/10.1016/j.lingua.2010.02.004>
- Sassenhagen, J., Schlewesky, M., & Bornkessel-Schlewesky, I. (2014). The P600-as-P3 hypothesis revisited: Single-trial analyses reveal that the late EEG positivity following linguistically deviant material is reaction time aligned. *Brain and Language*, *137*, 29–39. <http://dx.doi.org/10.1016/j.bandl.2014.07.010>
- Steinhauer, K., White, E. J., & Drury, J. E. (2009). Temporal dynamics of late second language acquisition: Evidence from event-related brain potentials. *Second Language Research*, *25*, 13–41. <http://dx.doi.org/10.1177/0267658308098995>
- Tanner, D., Inoue, K., & Osterhout, L. (2014). Brain-based individual differences in online L2 grammatical comprehension. *Bilingualism: Language and Cognition*, *17*, 277–293. <http://dx.doi.org/10.1017/S1366728913000370>
- 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. <http://dx.doi.org/10.1017/S0272263105050102>
- van de Meerendonk, N., Kolk, H. H. J., Vissers, C. T. W. M., & Chwilla, D. J. (2010). Monitoring in language perception: Mild and strong conflicts elicit different ERP patterns. *Journal of Cognitive Neuroscience*, *22*, 67–82. <http://dx.doi.org/10.1162/jocn.2008.21170>
- van Herten, M., Chwilla, D. J., & Kolk, H. H. J. (2006). When heuristics clash with parsing routines: ERP evidence for conflict monitoring in sentence perception. *Journal of Cognitive Neuroscience*, *18*, 1181–1197. <http://dx.doi.org/10.1162/jocn.2006.18.7.1181>
- Vanhove, J. (2019). When labeling L2 users as nativelike or not, consider classification errors. *Second Language Research*. Advance online publication. <http://dx.doi.org/10.1177/0267658319827055>
- Verhees, M. W. F. T., Chwilla, D. J., Tromp, J., & Vissers, C. T. W. M. (2015). Contributions of emotional state and attention to the processing of syntactic agreement errors: Evidence from P600. *Frontiers in Psychology*, *6*, 388. <http://dx.doi.org/10.3389/fpsyg.2015.00388>

- Weiss, S., & Mueller, H. M. (2012). "Too many betas do not spoil the broth": The role of beta brain oscillations in language processing. *Frontiers in Psychology, 3*, 201. <http://dx.doi.org/10.3389/fpsyg.2012.00201>
- Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience, 16*, 1272–1288. <http://dx.doi.org/10.1162/0898929041920487>
- Williams, J. N., & Lovatt, P. (2005). Phonological memory and rule learning. *Language Learning, 55*, 177–233. <http://dx.doi.org/10.1111/j.0023-8333.2005.00298.x>
- Xue, J., Yang, J., Zhang, J., Qi, Z., Bai, C., & Qiu, Y. (2013). An ERP study on Chinese natives' second language syntactic grammaticalization. *Neuroscience Letters, 534*, 258–263. <http://dx.doi.org/10.1016/j.neulet.2012.11.045>

Received January 8, 2019

Revision received May 4, 2020

Accepted May 6, 2020 ■

E-Mail Notification of Your Latest Issue Online!

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at <https://my.apa.org/portal/alerts/> and you will be notified by e-mail when issues of interest to you become available!