Lexical Contributions to Inflectional Variability in L2
Predictive Processing

Holger Hopp

1. Inflectional variability in L2 gender

Grammatical gender is one of the most difficult aspects to acquire in a late-learned second language (L2), especially for learners whose first language does not have grammatical gender, e.g. English (Franceschina, 2005). There are two aspects of gender that present challenges to learners: First, learners need to learn the target gender of individual nouns, i.e. they need to assign a noun (e.g. Spanish casa – ‘house’) to a particular gender category (e.g. feminine). In this regard, L2 learners often assign a non-target gender to an item or they waver between different gender classes for a specific item, i.e. gender assignment is indeterminate. In short, L2 learners display lexical variability in gender assignment. Second, learners need to acquire morphosyntactic agreement properties of grammatical gender. Gender marking is (also) realized on lexical items that are grammatically dependent on a head noun, e.g. determiners and adjectives (e.g. la casa roja – ‘theFEM houseFEM redFEM’; Corbett, 1991). Adult L2 learners frequently omit gender agreement marking in production or they are not sensitive to gender violations in comprehension. In short, L2 learners display syntactic variability in gender agreement.

In this study, we explore whether lexical and syntactic variability in L2 gender are causally related. We focus on real-time processing, as inflectional variability is more pronounced in on-line production and comprehension than in untimed judgment tasks (e.g. Grüter, Lew-Williams & Fernald, 2012).

In principle, there are three different ways in which lexical and syntactic variability may be related in (predictive) L2 gender processing. First, they may constitute different problems, with syntactic variability representing non-target syntactic computation and lexical variability denoting non-target item-based lexical learning. This view is traditionally adopted in studies on syntactic variability in gender processing (e.g. Sabourin & Stowe, 2008). Methodologically, this means that studies only analyze the trials for which L2 learners had assigned the target genders. In essence, this approach then abstracts away from lexical variability in testing for syntactic variability with L2 gender.

*Holger Hopp, University of Braunschweig, h.hopp@tu-bs.de.

Syntactic variability is then studied as the problems with gender that remain once lexical variability is excluded from analysis.

Second, more recent approaches take into account that lexical variability may be relevant for syntactic variability. In an ERP study, Lemhöfer, Schriefers and Indefrey (2014) report that L1 German learners of Dutch do not show sensitivity to Dutch determiner-noun gender agreement violations when the objective/target gender system of Dutch is used as the baseline. However, when the analysis is adjusted for the subjective lexical gender assignment of the participants, the L2 learners display target-like neurophysiological gender agreement processing. For instance, if a learner incorrectly assigned a neuter Dutch word common gender in their L2 lexicon, then trials that would be an objective gender violation in Dutch (*de huis – ‘the\textsubscript{COM} house\textsubscript{NEUT}’) are actually grammatical trials according to the subjective gender assignment by the learner (de huis – ‘the\textsubscript{COM} house\textsubscript{COM(subjective)}’). Once the trials are adjusted for the subjective lexical gender assignment by the learners as assessed in an off-line gender assignment task, the L2 learners are more target-like in their processing of syntactic gender agreement. Syntactic variability is then studied as the gender problems that remain once lexical variability is adjusted for in the analysis.

Third, a current line of research explores the possibility that lexical variability partially causes syntactic variability in gender processing (e.g. Grüter et al., 2012; Hopp, 2013). It thus contrasts with the two first approaches that treat lexical variability as separate from syntactic variability in that lexical variability is excluded or adjusted for when studying syntactic variability. Instead, the third approach holds that lexical variability creates instability in gender representations that prevents target gender agreement processing.

In order to test this last approach, we study how lexical variability affects the predictive processing of grammatical gender agreement, i.e. how listeners construct a forward syntactic dependency relation between the gendered determiner and the noun (la\textsubscript{FEM} \rightarrow casa\textsubscript{FEM}) in comprehension.

2. Predictive processing of gender agreement and lexical variability

In sentence comprehension, establishing forward, i.e. predictive, agreement relations between the determiner and the noun on the basis of gender marking on determiners can help narrow down the set of potential nouns that could follow, thus facilitating comprehension (e.g. Lew-Williams & Fernald, 2010). In visual-world eye tracking studies, participants view displays containing different objects while they listen to simple instructions that refer to the objects (Huettig, Rommers, & Meyer, 2011). Native speakers use gender on determiners to pre-activate noun labels of the objects that are presented, and they will show earlier gaze shifts to the target pictures when gender information on the determiner allows them to anticipate which noun will follow (e.g. Dahan et al. 2000; Lew-Williams & Fernald, 2010). In contrast, L1 English adult L2 learners do not
benefit from informative gender marking on determiners to predict the following noun (Grütter et al., 2012; Lew-Williams & Fernald, 2010; though see Dussias, Valdés Kroff, Guzzardo Tamargo, & Gerfen, 2013). Grütter et al. (2012) found that even near-native (L1 English) learners of Spanish did not demonstrate anticipatory processing of gender agreement for highly familiar Spanish nouns. Of note, the learners also showed variable lexical gender assignment.

Building on this finding, Hopp (2013) directly tested for contingencies between variability in lexical gender representations and predictive processing in adult L1 English learners of German. Lexical gender assignment accuracy was operationalized as the accuracy by which the correct gender form could be supplied in spoken production. For the learners, correct gender assignment in production significantly correlated with predictive gender agreement processing in comprehension. Hopp accounted for the link between lexical and syntactic variability in terms of recent psycholinguistic approaches to prediction (e.g. Dell & Chang, 2014; Phillips & Ehrenhofer, 2015). By making predictions based on observations of the probability of linguistic events in the input, listeners can test hypotheses about language against the input and thus acquire or adjust knowledge of the target language. If a prediction is not met in the input, prediction error allows for implicit learning by adaptation to the properties actually observed in the input (Kuperberg & Jaeger, 2015).

In language processing, listeners rapidly adapt to prediction error in that prediction is adjusted, attenuated or even abandoned (Fine, Jaeger, Farmer, & Qian, 2013). For instance, native readers adjust their parsing preferences according to the probabilities experienced in the experimental environment (e.g. Fine et al., 2013; Kleinschmidt & Jaeger, 2015). For gender agreement, Hopp (2013) argued that the relation between knowledge of gender assignment and predictive use of gender agreement follows from similar implicit learning on the basis of prediction error. If gender assignment is variable or non-target-like in the L2 lexicon, then L2 learners will encounter frequent mismatches between their (subjective) gender assignment and the actual gender of nouns in the input. Specifically, lexical variability paves the way to frequent prediction errors, since the noun predicted to occur by the listener does not occur in the (target) input. As a consequence, the parser likely adjusts prediction strength according to error-based implicit learning, such that gender agreement will not be used predictively by L2 learners due to variable lexical gender assignment.

3. Experiment 1

In Experiment 1, we directly test this approach by manipulating lexical variability in gender assignment in that we provide instruction on German gender to intermediate L1 English learners. We test whether training on lexical gender leads to predictive processing of gender agreement. Experiment 1 is a training study in a pre-posttest design of production and comprehension accuracy on gender in L2 German. In a visual-world eye tracking study, we investigate whether participants use a gendered determiner to predict the
subsequent noun. The pretest and the posttest were modelled on Hopp (2013) and included (i) a picture description task in order to probe lexical gender assignment in production and (ii) a visual-world eye tracking task to investigate the predictive processing of gender agreement.

3.1. Participants

Thirty-four late (L1 English) L2 learners of German took part in the experiment (M = 21.2 years, SD = 2.2; 18 females). All participants were students in German programs at US universities. They had all started acquiring German after age 10 (M = 13.9 years, SD = 2.5). They completed a standardized 30-item written placement test (Goethe Institute, 2010), scoring 17.7 points on average (range: 11-28), which placed them into the upper-intermediate bracket. Mean length of exposure was 7.3 years (SD = 2.7), and mean length of residence in a German-speaking country was 0.7 years (SD = 1.0)

3.2. Pretest and Posttest: Materials

We constructed stimuli containing a gender-marking definite determiner, a (gender-ambiguous) adjective and a noun.

(1) Wo ist der/die/das gelbe [Noun]?
   Where is theMASC/FEM/NEUT yellow [Noun]?

Twenty items in total were created for the gender condition in (1), i.e. five difference trials for each of the three gender forms, i.e. masculine, feminine and neuter, in (1), and five same trials. For each item, four-picture displays using coloured drawings or reduced photographs of picturable and easily identifiable inanimate objects were designed (Figure 1). All objects were highly frequent.

Fig. 1. Display.

For the difference trials, one object with a clearly identifiable colour was the target; there were two identically coloured competitor objects that bore one
of the other two genders each, and the fourth object in the display was a differently coloured distracter of a gender different from the target. For the same trials, five ambiguous displays were designed along the same lines as Figure 1; however, the three colour-matched objects all had the same gender and the differently coloured distractor had a different gender. In addition, ten filler trials were added in which the differently coloured object was the target or in which a number adjective clearly identified the target. The resulting 30 displays were assigned to one of three lists which counter-balanced target noun and object position for the experimental items. Trials were presented in pseudorandomized order. In all, 80 nouns were used.

All sentences were recorded by a male native speaker of High German at a slow-to-moderate pace. Determiners had a mean length of 392 ms (range: 388-411ms), adjectives of 621 ms (range: 566-705ms) and nouns of 592 ms (range: 535-763ms). The onset of the adjectives and the nouns was aligned to occur 650 and 1750 ms after the onset of the determiner respectively. Participants were informed that the experiment involved word learning in German. They were debriefed about the real nature of the experiment only after the posttest. Participants were tested individually. In the first session, the participants completed the proficiency test and a language history questionnaire. In addition, they took part in the pretest. For the second session with the training and posttest, they returned to the lab about one week later (M = 7.4 days, SD = 0.9).

3.3. Pretest and Posttest: Procedure

Participants sat in front of a 19-inch screen at a distance of approximately 70 cm. Participants read instructions, and they completed two practice trials before a calibration began. The calibration aimed for visual acuity below 0.5 degrees and was repeated in the course of the experiment if necessary. In the experiment, an SMI RED 500 eye tracker recorded gaze position at 60 Hz.

A) Production task: In the production task that immediately preceded the comprehension task, participants saw still images of the four-picture displays and were asked to name the four objects in each display, including their colour (see Figure 1). This way, we could (i) ensure that participants recognized the objects and knew their labels and (ii) assess which gender L2 participants assigned to the labels of the objects. In their descriptions of the objects, the participants used noun phrases containing determiners, colour adjectives and nouns. In total, participants named 80 different objects in the experimental condition. Participants’ responses were audio-recorded and transcribed. Responses were coded for gender due to the gender form used on the determiner (der, die, das) for definite NPs as in (1) or, if the determiner was indefinite, the adjective (e.g. grün-er, -e, -es – ‘greenMASC/FEM/NEUT’) for indefinite NPs. Overall accuracy was coded as percentage of target responses.

B) Comprehension task: After the production part, the screen changed immediately to show the same displays including a fixation cross. After a preview of 3,000 ms, a sound signal alerted participants to fixate on the cross.
The participants’ gaze was directed to the fixation cross prior to each trial. The auditory presentation of the sentences began 1,500 ms after the sound signal. In the analysis, gaze position after determiner onset was coded every 20 ms for a time-window of 3,000 ms. Items with looks to the target region within the first 200 ms after determiner onset were discarded because they could not reflect linguistically guided gaze shifts (e.g. Huettig et al., 2011). In all, this affected less than 1% of the data. In total, the experiment comprising the two tasks took about 20 minutes. Participants received a different list of the experiment in the pretest and the posttest, such that they would not encounter the same target nouns or see the target objects in the same position again.

3.4. Training experiment: Materials and procedure

The training experiment comprised a computer-based training and a testing phase: In the training phase, all 80 nouns used in the pretest were presented to the participants. The 80 nouns were assigned to two blocks of 30 items each for the nouns of the experimental trials and one block of 20 items for the nouns in the filler condition. In each block, a picture for noun was presented to the participants for a duration of 4,000 ms each, with the determiner-noun sequence (e.g. der Hut – theMASC hat) written underneath the picture. In addition, an audio-stimulus was played containing the determiner-noun sequence. Participants then repeated each noun phrase out loud. In each block, each noun was presented three times, order was randomized, and no feedback was given.

In the testing phase that immediately followed each training block, the objects were presented visually without any auditory or written information. Participants then had to say the determiner-noun sequence out loud. The experimenter logged their responses on the keyboard. In cases when they had produced a non-target determiner, the participants received negative feedback (“no”), and the item was tagged to reappear at the end of the block. The participants did not receive any correction. Items for which the participants had produced non-target determiners were presented until the participants produced the correct determiner. Hence, all participants reached 100% accuracy on grammatical gender assignment in production at the end of the testing blocks. In all, the training experiment took about 20 minutes for the participants. The posttest immediately followed.

3.5. Results

3.5.1. Production task

In order to test the extent to which the participants learned noun labels and gender assignment in the training, we first consider the production accuracy by the participants in the pretest and the posttest. In the pretest (Table 1), there was large variability in accuracy of picture naming. In contrast, errors or failures in picture naming were very low in the posttest, which indicates that participants had successfully learned the nouns during the training. For gender accuracy, we
observe variability between the participants in both the pretest and the posttest. A mixed logistic regression on the number of gender mistakes in the pretest and the posttest also yielded a highly significant effect (β = 1.440, SE = 0.071, t = 20.342, p < .001), which shows that the training improved performance on gender assignment. However, variability remained in the posttest demonstrating that the participants could retain accuracy in gender assignment from the training and testing blocks to different degrees.

Table 1. Accuracy in production in pretest and posttest (n = 34).

<table>
<thead>
<tr>
<th></th>
<th>Pretest</th>
<th></th>
<th>Posttest</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Range</td>
<td>Mean</td>
<td>Range</td>
</tr>
<tr>
<td>Noun naming (80 max.)</td>
<td>70.68</td>
<td>57-80</td>
<td>79.82</td>
<td>78-80</td>
</tr>
<tr>
<td>Gender assignment</td>
<td>67.2%</td>
<td>46%-93%</td>
<td>86.1%</td>
<td>36%-100%</td>
</tr>
</tbody>
</table>

3.5.2. Comprehension task

In the comprehension task, we tested the extent to which training on lexical gender affects predictive use of gender agreement. We analysed the mean reaction time, i.e. the first fixation on the target picture after determiner onset. Following Hopp (2013) and Lemhöfer et al. (2014), we computed reaction times for the subset of trials based on the production accuracy of gender assignment by the L2ers, i.e. the subjective gender assignment of the participants. Specifically, we categorized the comprehension trials of the L2ers into (i) difference trials, if the L2er had assigned the target-language lexical gender to the target item and other (target) genders to the other items. Hence, these were trials in which gender marking on determiners would be informative as to the target item according to the learner’s subjective gender assignment. For all experimental trials, the difference trials constituted 159 out of 496 difference trials (32%) in the pretest and 375 out of 500 difference trials (75%) in the posttest. Further, we defined (ii) same trials as trials in which the participant had assigned the target item and other items the same gender. These amounted to 91 out of 170 same trials (54%) in the pretest and to 146 out of 170 same trials (86%) in the posttest. In the same trials, gender marking could not act as an unambiguous cue for the target object. All other trials were excluded.

Figure 2 shows the mean RTs in each trial type for the pretest and the posttest. The mean difference between difference and same trials is 88ms in the pretest and 346ms in the posttest. We analysed the mean reaction time of looking towards the target region in a mixed linear regression model with the fixed factors Type (same versus difference) and Test (pretest versus posttest). Subject and item were crossed random factors. There were no significant main effects of Type (β = 216.28, SE = 152.39, t = 1.419, p = .156) or Test (β = 193.22, SE = 151.33, t = 1.277, p = .202), yet a highly significant interaction of Type and Test (β = 305.82, SE = 87.68, t = 3.488, p = .001). In pairwise comparisons for pretest and posttest, the learners showed no significant
difference between *difference* and *same* trials in reaction times in the pretest ($\beta = 72.16$, SE = 68.29, $t = 1.057$, $p = .294$); yet, there was a highly significant difference in the posttest ($\beta = 411.92$, SE = 59.53, $t = 6.919$, $p < .001$).

**Fig. 2:** Mean reaction times (in ms) in target region by Type (*same* vs *difference*) & Test (pre vs post). All participants (n = 34). 0ms = det. onset; 1750ms = noun onset (error bars = SE).

In order to test the relation between lexical and syntactic variability, we probed whether the difference between *same* and *difference* trials was affected by accuracy in lexical gender assignment. For the pretest, a mixed linear regression analysis with Type and lexical gender assignment accuracy in the production test as fixed factors and subject and item as crossed random factors yielded no effects of Type, Gender Accuracy or any interaction (all $t$s < 1). For the posttest, however, the same analysis returned significant main effects of Type ($\beta = 608.65$, SE = 75.49, $t = 8.063$, $p < .001$), Gender Accuracy ($\beta = 29.49$, SE = 10.41, $t = 2.832$, $p = .005$), and a significant interaction of Type and Gender Accuracy ($\beta = 22.79$, SE = 5.44, $t = 4.190$, $p < .001$). To identify the nature of the interaction, correlational analyses were performed between the mean predictive gender effect (i.e. the mean difference between *same* and *difference* trials) and the lexical gender assignment accuracy in the production test. In the posttest, the correlation between gender assignment and the predictive use of gender was highly significant ($r = .534$, $p = .001$), i.e. greater accuracy in lexical gender assignment led to larger prediction in processing gender agreement.

To illustrate the significant association of the predictive effect in grammatical gender agreement, we decided to separate the L2 participants into two near-equally sized groups on the basis of the overall accuracy in gender assignment in production in the posttest following the procedure in Hopp (2013): A group with variable lexical gender (n = 18), i.e. variable or non-target-
like gender assignment, that made at least seven gender mistakes (9%) for the 80 objects (mean accuracy on gender = 76%), and a group with consistent lexical gender (n = 16), i.e. close-to-target gender assignment, that made at most six (7.5%) gender mistakes (mean accuracy = 97.5%). One-way ANOVAs were run to check whether the groups differed on any factors other than gender assignment accuracy in the posttest. The groups did not show any significant differences, except for proficiency, with the gender-consistent group scoring somewhat higher (19.5 vs. 16.2; F(1,32) = 5.614, p = .024).

Figure 3 shows the mean reaction times in the pretest and in the posttest.

---

**Fig. 3:** Mean reaction times (in ms) in target region by Type (*Same* vs *Difference*) and Test (Pre vs Post). Top: Gender-variable group (n = 18). Bottom: Gender-consistent group (n = 16). 0 ms is determiner onset (error bars = SE).
For the comprehension data, a linear mixed analysis with Test, Type and Group as fixed factors and subject and item as crossed random factors returned a significant three-way interaction between Test, Type and Group ($\beta = 396.65$, SE $= 183.20$, $t = 2.165$, $p = .031$). Subsequent pairwise comparisons of the mean reaction times by group showed no significant difference between difference and same trials in the pretest for either group (all $t$s $< 1$). For the posttest, the effect of Type was not significant for the gender-variable group ($\beta = 121.25$, SE $= 76.40$, $t = 1.587$, $p = .116$). In contrast, the gender-consistent group demonstrated a highly significant effect of Type ($\beta = 604.79$, SE $= 79.36$, $t = 7.621$, $p < .001$). In sum, the gender-variable group did not benefit from the training on lexical gender with respect to the predictive use of grammatical gender. Despite a numerical advantage for the difference versus the same trials, this advantage did not increase compared to the pretest. By contrast, the gender-consistent group came to use gender as a predictive cue in the posttest.

3.6. Experiment 1: Discussion

Experiment 1 showed that massed exposure to grammatical gender in form-focused training activities leads to low-intermediate-to-advanced (L1 English) learners attaining on-line sensitivity in the predictive processing of grammatical agreement in German NPs even after short periods of training.

We investigated if differences in lexical gender assignment affect the predictive processing of gender. In the pretest, we did not find any relation between lexical gender assignment and predictive processing of grammatical gender. In contrast, the posttest revealed that the predictive use of gender was strongly related to the amount of target-like assignment of gender in production. In other words, changes in lexical gender assignment, i.e. the degree to which gender assignment in the training session had been acquired or reinforced affected the amount of gender prediction. These findings support the hypothesis that lexical and syntactic variability in gender are related. Specifically, the findings align with research about adaptivity in implicit learning according to prediction success and error (Fine et al., 2013). L2 learners who successfully learned or strengthened the target gender assignment for the nouns used in the experiment in the training session used gender as a predictive cue in agreement processing because the cue reliably leads to facilitative processing of the upcoming noun. In contrast, learners with more variable gender assignment do not come to use gender for prediction, even for those (many) trials in which it would lead to facilitatory prediction. If these learners consistently used their variable and non-target gender assignment for prediction, they would make erroneous predictions that entail costly reanalysis for the parser (Hopp, 2013). In terms of parsing efficiency, error-based implicit learning then serves to attenuate gender-based prediction in order to reduce the likelihood of future prediction error. As a consequence, the learners who continue to exhibit variable gender assignment fail to use gender as a predictive agreement cue overall. From a psycholinguistic perspective, then, the finding that variability in gender
assignment in production is associated with variability in gender agreement in predictive processing is a consequence of implicit learning based on prediction success and error.

4. Experiment 2

In Experiment 2, we probe if the correlation between gender assignment and gender agreement processing observed in Experiment 1 can be replicated in a different group of participants. It may be that the contingencies between gender assignment and agreement found in Experiment 1 reflect differences among the L2 learners in proficiency or other factors unrelated to gender assignment. To address the role of such potential confounds, Experiment 2 tests whether native speakers of German show a similar relation between lexical gender assignment accuracy and the predictive use of gender agreement. Unlike L2 learners, mature native speakers (i) are highly proficient in the target language and (ii) have fully target-like, stable and readily accessible gender representations in their lexicons (dialectal variation aside), and (iii) use gender for predictive agreement processing. In Experiment 2, we manipulate the input by introducing gender assignment errors in a visual-world eye-tracking task. The introduction of non-target gender assignment in the input would lead to prediction error provided native speakers use their (target) gender knowledge for anticipation. In this regard, Experiment 2 simulates the situation of the L2 group with variable lexical gender in Experiment 1. While gender assignment errors are a property of the (internal) lexicon of L2 speakers, gender assignment errors are introduced in the (external) input to native speakers in Experiment 2. Both types of gender errors could lead to prediction error. If prediction error drives the (non-)use of gender for agreement processing, we would expect it to engender attenuation in predictive gender processing also in native speakers.

Experiment 2 employed a between-subjects blocked design. Participants were assigned to one of two groups, the No-Error and the Error group. In Experiment 2, we test whether the introduction of lexical variability in gender assignment leads to attenuation of predictive gender processing in natives.

4.1. Participants, Materials and Procedure

We recruited 42 native speakers of German (M = 22.6 years, SD = 3.2; 32 females) for the experiment. All participants were students at a German university at the time of testing and received 5 Euros.

For Experiment 2, all items from Experiment 1 were used. To demonstrate that participants in both groups used gender for prediction before encountering some trials containing gender errors, we added more items to Experiment 2. In total, we constructed 30 difference items and 30 same items as in Experiment 1. In addition, 30 items were added in which the differently coloured object was the target (filler condition). In this condition, the adjective was an unambiguous lexical cue to the referent. The items were assigned to two blocks, comprising
45 items each (15 of each type). The 15 difference items from Experiment 1 were assigned to the second block, along with the five same items from Experiment 1, 10 additional same items and 15 filler items. Thus we ensured we could compare predictive processing for the same set of items in Experiment 1 and the second block of Experiment 2. In all, 160 nouns were used.

We created three lists which counter-balanced target noun and object position for the experimental items. Each of the lists was shown in two conditions to two participant groups: a No-error condition, in which the gender assignment of the objects named was always target-like, and an Error condition. In the Error condition, the target objects of 11 filler items in the second block were assigned a non-target gender, i.e. participants would hear 11 incorrectly gender marked determiner-noun combinations (e.g. das blaue Waage – the NEUT blue scales FEM). The proportion of non-target gender assignment (11/45) roughly corresponded to the mean proportion of non-target gender assignment among the variable-gender group in the posttest in Experiment 1 (76%). Importantly, none of the experimental trials in the Error-condition contained gender mistakes, since non-target gender assignment was restricted to filler items.

The stimuli were recorded by the same male native speaker of German as in Experiment 1. Determiners had a mean length of 198 ms (SD = 28 ms), adjectives of 606 ms (SD = 53 ms) and nouns of 681 ms (SD = 139 ms). Participants were tested with the same eye tracker as in Experiment 1.

4.2. Analysis, Results and Discussion

The data from one participant in the Error-group had to be excluded due to tracker loss, which left 41 participants for analysis (18 in the No-error and 23 in the Error group). Reaction times, i.e. the first fixation on the target picture after determiner onset, were computed as in Experiment 1. Since the gender mistakes accrued in the second block, we decided to break down the results by quartiles.

Figure 4 shows the predictive gender effect in the two groups across the quartiles. Figure 4 shows how the predictive gender effect increases in both groups over time, yet declines in the Error-group in the final quartile. Given our specific hypothesis about possible effects of Group in the last time window, we computed separate analyses for each quartile. In Quartile 4, the interaction between Type and Group became marginally significant ($\beta = 122.13$, SE = 72.84, $t = 1.677$, $p = .094$). Post-hoc pairwise comparisons in Quartile 4 show that the No Error group displayed a highly significant effect of Type ($\beta = 188.28$, SE = 57.66, $t = 3.265$, $p = .001$), whereas the effect was not significant for the Error Group ($\beta = 66.15$, SE = 45.99, $t = 1.438$, $p = .151$).

These findings show that the inclusion of gender assignment errors in the filler items in the Error Group affected the predictive gender agreement in the experimental trials. Experiment 2 showed that native speakers robustly use gender for predictive agreement. At the same time, the experiment supplies clear evidence that introducing non-target-like gender assignment in unrelated filler trials in the Error group engendered a sizeable attenuation of the predictive use
of gender agreement in native processing in the final quartile. The use of gender in predictive processing is attenuated after only a few gender mistakes.

Fig. 4: Development of predictive gender effect (in ms, incl. standard error) across quartiles. No Error group (n = 18) and Error group (n = 23).

5. General Discussion

In two experiments, we investigated the relation between lexical gender assignment and the predictive processing of gender agreement in non-native and native German. In Experiment 1, we found (i) that intermediate late (L1 English) learners of German can come to show predictive processing of gender agreement after training on lexical gender assignment and (ii) that the accuracy in gender assignment moderates predictive gender agreement.

L2 learners who acquire overall target lexical gender for nouns (used in the experiment) come to exploit their target lexical knowledge for predictive agreement processing, because predictively using gender reliably facilitates comprehension. In contrast, if gender assignment remains non-target-like in a sizeable part of the lexicon, L2 learners do not use gender for predictive processing. Using gender assignment for gender agreement prediction would incur prediction errors for the items with non-target subjective gender assignment. An adaptive parser would then adjust prediction according to experience (see also Fine et al., 2013). For the learners with variable gender assignment in Experiment 1, such adaptation would lead to the non-use of (subjective) gender as a predictive cue.

However, the correlation between lexical and syntactic variability observed in Experiment 1 does not imply a causal link. In Experiment 2, we therefore tested a group of native speakers of German, whose lexical and syntactic
representations of gender are target-like and who make robust use of gender agreement in predictive processing. In the between-subjects blocked design of Experiment 2, native speakers stopped using gender predictively once gender had become unstable in irrelevant parts of the experiment, i.e. the filler trials. Listeners appeared to adapt automatic processing routines based on the computations of probability estimates of gender reliability in the overall input.

Experiment 2 suggest that L2-like performance on predictive gender agreement can be emulated in native processing when the input conditions in the experiment resemble the structure of gender representations in the L2 lexicon: Specifically, (subjective) gender becomes an unreliable cue because gender assignment is not always target-like. Such a cue has decreased predictive value in language processing since the forward transitional probability of the subjective, but non-target, gender representation in the learner corresponds to zero in the input. Under such circumstances, native speaker performance parallels L2 behaviour in that prediction according to gender is suspended.

These findings are in line with the lexical gender learning hypothesis by Grüter et al. (2012) and Hopp (2013). The lexical gender learning hypothesis holds that gender representations are less stable in L2 adults because they create weaker links between nouns and gender nodes (e.g. Gollan, Montoya, Cera, & Sandoval, 2008). In consequence, gender assignment remains partially erroneous or variable, which, in turn, prevents gender from being used as a predictive cue in agreement processing. This account captures the data on gender assignment and agreement processing in Grüter et al. (2012), Hopp (2013), and the findings in this study for L2 learners and for native speakers.

In sum, the present experiments suggest that lexical variability and syntactic variability in L2 gender processing are causally related. Previous approaches that excluded or controlled for lexical variability when studying syntactic variability in L2 gender may have unwittingly set aside a major aspect that determines problems with gender agreement in adult L2 acquisition.

References


