Bilingualism: Language and Cognition 18 (3), 2015, 502–523 © Cambridge University Press 2014 doi:10.1017/S1366728914000534

Lexical constraints in second language learning: Evidence on grammatical gender in German*

SUSAN C. BOBB
Department of Psychology, The Pennsylvania State University
JUDITH F. KROLL
Department of Psychology, The Pennsylvania State University
CARRIE N. JACKSON
Department of Germanic and Slavic Languages and Literatures, The Pennsylvania State University

(Received: November 8, 2013; final revision received: August 8, 2014; accepted: August 11, 2014; first published online 23 October 2014)

The present study asked whether or not the apparent insensitivity of second language (L2) learners to grammatical gender violations reflects an inability to use grammatical information during L2 lexical processing. Native German speakers and English speakers with intermediate to advanced L2 proficiency in German performed a translation-recognition task. On critical trials, an incorrect translation was presented that either matched or mismatched the grammatical gender of the correct translation. Results show interference for native German speakers in conditions in which the incorrect translation matched the gender of the correct translation. Native English speakers, regardless of German proficiency, were insensitive to the gender mismatch. In contrast, these same participants were correctly able to assign gender to critical items. These findings suggest a dissociation between explicit knowledge and the ability to use that information under speeded processing conditions and demonstrate the difficulty of L2 gender processing at the lexical level.

Keywords: grammatical gender, second language acquisition, proficiency

In recent years, researchers from different theoretical perspectives have addressed second language (L2) grammatical gender acquisition. At issue is whether the well-documented difficulties that adult L2 learners show, particularly with syntactic gender agreement, are due to an incomplete gender representation in the mental lexicon or to processing difficulties. Several second language acquisition (SLA) theories incorporate the possibility that native language (L1) syntactic structures/features can be transferred into the L2 (e.g., Schwartz & Sprouse, 1996; MacWhinney, 1997). Crucially, transfer is not possible when gender does not exist in the L1, potentially posing a particular challenge to ultimate attainment. Several theories address how L2 learners may acquire or fail to acquire new grammatical features (e.g., FAILED FUNCTIONAL FEATURES HYPOTHESIS, Hawkins & Franceschina, 2004; MISSING SURFACE INFLECTION HYPOTHESIS, Prévost & White, 2000; FULL TRANSFER/FULL ACCESS HYPOTHESIS, Schwartz & Sprouse, 1996; see also DeKeyser, 2000; Ullman, 2001; Paradis, 2004). Most of the studies investigating these theories, however, have focused on sentence-level processing. In the present study, we ask whether or not difficulties in learning L2 features are revealed when processing individual noun phrases, in the absence of the computational demands required to process an entire sentence.

Difficulties with grammatical gender

The research to date is mixed as to whether native speakers of a language without grammatical gender can acquire gender in the L2 after a critical period. Some studies argue that gender agreement difficulties in L2 production are due to performance pressures that obscure an underlying representation that is intact (e.g., Prévost & White, 2000). Other studies have found continuing gender difficulties in L2 comprehension as well and have argued for representational deficits (e.g., McCarthy, 2008). The results of L2 online studies using ERPs and self-paced reading are also inconclusive, with some suggesting that late L2 learners are unable to process grammatical gender in a native-like manner unless gender systems are similar in L1 and L2 (Sabourin & Stowe, 2008). Other studies, however, suggest that even learners whose languages do not share gender features across languages can achieve native-like processing, at least at high levels of L2 proficiency (Gillon Dowens, Vergara, Barber &

* This research was supported in part by NSF BCS-0720155 to Susan C. Bobb and Judith F. Kroll. The writing of this article was also supported in part by NIH Grant HD053146 and NSF Grants BCS-0955090 and OISE-0968369 to Judith F. Kroll, NIH Grant HD058765 to Carrie N. Jackson, and the German Excellence Initiative to Susan C. Bobb. We thank the audiences of the 7th International Symposium on Bilingualism and Psychonomics 2008 as well as Victor Kuperman, Natasha Tokowicz, Theres Grüter, Amber MacMillan, Sarah Chabal, and Jared Linck. We would also like to thank Angélique Blackburn, Mark Connelly, and Shara Gress for assistance with data collection.

Address for correspondence:
Department of Communication Sciences and Disorders, Northwestern University, 2240 Campus Drive, Evanston, IL 60208, USA
scb207@gmail.com

Carreiras, 2010; Foucart & Frenck-Mestre, 2012). But there is also evidence that even beginning L2 learners can show sensitivity to grammatical gender (Tokowicz & MacWhinney, 2005; Sagarra & Herschensohn, 2010).

The role of syntactic complexity

Crucially, the majority of studies showing L2 gender agreement difficulties do so in sentence contexts, where multiple gender cues come into play: gender is frequently redundantly marked on determiners, adjectives, and personal pronouns as part of gender agreement regularities. Additional syntactic information, such as number agreement, also competes for memory resources during processing. It is therefore possible that processing demands may obscure L2 learners' underlying knowledge (Montrul, Foote & Perpiñán, 2008; Tokowicz & Warren, 2010) and syntactic complexities may contribute to difficulties in L2 gender processing. For example, Keating (2009) found similar performance for native Spanish speakers and advanced L2 learners for violations within the determiner phrase. When violations occurred in nonlocal contexts, however, L2 learners no longer showed sensitivity to agreement violations (see also Gillon Dowens et al., 2010; Foucart & Frenck-Mestre, 2012). These studies suggest that lexical-level experiments, which we define as experiments where participants are only required to process a single noun phrase (e.g., Salamoura & Williams, 2007), may provide a window into constraints on L2 grammatical acquisition, and gender agreement in particular.

Gender Difficulties at the Lexical Level

In contrast to the body of L1 literature demonstrating sensitivity to grammatical gender mismatches outside of sentential contexts (e.g., Bates, Devescovi, Hernandez & Pizzamiglio, 1996; La Heij, Mak, Sander & Willeboordse, 1998), only a few studies have investigated L2 gender at the lexical level (e.g., Guillelmon & Grosjean, 2001; Carroll, 2005; Grüter, Lew-Williams & Fernald, 2012; Hopp, 2013). Most of these studies tested explicit gender assignment (but see Grüter et al., 2012; Scherag, Demuth, Rösler, Neville & Röder, 2004), finding that even learners who do not have grammatical gender in their L1 appear to "get it" when explicitly asked to identify the gender of nouns (e.g., Guillelmon & Grosjean, 2001). But correct gender identification does not necessarily mean that L2 learners implicitly process L2 gender. The typical gender identification task is unspeeded, allowing L2 learners to use mnemonic strategies to assign the correct gender to a noun (e.g., Carroll, 1989). Hawkins and Franceschina (2004) suggest that L2 learners use phonological regularities to correctly identify L2 gender (see also Bordag, Opitz & Pechmann, 2006). Similarly, Carroll (2001) argues that native English speakers learning German as an L2 do not acquire the inherent gender feature on the noun but instead have an "emergent gender" – storing correct word strings in memory as "ready-made expressions to be activated when needed" (Carroll, 2001, p. 361). It is therefore unclear whether L2 learners automatically process gender during lexical access, even though they can identify the gender of nouns.

The present study asks whether there are lexical contexts in which L2 learners do show gender sensitivity. Particularly if the lack of sensitivity to grammatical gender violations in previous L2 research stems from increased processing demands, then L2 learners may be more likely to exhibit sensitivity to grammatical gender in a lexical-level task, in which processing demands are arguably reduced relative to a sentence-level or productive task. On the other hand, it is also possible that lexical-level processing issues may underlie the syntactic agreement difficulties that L2 learners show (e.g., Grüter et al., 2012), such that L2 learners would demonstrate continued difficulties with grammatical gender processing.

Our second goal is to determine whether and how L2 proficiency modulates sensitivity to gender. With increasing L2 proficiency, can L2 learners become increasingly sensitive to gender marking at the lexical level? According to the REVISED HIERARCHICAL MODEL (Kroll & Stewart, 1994), with increasing proficiency, the L1 lexicon decreasingly mediates the L2. We hypothesize that this decrease in reliance on the L1 lexicon and direct access to conceptual information in the L2 may coincide with the forming of a more stable L2 gender representation (e.g., Paolieri, Cubelli, Macizo, Bajo, Lotto & Job, 2010).

Finally, we investigate how grammatical gender becomes associated with specific lexical items (e.g., Carroll, 1989; Arnon & Ramscar, 2012). Phonological cues to gender, such as predictable noun endings, may support gender learning by linking a specific ending to a particular article (Bordag et al., 2006; Hawkins, 2009). If late acquisition of grammatical gender is possible, can L2 learners use phonological cues to acquire gender assignment of specific lexical items? In subsequent analyses, we return to this issue.

The present study

In the current study, we tested native English speakers who learned German as an L2. The German gender system comprises three genders, masculine (*der*), feminine (*die*), and neuter (*das*). Gender is marked on articles (both definite and indefinite) and on attributive adjectives, and the link between a noun and its gender appears relatively arbitrary. Forks are feminine (*die Gabel*), knives are neuter

https://doi.org/10.1017/S1366728914000534 Published online by Cambridge University Press

Table 1. *Example distribution of items across three conditions for the critical English words ROPE (das Seil), TENT (das Zelt), and GOAL (das Ziel).*

	Condition		
	1 (German is grammatical) translation mismatch (TM)	2 (German is ungrammatical) gender mismatch (GM)	3 (German is grammatical) translation & gender mismatch (TGM)
Participant 1:	THE ROPE - DAS MEER	THE GOAL - DER ZIEL	THE TENT - DER MORD
Participant 2:	THE TENT - DAS MEER	THE ROPE - DER SEIL	THE GOAL - DER MORD
Participant 3:	THE GOAL - DAS MEER	THE TENT - DER ZELT	THE ROPE - DER MORD

Note. We divided English translations into triads and item-matched them on gender, frequency and length. We assigned the same incorrect translations (e.g., DAS MEER (the ocean), DER MORD (the murder)) to these three items for conditions 1 and 3. Condition 2 ensured participants paid attention to gender because simply matching noun translations would lead to incorrect "yes" judgments. Across three participants, all items and all conditions were seen, and each participant saw all the English words, but would only see a given incorrect German translation form once. By virtue of the triads across participants sharing two of the three German incorrect translation forms, each participant saw almost all the same German incorrect translation forms, allowing us to control for lexical familiarity. In Condition 2, the presented German article and noun are ungrammatical, whereas in Conditions 1 and 3, the German article and noun pair are grammatical, although they are the wrong translation of the English word.

(*das Messer*), and spoons are masculine (*der Löffel*). However, Köpcke (1982) delineated 24 phonological, 5 morphological, and 15 semantic rules that accounted for the gender of 90% of a corpus of 1466 German words. We will discuss the issue of morphophonological regularities in more detail below. Adding to the complexity, articles and adjectives are also marked for number and case information. As a result, the same article does not consistently precede nouns. There is also case syncretism, so that there is no one-to-one mapping of forms.

We tested L2 learners on translation recognition (De Groot, 1992), a task that learners at varied levels of L2 proficiency can complete with reasonable accuracy. While translation *recognition* has not been previously used to investigate gender processing, lexical-level translation *production* tasks have successfully been used in several studies to investigate how both L1 speakers and L2 learners access gender information (Vigliocco, Lauer, Damian & Levelt, 2002; Salamoura & Williams, 2007; Paolieri et al., 2010).

We also included a gender decision task (Radeau & Van Berkum, 1996), in which participants were presented with a bare noun and identified whether the noun was masculine, feminine, or neuter in German. L1 studies using the gender decision task have shown a strong influence of gender transparency and regularity for languages such as French, Italian, and Hebrew (for a review, see Gollan & Frost, 2001). By focusing on German, in which gender assignment is generally less transparent, but in which some words contain transparent gender markings (e.g., Bordag et al., 2006; Wegener, 2000), we hope to identify the potential contribution of transparency and language regularity to L2 gender acquisition.

Overview of Experiments 1 and 2: speeded lexical task

In the translation recognition task, both the native German speakers (Experiment 1) and the L2 German learners (Experiment 2) saw only translations from English to German, to allow participants in both language groups to anticipate the German gendered article after seeing the English noun. The order of presentation (i.e., English first, German second) was not changed across language groups in order to keep any gender preparation in translation consistent between language groups. As a result, L2 German learners (i.e., native English speakers) engaged in forward translation, whereas native German speakers completed backward translation. Participants saw sequences such as THE ROPE – DAS SEIL and responded as to whether the second noun phrase was a correct translation of the first noun phrase. We used a version of the task in which participants respond "yes" to correct translations and "no" for incorrect translations via a button press on a keyboard. In all cases, the critical trials were the "no" trials.

Table 1 illustrates the three critical conditions in which participants had to reject German translations of English words as incorrect:

1) Translation Mismatch (TM): German words that matched the critical English word in gender but not translation.
2) Gender Mismatch (GM): German words that matched the critical English word in translation, but not in gender.
3) Translation and Gender Mismatch (TGM): German words that did not match the critical English word in either gender or translation.

https://doi.org/10.1017/S1366728914000534 Published online by Cambridge University Press

The critical comparison was between Condition 1 and 3, as these two conditions differ in whether the presented gender of the wrong translation matched or mismatched the anticipated gender of the right translation.

Experiment 1: native German speakers

The goal of the first experiment was to determine whether a novel adaptation of the translation recognition task is sensitive to the matched or mismatched gender conditions (cf. gender congruency effects, La Heij et al., 1998). Unlike a traditional translation-recognition task, we chose to present det+noun NPs and not bare nouns because gender effects for Germanic languages may only appear in det+noun contexts (e.g., La Heij et al., 1998; but see Paolieri et al., 2010). We tested native German speakers learning English. Results of the native German speakers will indicate whether the translation recognition task is sensitive to gender processing in the first place, and whether the comparison of the translation mismatch (TM) to the translation and gender mismatch (TGM) condition is sensitive to gender violations.

Predictions

If the translation-recognition paradigm indexes gender processing, we expect participants to show longer response latencies in rejecting translation mismatches (TM): the initial presentation of the gender should confirm the anticipated translation, only to be violated with the presentation of the noun, resulting in processing costs as participants revise their response. Responses in this condition would therefore contrast with response times to items in the translation and gender mismatch (TGM) condition that require no response revision.

Method

Participants

Thirty-five participants were included in the analyses: twenty-eight L1 German–L2 English learners were recruited from a large German university, and seven were recruited at a large American university. Their data did not pattern differently. See Table 2 for an overview of participant characteristics.

Table 2. *Characteristics of native German and L2 German learners.*

Measure	Native Germans (*n* = 35)	L2 Germans (n = 72)
Age (years)	27.66(5.77)	28.35(11.59)
Ospan	46.68(6.90)	46.66(7.86)
Simon (milliseconds)	39.88(20.73)	36.12(31.66)
Average L1 self-rating (10 pt scale)	9.41(0.70)	9.56(0.59)
Average L2 self-rating (10 pt scale)	7.49(1.33)	7.10(1.44)
Age of L2 acquisition (years)	10.37(1.59)	16.24(5.75)
Length of immersion (months)	6.54(6.30)	14.60(25.11)

*Note. SD*s are in parentheses. Self-ratings reflect average ratings on listening, reading, speaking, writing, where 0 is the lowest and 10 is the highest score. "Ospan" refers to the Operational Span Task (Tokowicz et al., 2004), "Simon" to the Simon Task (Bialystok et al., 2004).

Materials

We included sixty simple English nouns, each paired with three German translations as outlined above. The three possible German translations for each English word were matched pairwise as closely as possible on frequency (Baayen, Piepenbrock & Gulikers, 1995) and word length. We performed t-tests to ensure that across conditions and participant lists, as well as across conditions within a given participant list, German frequency[1] (Baayen et al., 1995; Quasthoff, 2002) and word length, as well as English frequency (Baayen et al., 1995; Kučera & Francis, 1967), age of acquisition (Coltheart, 1981), word length, familiarity (Coltheart, 1981) and imageability (Gilhooly & Logie, 1980) were not significantly different (all *ps* > .1; see Appendix A for the items and Table 1 for an example of item matching). Care was taken that the incorrect gender assignment could not form a plausible gender assignment in a different case (e.g., zero plurals or genitive case or dative case).[2]

We included another 172 items as fillers. Fifty percent of all trials were true "yes" trials, where participants were presented correct gender-noun translation pairs. The distribution of gender in the critical items corresponded to the natural distribution of gender in German in which 50% of all words are masculine, 30% feminine and 20% neuter (Hohlfeld, 2006). However, overall, including fillers, participants saw close to an equal distribution of gender across lists to avoid a strategic bias based on the presentation of the article prior to the German translation.

[1] We follow recent recommendations to use the Universität Leipzig Wortschatz corpus instead of Celex (see Brysbaert, Buchmeier, Conrad, Jacobs, Bölte & Böhl, 2011; Koester, Gunter, Wagner & Friederici, 2004).

[2] After reviewing the items, one item *die Rücken* (the.FEM back.MASC) was marked as an incorrect gender-noun pairing although it could be correct in the plural. Removing the item did not significantly change the models.

https://doi.org/10.1017/S1366728914000534 Published online by Cambridge University Press

Procedure

We tested participants individually in a quiet room on a PC using E-prime stimulus presentation software (Schneider, Eschman & Zuccolotto, 2002). Text appeared in white in bold Courier New 24 font size on a black background in all upper case letters. Prior to each trial, a fixation sign (+) appeared at the center of the computer screen. Participants pressed the spacebar to initiate the trial. An English article appeared on the screen for 200 ms followed by an inter-stimulus interval (ISI) of 100 ms, followed by an English noun for 700 ms. After an ISI of 700 ms, a German article then appeared for 200 ms, followed by an ISI for 100 ms, followed by a German noun for 700 ms. Participants were asked to indicate whether the German noun phrase was the correct translation of the English noun phrase by pressing a "yes" or "no" button on the keyboard. Participants had 4000 ms from onset of the German noun to respond before the next fixation sign appeared. Response hand was kept consistent across participants. Participants completed 15 practice items including all three types of incorrect translation types and were given verbal feedback on their performance. Following the translation recognition task, participants performed the Gender Decision task (see Experiments 3 & 4). They then completed a Language History Questionnaire, Operation Span Task (Tokowicz, Michael & Kroll, 2004), and Simon Task (Bialystok, Craik, Klein & Viswanathan, 2004; Linck, Hoshino & Kroll, 2008). These measures of cognitive functioning were used to match L1 and L2 groups to ensure that any apparent insensitivities to gender are due to language proficiency and not cognitive abilities, and also to investigate whether cognitive resources may modulate the ability to use gender in the L2.

Analysis

We analyzed RTs using mixed-effect linear regression models (e.g., Baayen, Davidson & Bates, 2008). We analyzed accuracy data using mixed-effect logistic regression models appropriate for categorical data (Jaeger, 2008). We performed regression analyses using the lme4 package in the statistical software application R (R Development Core Team, 2005). We also controlled for several variables known to modulate visual word recognition. Table 3 summarizes the full list of predictors. We natural-log-transformed RTs to reduce skewness in the distribution.

Having two or more collinear predictors in the model can have harmful consequences, inflating estimated standard errors and causing inaccurate inferences of those predictors. To avoid collinearity, for every pair of predictors for which the Pearson correlation index *r* exceeded the threshold of 0.5, we applied the mediation analysis procedure to establish whether the influence of one predictor was fully mediated by the other predictor (MacKinnon, 1994; Judd & Kenny, 1981). We retained the predictor that absorbed the variance explained by the other predictor, unless there were theoretical reasons to retain the other predictor, as noted below. We also tested collinearity of the fixed effects in the final model and we residualized measures correlated at r ≥ .3. We tested for interactions between significant variables and dropped non-significant interactions from the model. For all random and fixed effects included in the final model, we confirmed the improvement in the goodness-of-fit by the likelihood ratio test, comparing a model with the variable to one without the variable. Our exploration of the random effects structure included modeling random slopes and contrasts with all fixed effects and interactions in the final models. No random effect, beyond the random intercepts for word (stimulus1, stimulus2) and participant (subject), proved to significantly improve the model's performance, as indicated by the likelihood ratio test. Unless noted otherwise, we report only fixed effects that reached significance at the 5% level.

Table 3. *Additional predictors of interest*

Variable	Range (units)	Mean (*SD*)	Median
E_fam	413:636 (100 to 700)	551 (48)	558
E_frq	0.89:3.09 (log units)	1.69 (0.50)	1.61
E_img	232:650 (100 to 700)	514 (99)	556
G_len	4:14 (characters)	6.39 (2.30)	6
G_frq	7:15 (ranked)	11 (2)	11
GA_RT	760:1185 (ms)	942 (97)	930
GA_Acc	93.16:98.72 (%correct)	96.57 (1.35)	97.01

Note. Variables of interest include *E_fam*, familiarity rating for the English word (Coltheart, 1981); *E_frq*, frequency rating for the English word (Baayen et al., 1995); *E_img*, imageability rating for the English word (Gilhooly & Logie, 1980); *G_len*, German word length; *G_frq*, frequency rating for German words (Quasthoff, 2002); *GA_RT*, participants' response time for Gender Decision task; and *GA_Acc*, participants' accuracy for Gender Assignment task.

Results

Latencies

We excluded incorrect responses (48 items, or 2.25% of the data), as well as items with response latencies below 300 ms and above 3000 ms or deviating 2.5 *SD*s from a participant's mean (66 items, 3.09% of the data), from RT analyses.

Mean latencies are presented in Table 4, and the results of the final model are presented in Table 5. Testing collinearity, the condition number $\kappa = 12.79$ indicated medium, non-harmful, collinearity (Baayen, 2008, p. 198).

https://doi.org/10.1017/S1366728914000534 Published online by Cambridge University Press

Table 4. *Mean response times (RT) in ms across the three critical conditions of the translation recognition task for native German speakers and L2 German learners.*

	Condition					
	TM		GM		TGM	
Group	RT	*SD*	RT	*SD*	RT	*SD*
Native German (Experiment 1)	1004	278	900	302	967	309
L2 Learner (Experiment 2)	1054	393	1306	512	1074	420

Note. Response times (RT) and standard deviations (SD) for conditions TM (Translation Mismatch), GM (Gender Mismatch), and TGM (Translation and Gender Mismatch).

Native German speakers were indeed sensitive to grammatical gender across translation pairs. Participants rejected translation mismatches (TM) more slowly than gender mismatches (GM) or translation and gender mismatches (TGM). This difference between the TM and TGM conditions in particular suggests sensitivity to grammatical gender across translations. Releveling the factors showed the difference between TGM and GM to be significant as well (*$p < .0001$*). Including gender matching as a fixed factor significantly improved the explanatory value of the model ($\chi^2(2) = 51.79$, $p < .001$). There were no significant interactions, although the interaction between the reference level TGM and GM and imageability norms approached significance (*$p = .084$; all other ps $> .1$)*.

Participants produced faster response times on translation decisions the further into the experiment they advanced, indicating a practice effect. They also responded significantly faster to more imageable items. Both block order and imageability significantly improved the explanatory value of our model (block order: $\chi^2(1) = 176.69, p < .001$, imageability: $\chi^2(1) = 11.57, p < .001$). Including participant performance on the Operational Span Task and Simon Task did not significantly improve the fit of the model ($p > .1$).

Accuracy

Native German participants performed at ceiling for this task: 97.06%, 97.70% and 98.27% correct for the translation mismatch (TM), gender mismatch (GM), and translation and gender mismatch (TGM) conditions respectively. The model did not show any significant differences between conditions.

Discussion

Our results suggest that native German speakers with L2 knowledge of English are sensitive to grammatical gender across translations. The difference between the translation mismatch (TM) and the translation and gender mismatch (TGM) condition provides compelling evidence for this sensitivity. These results are analogous to Stroop-like effects, such that participants judging mismatched translations with the correct article must override prepotent responses to accept the translation as correct. Although the initial presentation of the gender-marked article confirmed the anticipated translation, presentation of the noun violated expectations, resulting in processing costs as participants revised their response. Alternatively, if the gender is correct, participants may be also waiting to verify the noun translation. For the GM and TGM condition, participants can judge immediately

Table 5. *Final RT model for native German speakers on Translation Recognition task*

Fixed Effects	Estimate	MCMCmean	HPD95lower	HPD95upper	pMCMC	Pr(>|t|)
(Intercept)	7.286	7.285	7.177	7.387	<.001	<.001
TranslationGM	−0.140	−0.140	−0.167	−0.110	<.001	<.001
TranslationTGM	−0.054	−0.054	−0.086	−0.025	.002	.001
Block	−0.071	−0.071	−0.082	−0.062	<.001	<.001
E_img	0.000	0.000	−0.001	0.000	<.001	<.001
Random Effects	Std.Dev.	MCMCmedian	MCMCmean	HPD95lower	HPD95upper	
1 Stimulus2	0.032	0.023	0.022	0.000	0.043	
2 Stimulus1	0.048	0.046	0.046	0.030	0.063	
3 Subject	0.202	0.154	0.155	0.128	0.184	
4 Residual	0.238	0.241	0.241	0.233	0.249	

Note. Table of fixed and random effects shows estimates of the regression coefficients; highest posterior density intervals (HPDs), a Bayesian measure of confidence intervals; Monte Carlo Markov chain (MCMC) estimates of the p-values using 5,000 samples; as well as p-values obtained with the t-test for fixed effects using the difference between the number of observations and the number of fixed effects as the upper bound for the degrees of freedom (cf. Baayen, 2008). The coefficients for translation condition in the model represent contrasts with the reference level, TranslationTM, which is mapped onto the intercept.

whether the NP is correct based on the gender alone, which would speed up latencies. The results align with gender congruency effects found in other studies investigating lexical access. Past studies have reported consistent priming effects between modifiers and nouns, where participants respond slower when an invalid gender prime precedes the target noun (e.g., Scherag et al., 2004). With respect to why the TGM condition is slower than the GM condition, participants may process the ungrammaticality more quickly than the translation mismatch in rejecting the translation. While the GM translation is ungrammatical as a stand-alone noun-phrase (der.MASC Seil.NEUT), the TGM translation is grammatical as a stand-alone noun-phrase (der.MASC Mord.MASC).

Experiment 2: L2 German learners

Experiment 1 validated the new translation-recognition paradigm with L1 German speakers. In Experiment 2, we asked whether L2 German learners would show similar sensitivity to gender mismatches.

Predictions

If gender processing is impaired at the lexical level in L2 learning, then L2 German learners should show no difference in processing the translation mismatch (TM) condition and the translation and gender mismatch (TGM) condition, because the defining difference between these conditions is the match or mismatch of gender to the correct translation.

In contrast, if L2 German learners process L2 gender at the lexical level, they should display sensitivity to the mismatch in gender between TM and TGM conditions. If L2 German learners become more sensitive to L2 grammatical gender as L2 proficiency increases, then they should display faster and more accurate processing of translations, and an increasing sensitivity to the mismatch of gender across conditions.

Method

Participants

Eighty-two native English speakers with intermediate to advanced proficiency in German were recruited from several American universities. Participants with exposure to German before the age of 11 (n = 10) were dropped from the analyses to focus on late L2 learners (e.g., Weber-Fox & Neville, 1996), so that 72 participants were included in the final analyses. See Table 2 for an overview of participant characteristics.

Materials

The materials were identical to those used as in Experiment 1.

Procedure

The same procedure was used as in Experiment 1.

Analysis

The same analytical approach was used as in Experiment 1.

Results

In addition to the predictors considered in Experiment 1, we also included several measures of L2 proficiency: self-reported estimates of L2 proficiency, the age at which participants first started learning their L2 German, and participants' RT and percent accuracy score on a gender decision task (see Experiments 3 and 4 for more detail). As an additional measure of proficiency, participants completed a speeded picture naming task in L1 and L2 (e.g., Jared & Kroll, 2001). Table 6 summarizes the main L2 proficiency predictors.

Latencies

Incorrect responses (1097 items; 24.99% of the total data), as well as items with response latencies below 300 ms and above 3000 ms or deviating 2.5 *SD*s from a participant's mean (191 items or 4.35% of the data), were excluded from RT analyses.

All of the L2 predictors were significantly correlated, so we applied a mediation analysis to identify non-mediated predictors. Although the mediation analysis showed the time to make the gender decision (GA_RT) to be the best predictor of the L2 proficiency variables considered here, we did not use it in our final model because it was unclear whether the strong effect of GA_RT on RTs was because it was an adequate predictor of L2 proficiency, or because it was a predictor of the general speed of response for a given individual. Thus, we retained the predictor identified as second best in the mediation analysis, the participants' self-rating of their L2 proficiency (L2Rating), as a predictor of L2 proficiency in the final model. Translation condition, block order, imageability, and English log frequency significantly improved the fit of the model and were included in the final model. No interactions were significant ($ps > .1$). The results of the model are presented in Table 7. The condition number $\kappa = 18.93$ indicated medium, non-harmful, collinearity of the fixed effects.

Native English-speaking participants rejected translation mismatches (TM) faster than gender mismatches

https://doi.org/10.1017/S1366728914000534 Published online by Cambridge University Press

Table 6. *Additional variables of interest for L2 German learners (n = 72)*

Variable	Range (units)	Mean (*SD*)	Median
L2Rating	1.60: 9.60 (10 pts)	7.10 (1.44)	7.40
L2Naming	6.67:100.00 (%correct)	67.59 (18.08)	70.00
GA_Acc	32.03: 84.19 (%correct)	61.90 (11.16)	62.82
Mos_Abroad	0.00:144.00 (months)	14.60 (25.11)	6.00
L2AoA	11.00: 52.00 (years)	16.24 (5.75)	14.00

Note. Variables of interest include *L2Rating*, participants' mean self-rating in reading, writing, speaking, and comprehension; *L2Naming*, %correct in an L2 picture naming task; *GA_Acc*, %correct on the Gender Assignment task; *Mos_Abroad*, months spent in a German-speaking country; and *L2AoA*, mean age at which participants started learning German.

Table 7. *Final RT model for L2 German learners on Translation Recognition task*

Fixed Effects	Estimate	MCMCmean	HPD95lower	HPD95upper	pMCMC	Pr(>|t|)
(Intercept)	7.932	7.931	7.696	8.192	<.001	<.001
TranslationGM	0.192	0.192	0.157	0.225	<.001	<.001
TranslationTGM	0.020	0.021	−0.011	0.053	.200	.233
Block	−0.040	−0.040	−0.048	−0.030	<.001	<.001
E_img	−0.001	−0.001	−0.001	0.000	<.001	<.001
E_frq	−0.082	−0.083	−0.122	−0.045	<.001	<.001
L2Rating	−0.062	−0.062	−0.089	−0.033	<.001	.001
Random Effects	**Std.Dev.**	**MCMCmedian**	**MCMCmean**	**HPD95lower**	**HPD95 upper**	
Subject	0.229	0.169	0.170	0.148	0.194	
Stimulus1	0.058	0.055	0.056	0.042	0.071	
Stimulus2	0.043	0.038	0.038	0.018	0.057	
Residual	0.253	0.256	0.256	0.250	0.263	

Note. Table shows estimates of the regression coefficients; highest posterior density intervals (HPDs); Monte Carlo Markov chain (MCMC) estimates of the p-values using 5,000 samples; as well as p-values obtained with the t-test for fixed effects. The coefficients for translation condition in the model represent contrasts with the reference level, TranslationTM, which is mapped onto the intercept.

(GM) in the analysis, controlling for block order, imageability, and English word frequency. Critically, there was no significant difference between the translation mismatch (TM) condition and the translation and gender mismatch (TGM) condition. See Table 4 for mean latencies. Unlike the native German speakers in Experiment 1, when the nouns mismatched, these English learners of German were sensitive to the semantics of the mismatch but not to the grammatical gender. Including gender matching as a fixed factor significantly improved the explanatory value of the model, $\chi^2(2) = 81.43$, $p <$ .001.

Participants were also faster the further into the experiment they advanced, indicating a practice effect. They responded significantly faster to translations that were more imageable and more frequent than those that were not. Finally, there was a significant effect of L2 proficiency. Participants who rated themselves as more proficient were faster than those who rated themselves as less proficient. Importantly, there were no significant interactions between translation conditions and proficiency, or with any of the other variables of interest (*ps* > .1). All four factors significantly improved the explanatory value of our model (block order: $\chi^2(1) =$ 74.58, $p < .001$; imageability: $\chi^2(1) = 29.26$, $p < .001$; frequency: $\chi^2(1) = 15.20$, $p < .001$; L2 proficiency: $\chi^2(1)$ $= 10.37$, $p < .01$). Including participant performance on the Operation Span Task and Simon Task did not significantly improve the fit of the model (*ps* > .1).

Accuracy

To test for sensitivity to grammatical gender in response accuracy, we fit a model using response accuracy (correct vs. incorrect) as the binary outcome variable and translation condition (TGM vs. TM vs. GM) as the main predictor variable. Translation, block order, imageability, English log frequency and L2 naming significantly improved the fit of the model and were included in the final model. The results are presented in Table 8.

https://doi.org/10.1017/S1366728914000534 Published online by Cambridge University Press

Table 8. *Final accuracy model for L2 German learners on Translation Recognition task*

Fixed Effects	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−3.262	0.731	−4.464	<.001
TranslationGM	0.253	0.407	0.621	.535
TranslationTGM	0.380	0.448	0.850	.396
L2naming	0.064	0.008	8.485	<.001
E_frq	0.351	0.166	2.119	.034
E_img	0.002	0.001	2.764	.006
Block	0.124	0.046	2.686	.007
TranslationGM:L2naming	−0.059	0.007	−8.663	<.001
TranslationTGM:L2naming	−0.009	0.008	−1.131	.258
Random effects	Variance	Std.Dev.		
Stimulus2	0.25335	0.50334		
Subject	0.63183	0.79487		
Stimulus1	0.11472	0.33871		

Note. The coefficients for translation condition in the model represent contrasts with the reference level, TranslationTM, which is mapped onto the intercept.

Positive coefficients (β) indicate a greater likelihood of the outcome 1 (correct answer), and negatives a greater likelihood of 0 (incorrect answer). Overall, the model showed a very good fit of the data. The index of concordance *C-statistic* = 0.906, and the rank correlation between observed responses and predicted probabilities was *Dxy* = 0.811 (for details see Baayen, 2008, pp. 223–224; Jaeger, 2008). The condition number, $\kappa = 17.76$, indicated medium, non-harmful, collinearity.

We first fit the model that showed significant contrasts between TGM and GM conditions, and between the reference level TM and the GM condition, and no additional interactions. Participants were more accurate rejecting translation and gender mismatches (TGM; 90.21%, *SD* = 29.60) than gender mismatches (GM; 39.37%, *SD* = 48.80), and also more accurate rejecting translation mismatches (TM; 90.74%, *SD* = 28.19) than gender mismatches (GM). Crucially, there were no significant differences between the reference level TM and the TGM condition, which corroborates the RT findings. The final model in Table 7 shows a significant interaction between translation condition and L2 naming (accuracy on the L2 picture naming task), where participants with higher accuracy on picture naming were also more accurate in judging translations in the GM condition. Including both translation condition and L2naming as fixed factors significantly improved the explanatory value of our model, as did the interaction (Translation: $\chi^2(2) = 105.47$, $p < .001$; L2naming: $\chi^2(1) = 12.00$, $p < .001$; Translation*L2naming: $\chi^2(2) = 186.89$, $p < .001$).

Again, participants were more accurate the further into the experiment they advanced. They were also significantly more accurate on more imageable and more frequent items. All three factors significantly improved the explanatory value of our model (block order: $\chi^2(1) = 93.64$, $p < .001$, imageability: $\chi^2(1) = 7.29$, $p < .01$, frequency: $\chi^2(1) = 4.35$, $p < .05$).

Discussion

Overall, the results suggest that native English speakers who were late L2 German learners were not sensitive to L2 grammatical gender mismatches in the translation recognition task. Their latencies and accuracy did not differ significantly between the translation mismatch condition (TM) and the translation and gender mismatch (TGM) condition. Sensitivity to gender also did not change as a function of proficiency. Furthermore, the difficulty learners faced in rejecting the gender mismatch (GM) items, as reflected in increased latencies, indicates that they were particularly tuned to the semantics of the translation task and did not focus on the gender component of the task. Alternatively, longer latencies may be an indication that late L2 learners, regardless of their proficiency, do not engage in automatic gender processing and rely instead on explicit gender knowledge in making gender decisions. Specifically, they may have adopted a strategy of waiting for the appearance of the noun before making a gender decision. If the noun was incorrect, they could simply reject the translation. If the noun was correct, however, they then needed to make a gender judgment, resulting in increased processing costs.

Although a direct comparison between native German and L2 German language groups cannot be made due to the difference in translation directions, the results leave open the possibility that more proficient L2 German

https://doi.org/10.1017/S1366728914000534 Published online by Cambridge University Press

Table 9. *Selective RT model for L2 German learners on Translation Recognition task*

Fixed Effects	Estimate	MCMCmean	HPD95lower	HDP95upper	pMCMC	PR(>|t|)
(Intercept)	7.778	7.788	7.519	8.051	0.000	0.000
Block	−0.035	−0.035	−0.046	−0.022	0.000	0.000
E_Img	−0.001	−0.001	−0.001	0.000	0.000	0.000
TranslationGM	0.446	0.445	0.267	0.640	0.000	0.000
TranslationTGM	0.153	0.146	−0.003	0.304	0.059	0.051
L2Rating	−0.050	−0.051	−0.080	−0.021	0.002	0.010
E_frq	−0.081	−0.081	−0.125	−0.036	0.001	0.001
TranslationGM: L2Rating	−0.035	−0.035	−0.060	−0.010	0.008	0.006
TranslationTGM:L2Rating	−0.020	−0.019	−0.039	0.002	0.070	0.055
Random Effects	Std.Dev.	MCMCmedian	MCMCmean	PD95lower	HPD95upper	
Subject	0.221	0.162	0.163	0.140	0.186	
Stimulus1	0.063	0.058	0.058	0.041	0.078	
Stimulus2	0.044	0.032	0.031	0.000	0.055	
Residual	0.248	0.254	0.254	0.245	0.263	

Note. The coefficients for translation condition in the model represent contrasts with the reference level, TranslationTM, which is mapped onto the intercept.

learners with a more native-like command of German might begin to show sensitivity to gender using this task. It is also possible that results are obscured by data from items that participants did not know well. L2 German learners may show sensitivity to gender for items that are well established in their lexicon. To address the possibility that L2 German participants were sensitive to gender on a subset of known words, we conducted a post-hoc item analysis. Participants performed a gender decision task (see Experiments 3 and 4) in which they assigned gender to the critical nouns. We repeated the latency model on only those nouns to which a given participant correctly assigned gender (see Hopp, 2013, for a similar approach).

As in the previous model, participants were significantly faster to reject translation mismatches (TM; $M = 1033$ ms, $SD = 381$) than gender mismatches (GM; $M = 1270$ ms, $SD = 511$). Crucially, the final model in Table 9 now also revealed a nearly-significant difference in the critical comparison between the translation mismatch (TM) condition and the translation and gender mismatch (TGM; $M = 1018$ ms, $SD = 369$) condition in the same direction as for L1 German speakers. The model also revealed a significant interaction of proficiency with translation conditions between the TM and GM condition, and more importantly a near-significant interaction of proficiency with translation conditions in the critical comparison between the TM and the TGM condition.

As a group, the L2 German learners showed little sensitivity to grammatical gender processing in the current experimental paradigm, although the post-hoc analysis of L2 German data trends in the direction of the L1 German speaker data. It is possible, however, that L2 German learners might nevertheless be able to use grammatical gender in a task more under their explicit control (e.g., Montrul, Davidson, de la Fuente & Foote, 2014). This is the focus of Experiments 3 and 4.

Overview of Experiments 3 and 4: Metalinguistic measures of gender sensitivity

Experiment 2 suggests that although learner performance as a group in rejecting nouns with incorrect gender is only barely above chance, some L2 German learners are able to recognize incorrect gender assignment. What mechanisms might help learners bootstrap their way into the gender system? One possibility is that morphophonological features provide particularly salient cues to the L2 learner in deciding the gender of a given word. Previous research with both monolinguals and bilinguals has suggested that speakers of a language with grammatical gender are sensitive to regularities, such as the phonological features of a noun (Gollan & Frost, 2001; Schiller, Münte, Horemans & Jansma, 2003; Bordag et al., 2006). In German, the *–e* ending on a noun typically indicates feminine gender. This cue is the first gender cue that children learning German acquire (Mills, 1986) and is also the most reliable phonological gender cue in German, whereby 90% of words ending in *–e* are feminine (Wegener, 2000). Previous research by Bordag and colleagues indicates that bilinguals are sensitive to phonological cues when determining L2 gender. Results of a speeded gender picture naming task and a speeded grammaticality judgment task with L1 German speakers showed no sensitivity to these morphophonological

https://doi.org/10.1017/S1366728914000534 Published online by Cambridge University Press

gender regularities, whereas intermediate L2 German learners revealed an influence of the phonological noun form in both reaction times and accuracy.

In Experiments 3 and 4 we designed a speeded metalinguistic task, in which we tested participants' ability to assign gender explicitly rather than detect agreement errors as Bordag and colleagues (2006) did. The same participants from Experiments 1 and 2 were tested, allowing a similar comparison between native speakers and L2 learners to the one made by Bordag and colleagues.

Experiment 3: native German speakers

Similar to Experiment 1, the goal of Experiment 3 was to determine whether our novel adaptation of the gender task is sensitive to previously documented gender effects. More importantly, Experiment 3 served as a comparison to results from Experiment 4 with L2 learners.

Predictions

Phonological cues may no longer play a role in typical adult L1 gender processing (cf. Bordag et al., 2006). As a result, native speakers of German may show reduced or completely absent cue sensitivity. Alternatively, they may continue to show sensitivity to morphophonological features of a noun (Schiller et al., 2003; Hohlfeld, 2006).

Method

Participants

The same 35 native German speakers participated as in Experiment 1.

Materials

Two hundred and twenty one nouns, critical simple nouns as well as constituents of filler compound nouns from Experiments 1 and 2, formed the basis of the materials for Experiments 3 and 4. Participants were tested on the full set of 221 nouns. Forty eight items were selected for further analysis according to the following criteria: Using the phonological gender categorization of nouns in Bordag et al. (2006), monomorphemic nouns from the gender decision task were divided into three categories: typical, ambiguous, and atypical. Feminine nouns ending in *–e* were assigned to the typical category. Masculine or neuter nouns ending in a consonant were assigned to the ambiguous category because two gender options map onto one phonological cue. Feminine nouns ending in a consonant, or masculine or neuter nouns ending in *–e,* were categorized as atypical because they do not follow typical gender assignment. Of the 221 items, 16 items corresponded to the criteria for the atypical category. We then matched items corresponding to the other two categories as closely as possible to these 16 atypical nouns for a total of 48 items, 16 in each category. Items were matched on German and English frequency, German word length, English word length of the translation, English Age of Acquisition, and English Imageability (*ps* $>$ .1). The full set of critical items is available in Appendix B.

Procedure

Participants completed a computer-based gender decision task in which they read German bare-stem nouns printed as black capital letters on a white background. The presentation of the items was semi-randomized such that no more than three nouns of the same gender were presented in a row. A fixation sign (+) appeared for 250 ms, followed by a German noun. Participants selected one of three keyboard keys (c, b, m) to indicate whether the given noun was masculine, feminine, or neuter. The noun stayed on the screen until participants made a decision or the trial timed-out after 5000 ms. Participants were told to make their selections as quickly and accurately as possible. Key assignment mapping was consistent between participants. Response latencies and accuracy were recorded and analyzed.

Analysis

The same analytical approach was used as in Experiment 1.

Results

Latencies

Incorrect responses (31 data points; 1.84% of the data), as well as items with response latencies below 300 ms and above 3000 ms or deviating 2.5 *SD*s from a participant's mean (55 data points; 3.27% of the data), were excluded from RT analyses.

Translation, familiarity, and German frequency significantly improved the fit of the model and were included in the final model. The index of collinearity, the condition number $\kappa = 34.28$, was relatively high, and so we residualized measures correlated at $r \geq .3$. For instance, we regressed G_freq out of E_fam, by obtaining the residuals of an ordinary regression with E_fam as a dependent variable and G_freq as its predictor: the residual E_fam (labeled as rE_fam) was strongly correlated with the original value of E_fam ($r = 0.95$, $p < .001$) and was not correlated with G_freq ($p = 0.96$). We entered residual rE_fam into the model for RT together with G_freq. All effects retained their significance even after residualization. The new index of collinearity,

https://doi.org/10.1017/S1366728914000534 Published online by Cambridge University Press

Table 10. *Final RT model for native German speakers on Gender Decision task*

Fixed Effects	Estimate	MCMCmean	HPD95lower	HPD95upper	pMCMC	Pr(>|t|)
(Intercept)	6.616	6.616	6.522	6.706	<.001	<.001
Ambiguous	0.077	0.077	0.033	0.118	.001	.001
Atypical	0.020	0.020	−0.020	0.063	.358	.364
rE_fam	−0.001	−0.001	−0.001	0.000	.018	.018
G_freq	0.013	0.013	0.006	0.021	<.001	.001
Random Effects	Std.Dev.	MCMCmedian	MCMCmean	HPD95lower	HPD95upper	
Word	0.052	0.049	0.049	0.035	0.064	
Subject	0.098	0.087	0.088	0.068	0.109	
Residual	0.185	0.186	0.186	0.180	0.193	

Note. The coefficients for gender category (ambiguous, atypical) in the model represent contrasts with the reference level, typical, which is mapped onto the intercept.

$\kappa = 9.53$, indicated non-harmful collinearity. The final model, with residualized measures, is presented in Table 10. No interactions were significant (*ps* > .1).

Participants responded faster to nouns typically marked for gender (*M* = 886 ms, *SD* = 207) than those ambiguously marked for gender (*M* = 940 ms, *SD* = 228). Releveling of the factors also showed a significant difference between ambiguous and atypical nouns (*M* = 889 ms, *SD* = 210). No differences were observed between typical and atypical nouns. Including gender category as a fixed factor significantly improved the explanatory value of our model, $\chi^2(2) = 12.52, p < .01$.

There was also a significant effect of frequency and familiarity, such that participants responded faster to more frequent than less frequent items and faster to more familiar than less familiar items. Both frequency and familiarity significantly improved the fit of our model (frequency: $\chi^2(1) = 0.19, p < .01$; familiarity: $\chi^2(1) = 5.78, p < .05$).

Accuracy

Native German participants performed at ceiling for this task: 98.71% correct (*SD* = 11.29) for typical, 97.56% correct (*SD* = 15.44) for ambiguous, and 98.00% correct (*SD* = 14.02) for atypical gender-marked nouns. The model did not show any significant differences between conditions.

Discussion

The results of Experiment 3 suggest that native German speakers are sensitive to gender regularities in their L1. Unlike in Bordag et al. (2006), native speakers of German in the present study showed sensitivity to morphophonological regularities, with significantly slower reaction times to ambiguously-marked nouns than to typical or atypical nouns. These findings are in line with other monolingual research (e.g., RT data: Schiller et al., 2003; off-line data: Hohlfeld, 2006), demonstrating sensitivity to morphophonological correlates of grammatical gender among native speakers of German. However, typical and atypical nouns were processed at similar latencies. We suggest that atypical nouns are most likely learned by rote memory as exceptions, allowing for quick access of these lexical items during processing (but see Gollan & Frost, 2001).

Experiment 4: L2 German learners

Is there evidence that L2 learners show sensitivity to grammatical gender at the lexical level in this paradigm? Regardless of how L2 gender is thought to be represented, all theories allow for the potential role of morphophonological cues in gender assignment. In L2 lexical accounts, in which gender features on the noun are impaired (Carroll, 1989), noun gender learning occurs via a reliance on compensatory mechanisms, such as phonological cues, and rote memorization. Even representational deficit accounts, which hypothesize that L1 and L2 gender agreement are fundamentally different (e.g., Hawkins & Franceschina, 2004), allow that metalinguistic strategies, including sensitivity to morphophonological cues, can support L2 gender assignment (e.g., Hawkins, 2009).

Predictions

With increasing proficiency, L2 learners might increasingly be able to take advantage of morphophonological cues to gender assignment. However, if L2 learners cannot apply metalinguistic strategies, we would anticipate continued gender decision errors, even for highly proficient L2 speakers.

Table 11. *Final RT model for L2 German learners on Gender Decision task*

Fixed Effects	Estimate	MCMCmean	HPD95lower	HPD95upper	pMCMC	Pr(>|t|)
(Intercept)	7.030	7.034	6.838	7.234	<.001	<.001
Ambiguous	0.200	0.200	0.156	0.246	<.001	<.001
Atypical	0.225	0.224	0.173	0.273	<.001	<.001
G_freq	0.033	0.033	0.025	0.042	<.001	<.001
rE_fam	−0.001	−0.001	−0.001	−0.001	<.001	<.001
L2Rating	−0.059	−0.059	−0.085	−0.036	<.001	<.001
Random Effects	Std.Dev.	MCMCmedian	MCMCmean	HPD95lower	HPD95upper	
Subject	0.181	0.145	0.145	0.123	0.169	
Word	0.050	0.048	0.048	0.031	0.067	
Residual	0.259	0.262	0.262	0.254	0.270	

Note. The coefficients for gender category (ambiguous, atypical) in the model represent contrasts with the reference level, typical, which is mapped onto the intercept.

Method

Participants

The same 72 native English L2 German learners participated as in Experiment 2.

Materials

Materials were the same as in Experiment 3.

Procedure

The procedure was the same as in Experiment 3.

Analysis

The same analytical approach was used as in Experiment 1.

Results

Latencies

Incorrect responses (1229 data points; 35.56% of the data), as well as items with response latencies below 300 ms and above 3000 ms or deviating 2.5 *SD*s from a participant's mean (137 data points; 3.96% of the data), were excluded from RT analyses.

Of the included predictors, only gender category, German frequency, English familiarity, and L2 self-rating significantly improved the model and will be further discussed. The index of collinearity, the condition number $\kappa = 39.76$, was relatively high. All effects retained their significance even after residualization. The new index of collinearity, $\kappa = 13.52$, indicated non-harmful collinearity. The final model, with residualized measures, is presented in Table 11. No interactions were significant (*ps* > .1).

Participants assigned gender significantly faster to typical ($M = 1115$ ms, $SD = 407$) than ambiguous ($M = 1309$ ms, $SD = 482$) and to typical than atypical ($M = 1294$ ms, $SD = 483$) nouns in the analysis controlling for German frequency, English familiarity, and L2 self-ratings. There was no significant difference between ambiguous and atypical items. Including gender category as a fixed factor significantly improved the explanatory value of our model, $\chi^2(2) = 58.82, p < .001$.

Participants with higher ratings of L2 proficiency were also faster overall on the task. Importantly, ratings of L2 proficiency did not interact with gender category, indicating that proficiency did not modulate sensitivity to the three gender categories as measured by RTs. In addition to gender category and proficiency, RTs were affected by German word frequency and English familiarity. RTs were faster for high frequency German words than for low frequency words, and for highly familiar words than for less familiar words. All three factors significantly improved the explanatory value of our model (proficiency: $\chi^2(1) = 13.71, p < .001$; frequency: $\chi^2(1) = 40.65, p < .001$, familiarity: $\chi^2(1) = 15.20, p < .001$).

Accuracy

To test for participants' sensitivity to phonological gender regularities in response accuracy, we fit a model using response accuracy (correct vs. incorrect) as the binary outcome variable and gender category (typical vs. ambiguous vs. atypical) as the main predictor variable. Because the initial condition number $\kappa = 50.00$ was high, we residualized measures correlated at r ≥ .3. All effects retained their significance even after residualization. The new index of collinearity, $\kappa = 14.90$, indicated

Table 12. *Final accuracy model for L2 German learners on Gender Decision task*

Fixed Effects	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4.039	0.751	5.376	<.001
Ambiguous	−1.686	0.489	−3.449	<.001
Atypical	−2.583	0.494	−5.226	<.001
L2naming	0.036	0.006	5.840	<.001
L2AoA	−0.030	0.012	−2.463	.014
Order	−0.005	0.001	−3.640	<.001
rE_fam	0.013	0.003	4.620	<.001
rE_img	0.005	0.002	2.526	.012
G_freq	−0.235	0.050	−4.726	<.001
ambiguous:L2naming	−0.015	0.007	−2.322	.020
atypical:L2naming	−0.019	0.007	−2.761	.006
Random Effects	Variance	Std.Dev.		
Subject	0.20141	0.44878		
Word	0.42655	0.65311		

Note. The coefficients for gender category (ambiguous, atypical) in the model represent contrasts with the reference level, typical, which is mapped onto the intercept.

non-harmful collinearity. The final model, with residualized measures, is presented in Table 12. Positive coefficients (β) indicate a greater likelihood of the outcome 1 (correct answer), and negatives a greater likelihood of 0 (incorrect answer). Overall, the model showed *C-statistic* = 0.867, *Dxy* = 0.735.

In line with our predictions and the RT results, participants were significantly more accurate on typical than ambiguous or atypical items in the analysis controlling for L2 picture naming accuracy, L2 age of acquisition, order of presentation, and German word frequency. Unlike the RT results, there was also a significant difference between ambiguous and atypical items in the predicted direction (model not shown), such that participants were more accurate on ambiguous than atypical items. Respective mean correct responses for each of the conditions were 89.30% (*SD* = 30.91) for typical items, 63.34% (*SD* = 48.21) for ambiguous items, and 36.64% (*SD* = 48.20) for atypical items. Including gender category as a fixed factor significantly improved the explanatory value of our model, $\chi^2(2) = 79.08$, $p < .001$.

Two L2 measures of proficiency also showed significant effects: the age at which participants first learned the L2, and their accuracy on the L2 picture naming (PN) task. This last predictor showed a significant interaction with gender category between the typical and ambiguous categories as well as between the typical and atypical categories. Figure 1 suggests that less-proficient learners took greater advantage of the gender cue provided in the typical category, as reflected in the asymmetry of the difference between typical vs. ambiguous accuracy scores for less-proficient compared to more-proficient participants. Especially less-proficient participants may benefit from phonological cues to gender, and with increasing proficiency, there is a qualitative shift in how gender is processed. However, the graph also shows that highly proficient learners were near ceiling in their performance on typically marked nouns, so the apparent asymmetry may be a function of this high level of accuracy. Including both L2AoA and L2PN accuracy, as well as the Gender Category*L2PN interaction, significantly improved the fit of our model (L2AoA: $\chi^2(1) = 5.79$, $p < .05$; L2PN: $\chi^2(1) = 79.61$, $p < .001$; Gender Category*L2PN: $\chi^2(2) = 8.07$, $p < .05$).

Participants responded less accurately the further into the experiment they proceeded, suggesting fatigue effects. Participants were significantly more accurate on more frequent, more imageable, and more familiar items. Including all of these factors significantly improved the fit of our model (order: $\chi^2(1) = 12.98$, $p < .001$; frequency: $\chi^2(1) = 18.91$, $p < .001$; imageability: $\chi^2(1) = 6.08$, $p < .05$; familiarity: $\chi^2(1) = 18.365$, $p < .001$).

Discussion

Experiment 4 replicates the production results of Bordag et al. (2006) with a gender decision task. Results indicated significantly faster response latencies for typical nouns than atypical and ambiguous nouns, similar to Bordag et al. (2006). Percent accuracy scores showed greater accuracy for typical nouns and particularly low accuracy in the atypical category.

https://doi.org/10.1017/S1366728914000534 Published online by Cambridge University Press

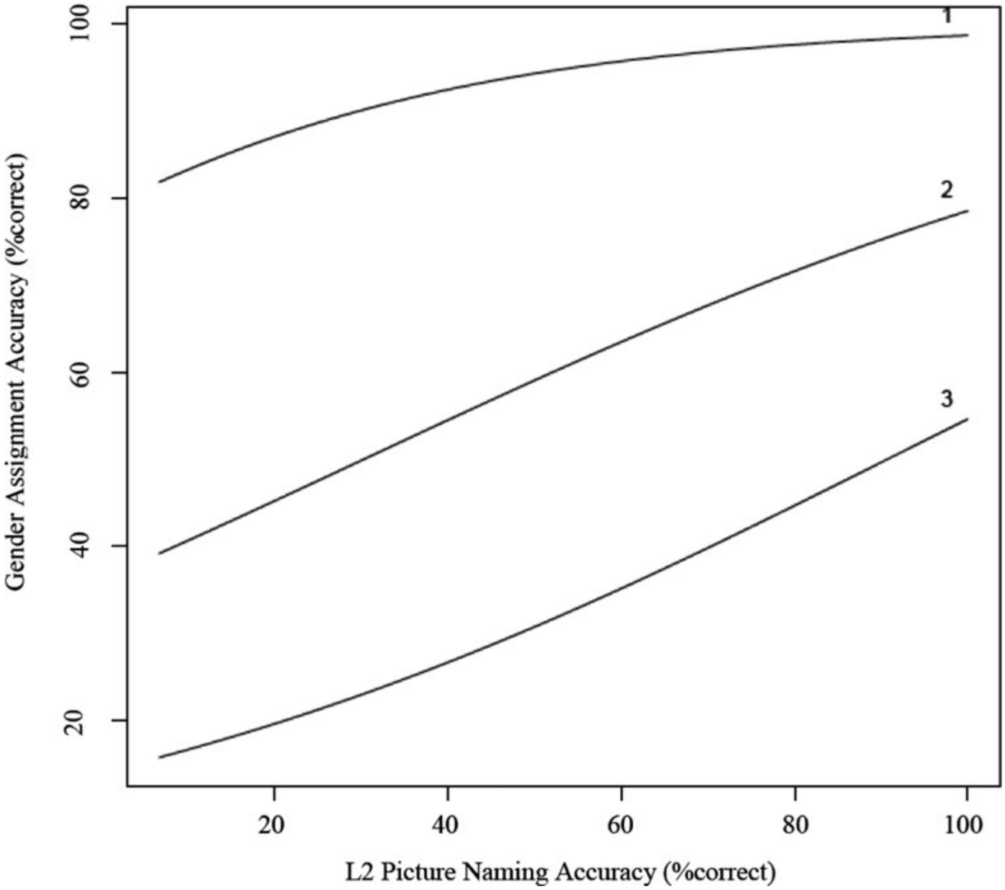

Figure 1. The effect of proficiency on gender assignment accuracy for L2 German learners. The graph depicts gender categories where 1 = typical, 2 = ambiguous, and 3 = atypical nouns.

The accuracy results also revealed significant noun type by proficiency interactions, suggesting that, contrary to results in Experiment 2, increased proficiency may change the way L2 learners assign noun gender. Although both the native German speakers and L2 German learners appeared sensitive to gender categories, and showed significant differences between the typical and ambiguous categories, only the L2 German learners additionally showed a significant difference in response latencies and accuracy between the typical and atypical conditions, with longer latencies and lower accuracy in the atypical condition. These results may suggest an increased sensitivity of L2 German learners, particularly of less-proficient learners, to the morphophonological distribution of noun endings when assigning gender in an effort to "make sense" of the seemingly arbitrary assignment of gender to nouns. In line with previous research (e.g., Hawkins & Franceschina, 2004), the particularly low accuracy on atypically marked nouns, even for highly proficient participants, may indicate an overgeneralization of morphophonological gender patterns.

General Discussion

Taken together, the lexical-level results of the present experiments parallel previous syntactic-level studies highlighting the difficulty of L2 gender processing (e.g., Sabourin, Stowe & De Haan, 2006), as well as a dissociation between tasks that require more automatic processing (Experiment 2) and those that are under the participant's control (Experiment 4) (Sabourin et al., 2006). The data show clear developmental patterns, in that English learners of German perform translation recognition faster and more accurately with increasing proficiency. However, none of the factors used as an index of L2 proficiency predicted sensitivity to gender. As a group, the L2 learners remained relatively insensitive to grammatical gender in the speeded translation recognition task. There was, however, a near-significant interaction with proficiency in the post-hoc analysis of the translation recognition data on items to which participants correctly assigned gender in the separate gender assignment task. Although these results should be interpreted with caution, we suggest, analogous to Hopp (2013), that L2 learners

https://doi.org/10.1017/S1366728914000534 Published online by Cambridge University Press

may need to reach a critical level of proficiency before using target gender knowledge in a native-like manner. As Hopp (2013) argues, "the ability to employ grammatical gender as a predictive cue hinges on overall mastery of lexical gender assignment" (p. 51). Additionally, the analysis of the gender decision data in Experiment 4 suggests that L2 learners may rely on distributional properties of gender to bootstrap their way into the gender system, displaying a more native-like *behavior* (distinct from native-like *processing*) on typically-marked nouns (Bordag et al., 2006). Particularly less-proficient L2 learners may benefit from these morphophonological gender regularities.

L2 gender sensitivity may, however, vary by gender type (Opitz, Regel, Müller & Friederici, 2013; Mills, 1986). In an additional post-hoc analysis, we included the gender of the critical items in Experiments 1 and 2 in the models. For the native German speakers, we found that participants responded fastest to feminine gender items (943 ms, *SD* = 297), followed by neuter (950 ms, *SD* = 280), and then masculine gender items (971 ms, *SD* = 310). The difference between feminine and masculine (*p < .01)* and feminine and neuter (*p < .01)* but not masculine and neuter gender (*p = .97)* reached significance in the model. There were no significant interactions with translation condition, however, (*p*s > *.1)* indicating that these differences had no bearing on gender sensitivity in translation recognition for native German speakers. For L2 German learners, we found that gender does not significantly account for any variance in the model (*p*s > .1).

The evidence provided by the current study offers constraints on the extent to which the sample of late L2 learners had learned grammatical gender at the point in language development at which they were tested. The observed effects of proficiency on other aspects of performance indicate that differences in proficiency within this range did not directly affect L2 gender processing. Taking these constraints into consideration, we consider three general explanations of the data.

First, L2 learners do not acquire a lexical representation of grammatical gender in German, providing evidence for hard constraints at the lexical level on the late acquisition of grammatical structures not present in the L1. The results from Experiment 1 and 2 might support such a view, although this view is qualified by the fact that post-hoc analyses of the translation recognition task suggest L2 German participants are sensitive to gender for items for which they can accurately assign gender. According to these findings, a strong version of hard constraints in L2 acquisition as posited for syntactic gender agreement (i.e., Hawkins & Chan, 1997) may not be tenable for lexical-level processing (see also Keating, 2009; Foucart & Frenck-Mestre, 2012; Hopp, 2013). The results support the views of Grüter et al. (2012) and Hopp (2013) that L2 learners develop weaker links between nouns and their respective genders in the L2 lexicon and underscore the importance of studies of L2 gender at the lexical or phrase level to better understand L2 difficulties with syntactic gender agreement. More research is clearly needed to fully develop these accounts.

A second possibility is that L2 learners process grammatical gender, but use a qualitatively different mechanism to do so than native speakers of German. This possibility would point to difficulties in online processing, but the ability to use strategies and rules for more successful explicit processing (Hawkins & Franceschina, 2004; Clahsen & Felser, 2006; Montrul et al., 2014). In fact, this alternative concurs with the data presented in the current study, which showed particular difficulties for L2 German learners in the translation recognition task in Experiment 2, where more automatic processing was key. In the less speeded and more explicitly gender-focused gender decision task of Experiment 4, participants showed a sensitivity to gender for a subset of nouns with particularly salient morphophonological markings.

A third possibility is that late L2 learners may approximate the behavior of native speakers, but do not have the cognitive resources to process L2 grammatical features as quickly as L1 speakers. McDonald (2006) showed that native speakers, when put under cognitive stress or load, perform like non-native speakers (see also Kilborn, 1991; Hopp, 2010). Processing load may play a crucial role in mediating learner effects. Hopp (2010) similarly showed a dramatic decrease in native German speakers' ability to identify gender mismatches when processing degraded stimuli. Our results raise the possibility that the speeded constraints of the translation-recognition task may have taxed L2 learner cognitive resources to such an extent as to negatively impact gender processing. While the pattern of data presented cannot fully distinguish between a difference in processing mechanisms or a cognitive resources alternative, we believe that the evidence is sufficiently compelling to reject the strong version of hard constraints in the acquisition of L2-specific features (see also Keating, 2009; Foucart & Frenck-Mestre, 2012: Hopp, 2013).

It is possible, however, that our translation recognition task, while speeded, was not sensitive enough to pick up on emerging gender sensitivity in this group of L2 learners. By investigating a structure particularly difficult to acquire, our approach to testing ultimate attainment was also conservative. Recent evidence suggests that even in related languages that both have grammatical gender, learning L2 gender is difficult (Lemhöfer, Schriefers & Hanique, 2010). Similar to research on gender agreement in syntactic contexts, in future research it will be important to investigate these lexical issues using measures that are particularly sensitive to the earliest time course of processing, such as ERPs (e.g., Frenck-Mestre, Foucart,

https://doi.org/10.1017/S1366728914000534 Published online by Cambridge University Press

Carrasco-Ortiz & Herschensohn, 2009; Foucart & Frenck-Mestre, 2012).

In light of the results of Experiments 1 and 2, Experiments 3 and 4 provide a plausible mechanism that L2 learners can use to bootstrap performance (see also Bordag et al., 2006). Grüter et al. (2012) raise the possibility that in late L2 learners, the mapping of gender to the lexical entry is weak. If this is true, language regularities such as morphophonological cues may help L2 learners to compensate for deficient lexical representations, or, alternatively, may serve to strengthen connections between the lexical entry and the gender node. Strikingly, even in a language whose morphophonological cues to gender are not as transparent, learners exploit these gender regularities and use them to make overt gender decisions. It is here that we also see a modulation by proficiency, suggesting that it is the less-proficient L2 learners who appear to rely more heavily on gender cues. At first blush, the direction of this finding may seem counter-intuitive. Why would L2 learners become decreasingly sensitive to cues that support gender access? As others have also argued, gender may initially be computed for a given lexical entry each time it is accessed before eventually being permanently stored (Corbett, 1991; Bordag et al., 2006). Note that L1 adults in our study continued to show some sensitivity to morphophonological cues (see also Schiller et al., 2003; Hohlfeld, 2006). More advanced L2 learners may therefore be on a continuum of cue sensitivity, relying decreasingly on morphophonological gender cues as gender information becomes more lexicalized (Taraban & Kempe, 1999).

Conclusions

The current study contributes to the growing body of literature documenting constraints to L2 learning by revealing these constraints even in contexts of reduced grammatical complexity and at lexical levels of processing. Particularly the lack of modulation of effects by proficiency in the translation recognition task points to the extreme difficulty in learning certain L2-specific language structures. However, the fact that L2 learners did show some sensitivity to grammatical gender in the translation recognition task for those items to which they could explicitly assign gender suggests that maturational constraints alone are inadequate to account for learner behavior. In this respect, the post-hoc analyses in the translation recognition task and learners' sensitivity to gender in the gender decision task are particularly hopeful. In future research, it will be important to assess the developmental trajectory of different grammatical features to determine whether there are principled reasons why some language features are particularly difficult for adult L2 learners to acquire and to distinguish between features that may require extremely high proficiency to reveal native-like performance and those that may never reach native-like performance.

Appendix A. *Critical items in Experiments 1 & 2*

E_Word		G_Translation		TM		TGM		GM	
THE	HOLE	das	Loch	DAS	OBST	DER	WITZ	DER	LOCH
THE	WOOD	das	Holz	DAS	OBST	DER	WITZ	DER	HOLZ
THE	BREAD	das	Brot	DAS	OBST	DER	WITZ	DER	BROT
THE	ROPE	das	Seil	DAS	MEER	DER	MORD	DER	SEIL
THE	TENT	das	Zelt	DAS	MEER	DER	MORD	DER	ZELT
THE	GOAL	das	Ziel	DAS	MEER	DER	MORD	DER	ZIEL
THE	SHIRT	das	Hemd	DAS	GEMÜSE	DER	FUCHS	DER	HEMD
THE	SHEEP	das	Schaf	DAS	GEMÜSE	DER	FUCHS	DER	SCHAF
THE	PALACE	das	Schloss	DAS	GEMÜSE	DER	FUCHS	DER	SCHLOSS
THE	SCARF	der	Schal	DER	KNOPF	DAS	BRETT	DAS	SCHAL
THE	RABBIT	der	Hase	DER	KNOPF	DAS	BRETT	DAS	HASE
THE	KNIGHT	der	Ritter	DER	KNOPF	DAS	BRETT	DAS	RITTER
THE	SPOON	der	Löffel	DER	SCHIRM	DAS	SCHACH	DAS	LÖFFEL
THE	SCREAM	der	Schrei	DER	SCHIRM	DAS	SCHACH	DAS	SCHREI
THE	COUGH	der	Husten	DER	SCHIRM	DAS	SCHACH	DAS	HUSTEN
THE	TIP	der	Hinweis	DER	DAUMEN	DAS	GEMÄLDE	DAS	HINWEIS
THE	FOAM	der	Schaum	DER	DAUMEN	DAS	GEMÄLDE	DAS	SCHAUM
THE	BELT	der	Gürtel	DER	DAUMEN	DAS	GEMÄLDE	DAS	GÜRTEL

https://doi.org/10.1017/S1366728914000534 Published online by Cambridge University Press

Appendix A. *Continued*

E_Word		G_Translation		TM		TGM		GM	
THE	BACK	der	Rücken	DER	WERT	DIE	GEFAHR	DIE	RÜCKEN
THE	HEAD	der	Kopf	DER	WERT	DIE	GEFAHR	DIE	KOPF
THE	LOOK	der	Blick	DER	WERT	DIE	GEFAHR	DIE	BLICK
THE	ENTRANCE	der	Eingang	DER	VORTRAG	DIE	AHNUNG	DIE	EINGANG
THE	ACCIDENT	der	Unfall	DER	VORTRAG	DIE	AHNUNG	DIE	UNFALL
THE	STOMACH	der	Magen	DER	VORTRAG	DIE	AHNUNG	DIE	MAGEN
THE	BONE	der	Knochen	DER	TEPPICH	DAS	ERLEBNIS	DAS	KNOCHEN
THE	PROOF	der	Beweis	DER	TEPPICH	DAS	ERLEBNIS	DAS	BEWEIS
THE	SWEAT	der	Schweiß	DER	TEPPICH	DAS	ERLEBNIS	DAS	SCHWEIß
THE	WING	der	Flügel	DER	KREBS	DAS	FASS	DAS	FLÜGEL
THE	CREEK	der	Bach	DER	KREBS	DAS	FASS	DAS	BACH
THE	GIANT	der	Riese	DER	KREBS	DAS	FASS	DAS	RIESE
THE	SNAKE	die	Schlange	DIE	FAHNE	DAS	ELEND	DAS	SCHLANGE
THE	DOLL	die	Puppe	DIE	FAHNE	DAS	ELEND	DAS	PUPPE
THE	FORK	die	Gabel	DIE	FAHNE	DAS	ELEND	DAS	GABEL
THE	LEMON	die	Zitrone	DIE	ABGASE	DAS	GESPENST	DAS	ZITRONE
THE	COIN	die	Münze	DIE	ABGASE	DAS	GESPENST	DAS	MÜNZE
THE	ONION	die	Zwiebel	DIE	ABGASE	DAS	GESPENST	DAS	ZWIEBEL
THE	INSURANCE	die	Versicherung	DIE	ERFAHRUNG	DAS	GEFÄNGNIS	DAS	VERSICHERUNG
THE	POPULATION	die	Bevölkerung	DIE	ERFAHRUNG	DAS	GEFÄNGNIS	DAS	BEVÖLKERUNG
THE	CONFIDENCE	die	Zuversicht	DIE	ERFAHRUNG	DAS	GEFÄNGNIS	DAS	ZUVERSICHT
THE	APOLOGY	die	Entschuldigung	DIE	SCHÖNHEIT	DAS	SCHICKSAL	DAS	ENTSCHULDIGUNG
THE	SCIENCE	die	Wissenschaft	DIE	SCHÖNHEIT	DAS	SCHICKSAL	DAS	WISSENSCHAFT
THE	REALITY	die	Wirklichkeit	DIE	SCHÖNHEIT	DAS	SCHICKSAL	DAS	WIRKLICHKEIT
THE	FUTURE	die	Zukunft	DIE	EINLADUNG	DAS	GEWICHT	DAS	ZUKUNFT
THE	HEALTH	die	Gesundheit	DIE	EINLADUNG	DAS	GEWICHT	DAS	GESUNDHEIT
THE	SERVICE	die	Bedienung	DIE	EINLADUNG	DAS	GEWICHT	DAS	BEDIENUNG
THE	AREA	das	Gebiet	DAS	BLATT	DER	HERBST	DER	GEBIET
THE	FACE	das	Gesicht	DAS	BLATT	DER	HERBST	DER	GESICHT
THE	LAW	das	Gesetz	DAS	BLATT	DER	HERBST	DER	GESETZ
THE	STEP	der	Schritt	DER	SCHUSS	DAS	GEHALT	DAS	SCHRITT
THE	PAIN	der	Schmerz	DER	SCHUSS	DAS	GEHALT	DAS	SCHMERZ
THE	SMOKE	der	Rauch	DER	SCHUSS	DAS	GEHALT	DAS	RAUCH
THE	NIGHT	die	Nacht	DIE	KIRCHE	DAS	MEHL	DAS	NACHT
THE	REST	die	Ruhe	DIE	KIRCHE	DAS	MEHL	DAS	RUHE
THE	VOICE	die	Stimme	DIE	KIRCHE	DAS	MEHL	DAS	STIMME
THE	LIE	die	Lüge	DIE	MIETE	DAS	KREUZ	DAS	LÜGE
THE	BOX	die	Kiste	DIE	MIETE	DAS	KREUZ	DAS	KISTE
THE	SKIN	die	Haut	DIE	MIETE	DAS	KREUZ	DAS	HAUT
THE	STRANGER	die	Fremde	DIE	ANKUNFT	DAS	EREIGNIS	DAS	FREMDE
THE	EXCEPTION	die	Ausnahme	DIE	ANKUNFT	DAS	EREIGNIS	DAS	AUSNAHME
THE	WEAKNESS	die	Schwäche	DIE	ANKUNFT	DAS	EREIGNIS	DAS	SCHWÄCHE

https://doi.org/10.1017/S1366728914000534 Published online by Cambridge University Press

Appendix B. *Critical items in Experiments 3 & 4*

Category	G_Word	Gender	Length	Freq_a	E_Translation	Length	Log Freq_b
Typical	KIRCHE	F	6	8	CHURCH	6	2.20
Typical	RUHE	F	4	9	REST	4	2.34
Typical	TÜTE	F	4	13	BAG	3	1.80
Typical	SCHLANGE	F	8	11	SNAKE	5	1.19
Typical	ZITRONE	F	7	15	LEMON	5	1.15
Typical	SCHEIBE	F	7	12	PANE	4	0.44
Typical	FAHNE	F	5	12	FLAG	4	1.32
Typical	HOSE	F	4	12	PANTS	5	1.22
Typical	KARTE	F	5	9	CARD	4	1.67
Typical	LIEBE	F	5	9	LOVE	4	2.55
Typical	STELLE	F	6	8	POSITION	8	2.29
Typical	MÜNZE	F	5	13	COIN	4	0.95
Typical	PUPPE	F	5	10	DOLL	4	1.27
Typical	REISE	F	5	9	TRIP	4	1.76
Typical	FLASCHE	F	7	11	BOTTLE	6	1.92
Typical	STIMME	F	6	9	VOICE	5	2.37
Ambiguous	BLATT	N	5	9	LEAF	4	1.22
Ambiguous	FENSTER	N	7	9	WINDOW	6	2.13
Ambiguous	HEMD	N	4	12	SHIRT	5	1.67
Ambiguous	KOPF	M	4	8	HEAD	4	2.66
Ambiguous	ZELT	N	4	12	TENT	4	1.58
Ambiguous	WINTER	M	6	9	WINTER	6	1.91
Ambiguous	MANTEL	M	6	12	COAT	4	1.73
Ambiguous	NAGEL	M	5	11	NAIL	4	1.11
Ambiguous	GÜRTEL	M	6	13	BELT	4	1.35
Ambiguous	BRIEF	M	5	9	LETTER	6	2.09
Ambiguous	SCHAL	M	5	13	SCARF	5	0.95
Ambiguous	TRAUM	M	5	10	DREAM	5	1.77
Ambiguous	KLAVIER	N	7	12	PIANO	5	1.43
Ambiguous	TANZ	M	4	11	DANCE	5	1.67
Ambiguous	BETT	N	4	10	BED	3	2.39
Ambiguous	BALL	M	4	9	BALL	4	1.97
Atypical	GEMÜSE	N	6	11	VEGETABLE	9	1.38
Atypical	GABEL	F	5	14	FORK	4	1.17
Atypical	HASE	M	4	13	RABBIT	6	1.07
Atypical	KÄSE	M	4	12	CHEESE	6	1.46
Atypical	STADT	F	5	6	CITY	4	2.34
Atypical	ZWIEBEL	F	7	14	ONION	5	1.02
Atypical	WELT	F	4	6	WORLD	5	2.87
Atypical	FAHRT	F	5	9	RIDE	4	1.55
Atypical	ARBEIT	F	6	7	WORK	4	2.92
Atypical	KARTOFFEL	F	9	14	POTATO	6	1.10
Atypical	HAUT	F	4	10	SKIN	4	1.96
Atypical	JUGEND	F	6	10	YOUTH	5	1.82
Atypical	RIESE	M	5	13	GIANT	5	1.56
Atypical	MUSIK	F	5	8	MUSIC	5	2.13
Atypical	NACHT	F	5	7	NIGHT	5	2.63
Atypical	GEBURT	F	6	10	BIRTH	5	1.78

Notes. a_Quasthoff, U. (2002); b_Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995)

https://doi.org/10.1017/S1366728914000534 Published online by Cambridge University Press

References

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, *122,* 292–305.

Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.

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

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The Celex Lexical Database*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Bates, E., Devescovi, A., Hernandez, A., & Pizzamiglio, L. (1996). Gender priming in Italian. *Perception & Psychophysics, 85,* 992–1004.

Bialystok, E., Craik, F. I. M., Klein, R., & Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: Evidence from the Simon task. *Psychology and Aging, 19,* 290–303.

Bordag, D., Opitz, A., & Pechmann, T. (2006). Gender processing in first and second langauges: The role of noun termination. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 1090–1101.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. Experimental Psychology, *58*, 412–424.

Carroll, S. (1989). Second-language acquisition and the computational paradigm. *Language Learning, 39,* 535–594.

Carroll, S. (2001). *Input and evidence: The raw material of second language acquisition*. Amsterdam: John Benjamins Publishing Company.

Carroll, S. (2005). Input and SLA: Adults' sensitivity to different sorts of cues to French Gender. *Language Learning, 55,* 79–138.

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics, 27,* 3–42.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology, 33a,* 497–505.

Corbett, G. (1991). *Gender*. Cambridge: Cambridge University Press.

De Groot, A. M. B. (1992). Determinants of word translation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1001–1018.

DeKeyser, R. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22,* 499–533.

Foucart, A., & Frenck-Mestre, C. (2012). Can late L2 learners acquire new grammatical features? Evidence from ERPs and eye-tracking. *Journal of Memory and Language, 66,* 226–248.

Frenck-Mestre, C., Foucart, A., Carrasco-Ortiz, H., & Herschensohn, J. (2009). "Processing of grammatical gender in French as a first and second language: Evidence from ERPs". In L. Roberts, G.D. Véronique, A. Nilsson, & M. Tellier (Eds.), *EUROSLA Yearbook 9*, (pp.76–106). Amsterdam: John Benjamins Publishing Company.

Gilhooly, K. J., & Logie, R. H. (1980). Age of acquisition, imagery, concreteness, familiarity and ambiguity measures for 1944 words. *Behaviour Research Methods and Instrumentation, 12,* 395–427.

Gillon Dowens, M., Vergara, M., Barber, H. A., & Carreiras, M. (2010). Morphosyntactic processing in late second-language learners. *Journal of Cognitive Neuroscience, 22,* 1870–1887.

Gollan, T. H., & Frost, R. (2001). Two Routes to Grammatical Gender: Evidence from Hebrew. *Journal of Psycholinguistic Research, 30,* 627–651.

Grüter, T., Lew-Williams, C., & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research, 28,* 191–215.

Guillelmon, D., & Grosjean, F. (2001). The gender marking effect in spoken word recognition: The case of bilinguals. *Memory and Cognition, 29,* 503–511.

Hawkins, R. (2009). Statistical learning and innate knowledge in the development of second language proficiency: Evidence from the acquisition of gender concord. In: Benati AG (ed.) Issues in second language proficiency, 63–78. London: Continuum International Publishing.

Hawkins, R., & Chan, C. (1997). The partial availability of Universal Grammar in second language acquisition: The 'failed functional features hypothesis.' *Second Language Research, 13,* 187–226.

Hawkins, R., & Franceschina, F. (2004). Explaining the acquisition and nonacquistion of determiner noun gender concord in French and Spanish. In Prévost, P. & Paradis, J. (eds.). *The Acquisition of French in Different Contexts: Focus on functional categories*, pp 175–205. Philadelphia: John Benjamins Publishing Company.

Hohlfeld, A. (2006). Accessing grammatical gender in German: The impact of gender-marking regularities. *Applied Psycholinguistics, 27,* 127–142.

Hopp, H. (2010). Ultimate attainment in L2 inflectional morphology: Performance similarities between non-native and native speakers. *Lingua*, *120,* 901–931.

Hopp, H. (2013). Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research*, *29,* 33–56.

Jaeger, T. F. (2008). Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of Memory and Language, 59,* 434–446.

Jared, D., & Kroll, J. F. (2001). Do bilinguals activate phonological representations in one or both of their languages when naming words? *Journal of Memory and Language, 44,* 2–31.

Judd, C. M., & Kenny, D. A. (1981). Process Analysis: Estimating mediation in treatment evaluations. *Evaluation Review, 5,* 602–619.

Keating, G. (2009). Sensitivity to violations of gender agreement in native and nonnative Spanish: An eye-movement investigation. *Language Learning, 59,* 503–535.

Kilborn, K. (1991). Selective impairment of grammatical morphology due to induced stress in normal listeners:

https://doi.org/10.1017/S1366728914000534 Published online by Cambridge University Press

Implications for aphasia. *Brain and Language, 41,* 275–288.

Köpcke, K.-M. (1982). *Untersuchungen zum Genussystem der deutschen Gegenwartssprache*. Tübingen: Niemeyer.

Koester, D., Gunter, T. C., Wagner, S., & Friederici, A. D. (2004). Morphosyntax, prosody, and linking elements: the auditory processing of German nominal compounds. *Journal of Cognitive Neuroscience, 16,* 1647–1668.

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language, 33,* 149–174.

Kučera, F., & Francis, W. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

La Heij, W., Mak, P., Sander, J., & Willeboordse, E. (1998). The gender-congruency effect in picture-word tasks. *Psychological Research, 61,* 209–219.

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

Linck, J. A., Hoshino, N., & Kroll, J. F. (2008). Cross-language lexical processes and inhibitory control. *The Mental Lexicon, 3,* 349–374.

MacKinnon, D.P. (1994). Analysis of mediating variables in prevention and intervention research. In A. Cazares and L. A. Beatty, *Scientific methods in prevention research*, pp. 127–153. NIDA Research Monograph 139. DHHS Pub. No. 94–3631. Washington, DC: U.S. Govt. Print. Office.

MacWhinney, B. (1997). Second language acquisition and the Competition Model. In J. Kroll & De Groot (Eds.), *Tutorials in bilingualism*, Mahwah, NJ: Lawrence Erlbaum.

McCarthy, C. (2008). Morphological variability in the comprehension of agreement: An argument for representation over computation. *Second Language Research, 24,* 459–486.

McDonald, J. L. (2006). Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language, 55,* 381–401.

Mills, A. E. (1986). The acquisition of gender: a study of German and English. *Springer Series in Language and Communication; 20*. Berlin: Springer.

Montrul, S., Davidson, J., de la Fuente, I., & Foote, R. (2014). Early language experience facilitates the processing of gender agreement in Spanish heritage speakers. *Bilingualism: Language and Cognition, 17,* 118–138.

Montrul, S., Foote, R., & Perpiñán, S. (2008). Gender agreement in adult second language learners and Spanish heritage speakers: The effects of age and context of acquisition. *Language Learning, 58,* 503–553.

Opitz, A., Regel, S., Müller, G., & Friederici, A. D (2013). Neurophysiological evidence for morphological underspecification in German strong adjective inflection. *Language, 89,* 231–264.

Paolieri, D., Cubelli, R., Macizo, P., Bajo, M. T., Lotto, L., and Job, R. (2010). Grammatical gender processing in Italian and Spanish bilinguals. *Q. J. Exp. Psychol., 63,* 1631–1645.

Paradis, M. (2004). *A neurolinguistic theory of bilingualism*. Amsterdam: John Benjamins.

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.

Quasthoff, U. (2002). Deutscher Wortschatz im Internet [Online database]. http://www.wortschatz.uni-leizpig.de/ (2002). Leipzig, Germany: University of Leipzig.

R Development Core Team (2005), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org.

Radeau, M., & Van Berkum, J. J. A. (1996). Gender decision. *Language and Cognitive Processes, 11,* 605–610.

Sabourin, L., & Stowe, L. A. (2008). Second language processing: When are L1 and L2 processed similarly. *Second Language Research, 24,* 397–430.

Sabourin, L., Stowe, L., & De Haan, G. J. (2006). Transfer effects in learning a second language grammatical gender system. *Second Language Research, 22,* 1–29.

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.

Salamoura, A., & Williams, J. N. (2007). The representation of grammatical gender in the bilingual lexicon: Evidence from Greek and German. *Bilingualism: Language and Cognition, 10,* 257–275.

Scherag, A., Demuth, L., Rösler, F., Neville, H., & Röder, B. (2004). The effects of late acquisition of L2 and the consequences of immigration on L1 for semantic and morpho-syntactic language aspects. *Cognition, 93,* B97–B108.

Schiller, N. O., Münte, T., Horemans, I., & Jansma, B. M. (2003). The influence of semantic and phonological factors on syntactic decisions: An event-related brain potential study. *Psychophysiology, 40,* 869–877.

Schneider, W., Eschmann, A., & Zuccolotto, A. (2002). E–Prime v1.1. Pittsburgh, PA: Psychology Software Tools Inc.

Schwartz, B. D., & Sprouse, R. (1996). L2 cognitive states and the full transfer/ full access model. *Second Language Research, 12,* 40–72.

Taraban, R., & Kempe, V. (1999). Gender processing in native and non-native Russian speakers. *Applied Psycholinguistics, 20,* 119–148.

Tokowicz, N., & MacWhinney, B. (2005). Implicit and explicit measures of sensitivity to violations in second language grammar. *Studies in Second Language Acquisition, 27,* 173–204.

Tokowicz, N., Michael, E., & Kroll, J. F. (2004). The roles of study abroad experience and working memory capacity in the types of errors made during translation. *Bilingualism: Language and Cognition, 7,* 255–272.

Tokowicz, N., & Warren, T. (2010). Beginning adult L2 learners' sensitivity to morphosyntactic violations: A self-paced reading study. *European Journal of Cognitive Psychology, 22,* 1092–1106.

Ullman, M.T. (2001). The neural basis of lexicon and grammar in first and second language: The declarative/procedural model. *Bilingualism: Language and Cognition, 4,* 105–122.

https://doi.org/10.1017/S1366728914000534 Published online by Cambridge University Press

Vigliocco, G., Lauer, M., Damian, M. F., & Levelt, W. J. M. (2002). *Semantic and syntactic forces in noun phrase production*. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 46–58.

Weber-Fox, C. M., & Neville, H. J. (1996). Maturational constraints on functional specializations for language processing: ERP and behavioral evidence in bilingual speakers. *Journal of Cognitive Neuroscience, 8,* 231–256.

Wegener, H. (2000). German gender in children's second language acquisition. In B. Unterbeck & M. Rissanen (Eds.), *Gender in grammar and cognition*, pp. 511–544. Berlin: Mouton de Gruyter.

https://doi.org/10.1017/S1366728914000534 Published online by Cambridge University Press