Second Language Research
http://slr.sagepub.com/

Grammatical gender in L2: A production or a real-time processing problem?

Theres Grüter, Casey Lew-Williams and Anne Fernald

Second Language Research 2012 28: 191
DOI: 10.1177/0267658312437990

The online version of this article can be found at:
http://slr.sagepub.com/content/28/2/191

Published by:
SAGE
http://www.sagepublications.com

Additional services and information for *Second Language Research* can be found at:

Email Alerts: http://slr.sagepub.com/cgi/alerts

Subscriptions: http://slr.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations: http://slr.sagepub.com/content/28/2/191.refs.html

>> Version of Record - May 23, 2012

What is This?

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

second
language
research

Grammatical gender in L2: A production or a real-time processing problem?

Second Language Research
28(2) 191–215
© The Author(s) 2012
Reprints and permission: sagepub.
co.uk/journalsPermissions.nav
DOI: 10.1177/0267658312437990
slr.sagepub.com
SAGE

Theres Grüter
University of Hawai'i, USA

Casey Lew-Williams
University of Madison-Wisconsin, USA

Anne Fernald
Stanford University, USA

Abstract
Mastery of grammatical gender is difficult to achieve in a second language (L2). This study investigates whether persistent difficulty with grammatical gender often observed in the speech of otherwise highly proficient L2 learners is best characterized as a production-specific performance problem, or as difficulty with the retrieval of gender information in real-time language use. In an experimental design that crossed production/comprehension and online/offline tasks, highly proficient L2 learners of Spanish performed at ceiling in offline comprehension, showed errors in elicited production, and exhibited weaker use of gender cues in online processing of familiar (though not novel) nouns than native speakers. These findings suggest that persistent difficulty with grammatical gender may not be limited to the realm of language production, but could affect both expressive and receptive use of language in real time. We propose that the observed differences in performance between native and non-native speakers lie at the level of lexical representation of grammatical gender and arise from fundamental differences in how infants and adults approach word learning.

Keywords
grammatical gender, L2, Spanish, online processing, eye-tracking, production, ultimate attainment

Corresponding author:
Theres Grüter, University of Hawai'i, Department of Second Language Studies, 1890 East-West Road, Honolulu, HI 96822, USA
Email: theres@hawaii.edu

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

I Introduction

Grammatical gender has played a prominent role in research on second language (L2) acquisition and processing. This is not surprising given that grammatical gender is a widespread linguistic phenomenon, is captured relatively easily with a limited set of descriptive statements or rules, yet at the same time, appears to constitute a major challenge when it comes to mastering a non-native language. The persistent difficulty with aspects of grammatical gender often observed in L2 acquisition differs sharply from observations on first language (L1) acquisition, where difficulty with grammatical gender is rare, at least in typically developing populations (e.g. Carroll, 1989; Pérez-Pereira, 1991). Why is it that the same linguistic property seems to be a snap for one group of learners, i.e. young children, but not for others, namely adults learning a non-native language?

Different answers have been proposed, including age-related insensitivity to grammatical features not instantiated in the L1 (Franceschina, 2005; Hawkins, 2009; Sabourin and Stowe, 2008), performance limitations specifically affecting L2 production (Alarcón, 2011; Montrul et al., 2008), and more general limitations on L2 processing due to increased demands on working memory (Gillon Dowens et al., 2010; Keating, 2009, 2010) or to different language learning environments (Lew-Williams and Fernald, 2010). Numerous studies have reported comparisons between native and non-native speakers on tasks manipulating grammatical gender, with varied outcomes, and conclusions that are not entirely convergent. Yet given the many differences between these studies in terms of learner characteristics (e.g. proficiency and L1), and the wide range of methodologies employed, it is virtually impossible to compare outcomes directly. This makes it difficult to determine what aspects of language use involving grammatical gender, if any, may be susceptible to persistent differences between native and (highly proficient) non-native speakers. To understand better the nature of ultimate attainment in the domain of grammatical gender in L2 acquisition, research is needed that employs a wider variety of experimental tasks including both expressive and receptive, as well as on- and offline measures *with the same group of learners*, and that rigorously controls for learner characteristics, such as L2 proficiency and L1 background.

The study we present here employs three different experimental measures targeting grammatical gender, including offline sentence–picture matching (Experiment 1), elicited production (Experiment 2), and online processing using an eye-tracking paradigm (Experiment 3), as well as three independent measures of proficiency (self-rating, a written cloze test, and an automated oral language assessment). All three experiments were conducted with the same two groups of participants: native Spanish speakers (n = 19) and, given the focus of this study on ultimate attainment, the most highly proficient native English-speaking L2 learners of Spanish that we could recruit (n = 19). This design will yield a clearer understanding of the nature of potentially persistent L1–L2 differences in the realm of grammatical gender. Identifying and delineating such differences is a prerequisite for understanding the mechanisms that may underlie them. After presenting results of these three experiments, we consider implications of their combined outcomes for understanding the nature of L1–L2 differences pertaining to grammatical

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

gender, and outline an account focusing on potential differences in the lexical representation of gender, arising as a result of fundamental differences in L1 vs. L2 learning.

II Grammatical gender

Many languages categorize nouns into subclasses according to phonological, morphological, semantic, or, in many cases, mostly arbitrary criteria (Corbett, 1991). In Spanish, the target language of the present study, all nouns are assigned to either the feminine or the masculine noun (or gender) class. While biological gender plays a role in this assignment for nouns denoting human referents, semantic criteria are largely irrelevant in Spanish for inanimate nouns (Harris, 1991). In many cases, morpho-phonological properties of the noun can provide a cue to its gender, with nouns ending in *-o* overwhelmingly belonging to the masculine, and those ending in *-a* to the feminine class (Teschner and Russell, 1984). Importantly, however, this cue is: (i) not fully reliable: masculine nouns ending in *-a* (e.g. *el fantasma* 'the-MASC ghost') and feminine nouns ending in *-o* (*la mano* 'the-FEM hand'), albeit rare, do exist; and (ii) often absent, as in the large number of nouns ending in *-e* or a consonant, which may belong to either gender class.

The only consistent and fully reliable cue to a noun's gender class is distributional, consisting of its co-occurrence relations with transparently gender-marked modifiers, such as determiners and (attributive and predicative) adjectives, illustrated in (1).

(1) a. el árbol pequeño
the-MASC tree(-MASC) small-MASC
'the small tree'

b. la llave antigua
the-FEM key(-FEM) old-FEM
'the old key'

c. El árbol es pequeño.
the-MASC tree(-MASC) is small-MASC
'The tree is small.'

d. La llave es antigua.
the-FEM key(-FEM) is old-FEM
'The key is old.'

While membership in a gender class, or 'gender assignment', is a lexical property of nouns, gender-marking on determiners and adjectives is a derivative property that depends on the noun they modify. The descriptive generalization that determiners and adjectives must be marked for the same gender as the noun they are associated with is known as 'gender agreement'. There is general consensus within the syntactic literature that gender agreement is best captured in terms of checking relations between the lexical gender feature on the noun, and functional features on determiners and adjectives (Bernstein, 1993; Carstens, 2000; Chomsky, 1995; Ritter, 1993). Specific proposals differ on the precise nature of the functional categories presumed to be involved in this phenomenon, as well as on the exact mechanisms of feature checking. As none of these differences affect the research questions in this study, we remain agnostic on these distinctions for our present purpose, and refer the reader to Franceschina (2005) for a comprehensive overview of the syntax of gender agreement in the context of L2 acquisition.

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

III Learning grammatical gender

For the purpose of language learning, acquiring grammatical gender involves acquiring both lexical and syntactic knowledge. 'Mastery' of grammatical gender involves both kinds of knowledge, but also the ability to consistently access and deploy this knowledge in various contexts of language use. An L2 learner's difficulty with grammatical gender in a specific context of language use could thus stem from a number of different sources: (i) difficulty at the level of gender assignment (lexical knowledge); (ii) difficulty at the level of gender agreement (syntactic knowledge); (iii) or difficulty with accessing and/or deploying this lexical and/or syntactic knowledge within the real-time constraints imposed by the specific context of use.

Previous studies investigating the acquisition of grammatical gender by L2 learners differ substantially with regard to the relative emphasis they place on each of these three aspects. In research taking a generative perspective, the main focus has been on gender agreement, that is, syntactic knowledge. The key question within this paradigm has been whether L2 learners whose L1 does not instantiate grammatical gender can acquire the formal gender features that drive agreement relations in the syntax. Given that evidence from language production is often difficult to interpret with regard to the presence or absence of functional features (e.g. Lardiere, 1998), studies in this tradition have typically investigated learners' performance on carefully constructed offline comprehension tasks, where good performance is argued to be possible only if the learner's grammar includes the relevant functional feature (e.g. Alarcón, 2011; McCarthy, 2008; Montrul et al., 2008; White et al., 2004). These studies have shown that while variability of the kind observed in production may extend to comprehension in the case of intermediate-proficiency learners (McCarthy, 2008), advanced-proficiency L2 learners (including those in McCarthy, 2008) perform at ceiling on these tasks, regardless of whether their L1 has gender.

Such results have led to the conclusion that the functional gender feature in question, and thus gender agreement, is acquirable in L2, and that any remaining errors with grammatical gender in production must be due to a production-specific performance problem. This conclusion aligns with the Missing Surface Inflection Hypothesis (MSIH; Prévost and White, 2000; see also Haznedar and Schwartz, 1997), which holds that errors with inflectional morphology in spoken production do not reflect the absence or deficiency of the corresponding functional feature(s) in the underlying syntactic representation. Instead, such errors are attributed to a failure of appropriate vocabulary insertion at the level of morphology. This difficulty is expected to arise specifically during oral production, when learners have to actively select and insert inflectional morphology while they speak. Considerable emphasis has been and continues to be placed on spoken production as the locus of difficulty under the MSIH, as illustrated by a recent summary of this position by White (2011: 585, our emphasis):

> [P]roponents of the Missing Surface Inflection Hypothesis argue that L2ers appropriately represent features at an abstract level, attributing failure to produce consistent inflection to temporary difficulties in accessing the relevant lexical items by which inflection is realized, *particularly when speaking*.

This emphasis has led to L2 studies on a variety of inflectional properties, contrasting learners' performance on production vs. comprehension tasks, where a dissociation

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

between the two in favor of comprehension is typically interpreted as support for the MSIH (for grammatical gender, see for example Alarcón, 2011). However, it appears that a potentially important confound has been disregarded: the comprehension tasks used in the vast majority of these studies were not time constrained. In contrast, spoken language production occurs in real time by its very nature, and is thus always affected by the pressures of real-time processing. It therefore remains unclear whether persistent difficulty with grammatical gender in production is really a production-specific problem, or whether it might be a result of difficulty with the retrieval of gender information in real-time language use (both expressive and receptive) more generally.

The retrieval and processing of grammatical gender cues in real-time have been investigated extensively within a psycholinguistic research tradition. Both child and adult native speakers of gender-marking languages make use of gender cues to facilitate online processing, and show sensitivity to violations of gender agreement in a variety of experimental paradigms (Bates et al., 1996; Dahan et al., 2000; Grosjean et al., 1994; Lew-Williams and Fernald, 2007; Wicha et al., 2004). Research with non-native speakers has been less conclusive. It appears that in cases where learners' L1 and L2 make similar gender-based distinctions, L2 learners show sensitivity to gender violations comparable to that observed in native speakers (Foucart and Frenck-Mestre, 2011; Sabourin and Stowe, 2008). Results from learners whose L1 does not encode grammatical gender, the scenario of interest in the present study, have been more mixed. Some authors have reported sensitivity to violations of gender agreement even in learners with relatively limited proficiency (Foote, 2011; Sagarra and Herschensohn, 2010; Tokowicz and MacWhinney, 2005). Keating (2009) similarly reports that advanced L2 learners of Spanish were sensitive to violations of gender agreement in a self-paced reading experiment, yet only for agreement mismatches on attributive adjectives immediately adjacent to the noun. Unlike native speakers, Keating's advanced L2 learners did not show sensitivity to agreement mismatches on predicative adjectives that are more distant from the corresponding noun. Similarly, Gillon Dowens et al. (2010) observed different ERP signatures in native vs. near-native L2 speakers of Spanish for sentences with agreement mismatches on predicative adjectives. By contrast, they observed no differences between their L1 and L2 groups for sentences containing a gender mismatch between a determiner and an immediately following noun. These findings suggest that advanced L2 learners may perform like native speakers in their sensitivity to violations of gender agreement between adjacent words, indicating that gender agreement per se is operative in advanced L2 learners' real-time language use. However, sensitivity to violations of gender agreement is apparently more susceptible to disruption in L2 learners when mismatches are non-local, suggesting that non-linguistic factors such as working memory may play a role (for further evidence on the contribution of working memory, in both non-native as well as native speakers, see also Keating, 2010).

The psycholinguistic studies reviewed so far all relied on experimental paradigms involving ungrammatical sentences, testing whether learners are sensitive to mismatches in gender marking between nouns and adjectives. While mismatch paradigms have dominated psycholinguistic investigations of grammatical gender in L2 acquisition, few studies have looked at potentially facilitative effects of gender cues on lexical access in fully grammatical sentences. Working with L2 learners with an average exposure to French of more than 20 years, Guillelmon and Grosjean (2001) asked whether a gender-marked

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

determiner preceding a noun would decrease response latencies when participants were asked to repeat that noun. Native speakers and L2 learners exposed to French since childhood were indeed faster to repeat nouns preceded by gender-marked determiners (e.g. *le joli bateau* 'the-MASC nice boat') than those without any preceding gender-marking (e.g. *leur joli bateau* 'their nice boat'). By contrast, the L2 learners who were not exposed to French until their teenage years or later showed no processing advantage on gender-marked trials, despite (self-reported) high proficiency in spoken French. In a related study, Lew-Williams and Fernald (2010) used an eye-tracking procedure to investigate whether L2 learners of Spanish would take advantage of gender-marked determiners in identifying a referent in a two-picture visual display. Participants saw two objects referred to by nouns of either the same (*la pelota – la galleta* 'the-FEM ball – the-FEM cookie') or different gender (*la pelota – el zapato* 'the-FEM ball – the-MASC shoe'), while listening to a sentence asking them to identify one of them (*¿Dónde está la pelota?* 'Where is the-FEM ball'). They found that native speakers were faster to orient to the target picture on different-gender trials, where the gender-marked determiner constituted an informative cue, than on same-gender trials, where no prenominal cues were available. For the L2 group, no facilitative effect was observed.

While the absence of a facilitative effect in the L2 group studied by Lew-Williams and Fernald (2010) may be attributable to these learners' limited proficiency, such an explanation seems less likely for Guillelmon and Grosjean's (2001) late bilinguals, who had been immersed in the L2 for an average of 24 years. Instead, the generalization emerging from these findings may be that while advanced L2 learners can achieve sensitivity to gender mismatches, the use of gender-marking on determiners to facilitate retrieval of a noun is an ability in which L2 learners consistently differ from native speakers. If this generalization is correct, even highly proficient learners would not be expected to show a facilitative effect on different-gender trials in Lew-Williams and Fernald's experiment.

To the extent that the different strands of previous research on grammatical gender in L2 acquisition can be summarized, the findings that have emerged from studies of learners with advanced proficiency seem to converge on the following generalizations:

- Advanced-proficiency learners demonstrate sensitivity to gender agreement in offline comprehension tasks.
- Advanced-proficiency learners demonstrate sensitivity to violations of gender agreement in online processing, at least for agreement relations within a local domain.
- Advanced-proficiency learners do not appear to be able to use gender marking as a facilitative cue in processing grammatical sentences online.
- Advanced-proficiency learners continue to make occasional gender-related errors in spoken production.

It is important to bear in mind, however, that these generalizations are based on the findings of separate studies with different learner groups. This makes it difficult to determine whether the specific weaknesses with spoken production and use of gender cues in online processing observed *between* studies also co-occur *within* the same learner group. The present study is the first to investigate grammatical gender in spoken production and

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

Table 1 Summary of experiments in the present study

	Offline	Online
Production	–	Experiment 2: Elicited production
Comprehension	Experiment 1: Sentence–picture matching	Experiment 3: Eye-tracking

online processing concurrently within the same learner group. Such combined evidence is required in order to determine whether the persistent difficulty with grammatical gender observed in the speech of highly proficient L2 learners is best characterized as a production-specific performance problem in the spirit of the MSIH, or as difficulty with the retrieval of gender information in real-time language use more generally. Under a MSIH account, we would expect to see differences between native and highly proficient non-native speakers in production (Experiment 2), but, crucially, no differences are predicted for receptive language use, whether assessed offline (Experiment 1) or online (Experiment 3). Alternatively, if difficulty is due to problems with the retrieval of gender information in real-time language use, between-group differences are expected in both spoken production and online processing, but not in offline comprehension. Thus the key question we seek to address here is whether L1–L2 differences pattern along a production/comprehension or an online/offline divide when both these factors are crossed, as summarized in Table 1.[1]

IV The present study

1 Participants

Nineteen native English-speaking L2 learners of Spanish (M = 42 years, SD = 9.0) and 19 native Spanish-speaking adults (M = 32 years, SD = 5.7) participated in three experiments. At the time of testing, all participants were residing in California. Participants in the L1 group were born in a Spanish-speaking country (Argentina, Chile, Mexico, Peru, or Spain), raised in Spanish-only home environments, and received most of their education up to the secondary school level in Spanish. Participants in the L2 group were raised in English-only home environments, with no exposure to Spanish or other gender-marking languages before the age of 11 years (mean age of first exposure = 16 years, range = 11–25). Criteria for inclusion in the L2 group were an overall score within the top two tiers of the Versant Spanish Test (Pearson, 2009; described below), indicating advanced to near-native oral proficiency, as well as self-report of daily use of Spanish. Thirteen of the 19 participants in the L2 group indicated that they were currently using Spanish in a professional capacity, as a translator, interpreter, or language teacher. The participants included in this final sample represent the most highly proficient late L2 learners of Spanish that we could recruit in the wider San Francisco Bay area.

Participants completed all three experiments as well as the written proficiency measure during a single, approximately 90-minute lab visit. All participants completed the tasks in the same order:

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

1. looking-while-listening (Experiment 3);
2. cloze test (written proficiency measure);
3. elicited production (Experiment 2); and
4. sentence–picture matching (Experiment 1).

Prior to the lab visit, participants were asked to complete a web-accessible language background questionnaire, and they were asked to take the oral proficiency test by phone (see below for details).

2 Proficiency measures

Three measures of Spanish proficiency were administered to participants in both groups:

1. self-rating;
2. a written cloze test; and
3. the Versant Spanish Test (Pearson, 2009).

For the self-rating measure, participants were asked to rate their proficiency in four domains of language use (speaking, understanding, reading, writing), as well as their overall ability in Spanish, on a scale of 0 to 10. The cloze test, used in several previous studies on L2 Spanish (McCarthy, 2008; Montrul et al., 2008; White et al., 2004), consisted of two short texts in which a total of 50 phrases had been replaced with blanks. The participants' task was to select the correct word or phrase for each blank from a choice of four. Finally, in the Versant Spanish Test – a commercially available assessment tool – participants' Spanish skills were assessed during a 20-minute phone call with an automated speech-recognition system. Test-takers were prompted to respond orally to a variety of tasks in Spanish. Responses were analysed automatically, yielding an overall score as well as four subscores: Sentence Mastery, Vocabulary, Fluency, and Pronunciation (for further detail on test characteristics, see Pearson, 2009). A minimum overall score of 69 (out of 80), indicating performance in the top two tiers as defined by the test developers, was set as an inclusionary criterion for participation in this study.

Results on these proficiency measures, summarized in Table 2, showed near-ceiling performance in both the L1 and L2 groups. Yet despite the overall high proficiency scores in both groups, L2 participants showed slightly lower proficiency than the L1 participants. Non-parametric Mann–Whitney tests indicated significant differences between the native and the non-native speakers with regard to self-rating of their expressive and receptive abilities, performance on the written cloze test, as well as oral fluency as measured by the Versant test (all $U < 80$, $p < .01$). Overall scores on the Versant test differed marginally between groups ($U = 120.0$, $p = .08$), while no significant difference emerged for the vocabulary subscale ($U = 165.0$, $p = .66$).

In sum, results from three independent measures of proficiency indicated that the participants in the L2 group had highly advanced to near-native ability across various domains of language use in Spanish. Yet despite our best efforts to recruit the most highly proficient, late L2 learners of Spanish, their performance on several measures remained slightly but significantly below that of native speakers, an observation

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

Table 2 Means (and ranges) of scores on proficiency measures in both groups, and significance of between–group comparisons (Mann–Whitney)

	L1 group (n = 19)	L2 group (n = 19)	p-value
Self-rating (speaking)[a]	9.9 (9–10)	8.5 (6–10)	< .01
Self-rating (understanding)[a]	9.9 (9–10)	8.6 (6–10)	< .01
Cloze test[b]	48.6 (46–50)	44.9 (37–49)	< .01
Versant (overall)[c]	79.9 (79–80)	76.8 (69–80)	.08
Versant (fluency)[c]	79.0 (69–80)	71.1 (49–80)	< .01
Versant (vocabulary)[c]	78.3 (71–80)	76.8 (58–80)	.66

Notes: [a] On a scale of 0 to 10; [b] Out of 50; [c] On a scale of 20 to 80

noteworthy in itself, and an important consideration in interpreting the experimental results.[2]

3 *Experiment 1: Sentence–picture matching*

a Method: Experiment 1 constitutes a replication of Montrul et al. (2008, Experiment 1), using a task originally developed by White et al. (2004).[3] Taking advantage of the fact that Spanish allows nouns to be omitted from a noun phrase when their referent is salient in the discourse (nominal ellipsis, or N-drop; Bernstein, 1993; Snyder et al., 2001; see (2)), this task was designed to investigate whether learners can identify the referent of a null noun in such a construction solely on the basis of gender (and number) marking on the determiner and/or the adjective. For example, the modifier *otra* ('other') in (2) is morphologically marked for feminine gender (and singular number), thus constraining the choice of referent to those denoted by feminine (singular) nouns.

(2) Tenemos que buscar otra
must-1-PL COMP find other-SG-FEM
'We must find another (one).'

As White et al. (2004: 116) argued, 'this task provides a means of determining, via comprehension rather than production, whether abstract [gender and number] features are present in learner grammars.'

On 32 experimental trials, participants were presented with a written sentence containing a noun-drop construction as in (2), with gender (and number) marked on the determiner and/or the adjective, together with a choice of three pictures, each labeled with a bare noun. Only one of these nouns was of the gender class that matched the gender marked on the modifier(s) in the sentence. For example, the choice of objects in the case of (2) consisted of a key (showing the written label *llave*), a jacket (*abrigo*) and a watch (*reloj*), only the first of which is consistent with the feminine gender-class information on the modifier *otra*. Half of the target nouns were masculine, half were feminine, and both gender conditions included singular as well as plural targets. As the number manipulation is not relevant here, we follow Montrul et al. (2008) in collapsing singular

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

and plural targets for analysis. An additional 13 filler trials consisted of sentences without noun-drop, also accompanied by a choice of three labeled pictures.

The task was presented via a web interface. Participants were instructed to read sentences that were part of an overheard conversation between two people packing suitcases for a trip, and to guess what the people were talking about. The experimenter provided feedback on three practice items, which did not include noun-drop constructions. No feedback was provided during the experimental phase, where each trial was presented on a separate screen, and participants could advance to the next trial at their own pace after selecting one of the three items. As in Montrul et al. (2008), experimental items were presented to all participants in the same semi-randomized order.

b Results: Both groups performed at ceiling on this task, indicated by a mean accuracy of 98% (SD = 4.5) in the L1 and 96% (*SD* = 7.5) in the L2 group. A Mann–Whitney test indicated no significant between-group differences ($U = 138.8$, $p > .2$). As expected under both hypotheses, the highly proficient L2 learners in this study performed like native speakers on this offline receptive measure. These results replicate previous findings by White et al. (2004) and Montrul et al. (2008), confirming that advanced-proficiency learners can show native-like performance on an offline comprehension task targeting gender agreement, an outcome that may be interpreted (following White et al. 2004) as learners having successfully acquired the abstract gender feature posited in linguistic theory.

4 Experiment 2: Elicited production

a Method: The goal of this task was to assess gender assignment and agreement in participants' spoken production by eliciting determiner–noun–adjective sequences, thus providing reflections of gender-marking on both determiners and adjectives. Modeled after Montrul et al. (2008, Experiment 3), the items in this task consisted of 50 Spanish nouns (25 feminine, 25 masculine). Within each gender class, equal numbers of nouns with transparent endings (masculine: *-o*, feminine: *-a*), non-transparent endings (*-e*, –[consonant]), and irregular endings (masculin: *-a*, feminine: *-o*) were included.[4] However, unlike in the Montrul et al. experiment, linguistic stimuli were presented exclusively in the auditory modality. On each trial, participants saw two images of the target noun, which contrasted in at least one salient dimension (e.g. color). For example, for the target noun *mariposa* ('butterfly'), an image of a red butterfly and an image of a yellow butterfly were presented. Participants were then asked to make a choice between the two pictures by naming one of them (*¿Cuál mariposa prefieres?* 'Which butterfly do you like better?'). Elicitation questions were constructed using the determiners *cuál* and *qué* ('which'), which do not inflect for gender. Thus no gender agreement was present in the questions. The 50 target nouns were interspersed with 15 distractors, consisting of people, places or activities (e.g. *¿Cuál muchacha crees que es más simpática?* 'Which girl do you think is nicer?'). The goal of the distractors was to focus participants' attention on the reasons for their choices, rather than the form of their responses. Questions were prerecorded by a female native speaker of Spanish. There were no time constraints for

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

participants to provide a response. When the participant had completed his or her response, the experimenter manually initiated the next trial. Items were presented in the same order to all participants. Responses were audio recorded and subsequently transcribed and coded for analysis.

Prior to analysis, responses on two items (*radio* 'radio' and *modelo* 'model') were eliminated, due to confounds for assessing the correctness of gender assignment and agreement.[5] Responses to the remaining 48 experimental items were coded for the presence/absence of (i) a determiner, (ii) a noun, and (iii) an adjective. Responses not containing either the target noun or a noun-drop construction were excluded from further analysis. Also excluded were responses without a determiner and responses without an adjective. This led to the exclusion of 411/1,824, or 22.5%, of the data overall (22% for the L1 group, 23% for the L2 group). The remaining 1,413 responses (77.5%) all minimally contained a determiner and an adjective, showing that the experimental method employed here was generally effective in eliciting the targeted constructions. Although experimental stimuli were designed to avoid the elicitation of adjectives that do not inflect for gender (e.g. *grande* 'big'), 79 responses contained adjectives of this type (L1: 41, L2: 38). As these responses do not allow for an investigation of gender agreement, they were also excluded. Thus, final analyses were based on 1,334 responses (L1: 670, 70.5%; L2: 664, 69.9%), all of which minimally contained both a determiner and an adjective overtly inflected for grammatical gender.

Responses were coded for accuracy of (i) determiner–noun agreement and (ii) determiner–adjective agreement. This coding yielded four logically possible categories of responses, illustrated in Table 3.

Type (a) represents a correct response, while (b)–(d) constitute possible error types. Errors of type (b) consist of a correct determiner but incorrect adjective. On the assumption that determiner choice is the most immediate reflection of a noun's lexical gender (Carroll, 1989), this error type has been taken as evidence that the speaker correctly classified the noun with regard to its gender class, but failed to compute or express gender agreement on the adjective (e.g. Dewaele and Véronique, 2001; Montrul et al., 2008). Alternatively, the noun may have been misclassified, and agreement computed correctly with the adjective but incorrectly with the determiner. Under both scenarios a failure of agreement is involved, either with the adjective or the determiner. Thus we follow previous work in calling these errors of gender agreement, and assume that they reflect a problem at a syntactic level. Errors of type (c), where determiner and adjective agree with each other but do not instantiate the gender required by the noun, are interpreted most parsimoniously as errors of gender assignment, assuming that speakers have misclassified the noun with regard to its gender class, and agreement has taken place as expected given the lexical (mis)classification of the noun.[6] Following previous authors, we thus take this error to be a reflection of a problem at the lexical level. Finally, errors of type (d), with an incorrect determiner but what appears to be correct agreement between the noun and the adjective, are somewhat difficult to interpret. One might consider them errors of both gender assignment and agreement, consistent with the interpretation of types (b) and (c). However, the correct agreement between noun and adjective may be reflective of both correct gender assignment and agreement, while the

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

Table 3 Possible response types in Experiment 2

	Det–N match	Det–N mismatch
Det–Adj match	Target a. *la (mariposa) roja*	Assignment error c. *el (mariposa) rojo*
Det–Adj mismatch	Agreement error b. *la (mariposa) rojo*	(Both/?) d. *el (mariposa) roja*

incorrect choice of the determiner was due to some other performance factor. In view of this difficulty, together with the low incidence of this error type in our data, we do not analyse errors of this type further.

b Results: Table 4 presents the distribution of response types in both groups. As expected, errors of any kind were exceedingly rare in the L1 group, where all but 9 of 670 responses (mean over participants = 98.7%) were of type (a). In the L2 group, by contrast, significantly fewer responses were of type (a) (80.0%, Mann–Whitney U = 15.0, $p < .001$), with most of the remaining responses (114/664, 17.2%) constituting assignment errors: type (c). These assignment errors occurred more often among irregularly marked nouns (masculine: *-a*, feminine: *-o*; 33.7% assignment errors) than among those with transparent (masculine: *-o*, feminine: *-a*; 12.3%) and non-transparent (*-e*, –[consonant]; 17.0%) endings, and they were somewhat more frequent for feminine (20.4%) compared to masculine (14.4%) targets ($\chi^2 = 4.21$, $p < .05$). Agreement errors – type (b) – on the other hand, were rare in both groups (L1:.3%, L2: 1.5%), with no significant between-group differences ($U = 132.0$, $p > .1$). Errors of type (d) were non-existent in the L1 group, and rare in the L2 group (8/670, 1.3%).

These findings indicate that difficulty with grammatical gender in production persists in L2 learners with advanced to near-native levels of proficiency. Interestingly, we observed an asymmetry between assignment and agreement errors in the L2 group, with more than 10 times as many cases of assignment (114/664) compared to agreement errors (10/664). The observed asymmetry suggests that the persistent difficulty with grammatical gender experienced by highly proficient L2 learners is more likely attributable to lexical, rather than syntactic properties of gender. We return to this observation in the discussion.

5 Experiment 3: Looking-while-listening

a Method: Experiment 3 employed the looking-while-listening procedure (Fernald et al., 2008). Participants viewed pairs of pictures projected onto a screen while listening to sentences naming one of the pictures. Participants' eye movements were videotaped (sampling every 33 ms), with a digital time-code time-locked to the auditory stimuli. Using custom software, eye movements were coded offline, frame-by-frame, enabling a precise time-course analysis of participants' looking patterns as the speech signal unfolds. This experiment was constructed by combining materials from two previous

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

Table 4 Response types by group in elicited production (overall counts, means by participants, standard deviations)

	Det–N match	Det–N mismatch
Det–Adj match	A. Target L1: 661/670 (98.7%, 1.7) L2: 532/664 (80.0%, 14.6)	C. Assignment error L1: 7/670 (1.0%, 1.6) L2: 114/664 (17.2%, 12.8)
Det–Adj mismatch	B. Agreement error L1: 2/670 (0.3%, 0.9) L2: 10/664 (1.5%, 2.3)	D. Both/? L1: 0/670 L2: 8/670 (1.3%, 2.6)

experiments, reported in Lew-Williams and Fernald (2010: Experiments 1 and 3), which we treat here as two experimental conditions: the familiar-noun and the novel-noun condition.

The purpose of this experiment was to test whether gender-marking on the determiner facilitates interpretation of the following noun. Speech stimuli in the familiar-noun condition consisted of simple Spanish sentences starting with one of two carrier frames – *Encuentra ...* ('Find …') or *¿Dónde está ...?* ('Where is …?') – followed by one of eight determiner–noun sequences, four masculine and four feminine: *la pelota* ('ball'), *la galleta* ('cookie'), *la vaca* ('cow'), *la rana* ('frog'), *el zapato* ('shoe'), *el carro* ('car'), *el pájaro* ('bird'), and *el caballo* ('horse'). These sentences were recorded by a female native speaker of Spanish, with sound files edited to control for the duration of the carrier phrase (M = 914 ms, range = 900–31), determiner (M = 280 ms, range = 268–99), and noun (M = 720 ms, range = 670–770). Visual stimuli were colorful digital images of animals and objects on gray backgrounds. Each picture served as target on four trials and as distracter on four trials, with side of target picture presentation counterbalanced.

These stimuli were presented in two trial types: on same-gender trials the two images depicted objects referred to by nouns of the same grammatical gender (e.g. *la pelota, la galleta*), while on different-gender trials images depicted objects referred to by nouns of different gender (e.g. *la pelota, el zapato*). In the latter case, the gender-marked determiner served as a potentially informative cue to the object being named. On same-gender trials, in contrast, no disambiguating information appeared before the onset of the noun itself. Thus, if listeners process gender-marking on determiners incrementally, we expect faster shifting to the target image on different- than on same-gender trials, as was observed in previous experiments using these same stimuli with adult as well as 3-year-old native speakers of Spanish (Lew-Williams and Fernald, 2007, 2010). Here we asked whether highly proficient L2 learners would demonstrate the same processing advantage.

Interspersed with trials in the familiar-noun condition were trials in the novel-noun condition. This condition was included to control for the amount of previous experience participants in both groups had with each noun. While adult native speakers have typically heard familiar nouns such as *la pelota* many thousands of times, it appears safe to assume that L2 learners have heard fewer instances of them overall. As cumulative frequency of exposure is known to affect speed of lexical access, differences in the amount

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

of previous experience with familiar nouns may contribute towards differential processing efficiency in the L1 vs. L2 group. The goal of the novel-noun condition was to even the playing field in this regard. During the first part of the experiment, participants were exposed to four novel nouns, each paired with a novel object. Each novel noun-object pair was presented five times, for a total of 20 teaching trials. On these teaching trials, speech stimuli consisted of a simple sentence frame (*¡Mira, es ...!* 'Look, it's …!') followed by a novel noun preceded by an indefinite determiner (*una catela, una pifa, un durino, un tebo*). The indefinite determiner (*un*/*una*), together with canonical noun endings (*-o*/*-a*), marked two of these nouns as masculine and two as feminine. Visual stimuli consisted of a single image of the respective novel object. During the second part of the experiment, participants were exposed to 24 novel-noun test trials, structured analogously to the trials in the familiar-noun condition. Each novel noun served as target on three same-gender and three different-gender trials. Sentences consisted of a frame *¿Dónde está ...?* ('Where is …'), a definite article (*el*/*la*), and a novel noun. Duration of sentence frames, determiners, and nouns used on novel-noun test trials were edited to closely parallel those in the familiar-noun condition. On same-gender trials, visual stimuli consisted of two images depicting objects referred to by novel nouns of the same gender (e.g. *la catela, la pifa*), while on different-gender trials, images depicted objects referred to by novel nouns of different gender (e.g. *la catela, el durino*). Importantly, prior to the test trials, participants never encountered a definite article in combination with a novel noun. Thus, in contrast to the familiar-noun condition, a processing advantage on different-gender trials in the novel-noun condition could not result from learners relying on co-occurrence relations between the definite determiner and the noun. Crucially, if an effect is obtained in the novel-noun condition, this would indicate that speakers have classified the novel nouns into abstract gender classes based on the relevant information provided in the teaching trials, and are drawing on this information when encountering the definite determiner in the test trials.

Immediately prior to the eye-tracking procedure, participants were briefly introduced to the eight familiar object names used in the task. This was done through the experimenter reading each noun in Spanish (without a determiner, e.g. *pelota*), and participants providing both an English translation and indicating their familiarity with this noun on a four-point scale. The purpose of this introductory phase was to increase the likelihood that participants would associate the visual stimuli (e.g. picture of a cow) with the lexical item used in the experiment (e.g. *la vaca*, 'cow', rather than *el toro*, 'bull'). All participants correctly translated all items, and indicated that they were familiar or very familiar with these words. These high familiarity ratings, together with the fact that all nouns used in this experiment had transparent noun endings (masculine: *-o*, feminine: *-a*), also provide strong indication that the gender class of these items was known to participants.

During the eye-tracking phase, participants were presented with a total of 79 trials in one of two counterbalanced lists, including 32 trials in the familiar-noun condition, 20 teaching and 24 test trials in the novel-noun condition, and 3 filler trials.[7] On each trial pictures were visible for 2 s prior to the speech signal, for the duration of the 3 s speech signal, and for 1 s following the speech signal. Trials were separated by an 800 ms interval. The eye-tracking phase of the experiment lasted approximately 8 mins. Highly

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

trained coders, blind to trial type, subsequently viewed the video-record, and indicated for each 33 ms frame whether the subject was looking left, right, between the pictures, or away. Reliability coding was conducted on 21% of all participants. Intercoder agreement within a single frame was 99.4% for native speakers and 98.9% for L2 learners.

Since participants could not know in advance which picture would be named, they were by chance equally likely to be looking at the target or distracter picture at the onset of the determiner. If they were already looking at the correct picture (target-initial trials), they should maintain fixation; but if they were looking at the distracter picture (distracter-initial trials), they should shift quickly to the named picture. Distracter-initial trials were used to calculate reaction time (RT), the latency to initiate an eye movement toward the target picture. RT was calculated from determiner onset, that is, the first moment in the unfolding sentence where participants received potentially relevant acoustic information. Shifts initiated in the first 200 ms were not included in analyses, because they were likely to represent random shifting that occurred prior to the possible influence of the determiner (Allopenna et al., 1998). RTs were included in analyses if they occurred between 200 and 800 ms from determiner onset, a time window that corresponds roughly to the time windows in which significant effects were observed in related adult processing studies (e.g. Dahan et al., 2000).

b Results: The time course of participants' orienting to the target picture on distracter-initial trials is illustrated in Figure 1 for familiar nouns and Figure 2 for novel nouns. Visual inspection of the time course of participants' looking behavior in the familiar-noun condition suggests that native speakers were faster to converge on the target picture on different- compared to same-gender trials, consistent with previous findings by Lew-Williams and Fernald (2007, 2010). In the L2 group, however, the difference in performance between the two trial types appears to be considerably smaller, with the two curves barely diverging from each other. In the case of novel nouns, visual inspection

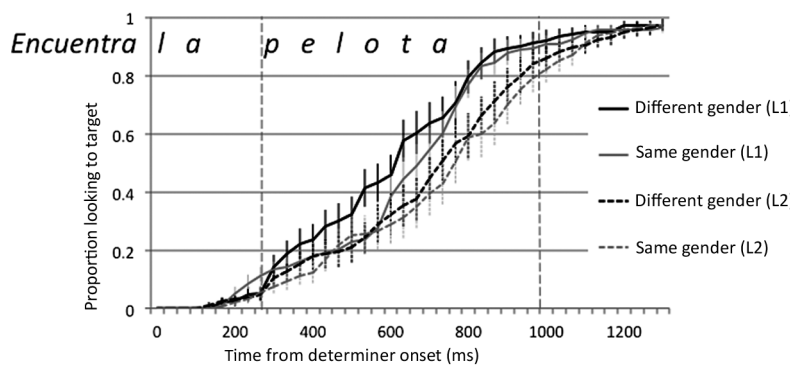

Figure 1 Time course of participants' looking to the target image on distracter-initial trials in the familiar-noun condition

Notes: Vertical dashed lines indicate onset and offset of the noun. Error bars indicate standard errors of the means.

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

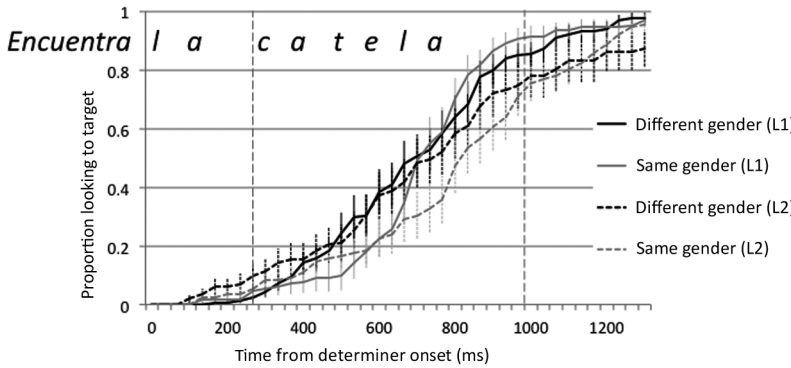

Figure 2 Time course of participants' looking to the target image on distracter-initial trials in the novel-noun condition

Notes: Vertical dashed lines indicate onset and offset of the noun. Error bars indicate standard errors of the means.

indicates first of all that participants in both groups successfully learned the novel nouns, as shown by the convergence of their looks on the target image by approximately one second after the onset of the noun in both conditions. Furthermore, it appears that native speakers were again faster to converge on the target picture on different- compared to same-gender trials, consistent with the findings of Lew-Williams and Fernald (2010). Interestingly, in the L2 group, the two curves also seem to diverge, although this difference occurs at a somewhat later time than in the L1 group.

Figures 3 and 4 show mean RTs on each trial type for familiar and novel nouns. For statistical analysis, mean RTs were entered into a 2 (group) × 2 (noun-type) × 2 (trial-type) mixed ANOVA. This analysis yielded a highly significant main effect for trial type ($F(1,33) = 16.6$, $p < .001$), showing that participants overall responded faster on different- than on same-gender trials. A significant main effect for noun type ($F(1,33) = 4.9$, $p = .03$) also emerged, indicating that familiar nouns were recognized faster overall than novel nouns. However, the main effect for group ($F(1,33) = .1$, $p > .8$) was not significant, nor were any interactions (all $F < 1$, $p > .3$). Thus a full analysis of variance did not reveal any significant differences between the performance of native and non-native speakers on this task. This finding may be interpreted as evidence that native and non-native speakers take equal advantage of gender-marking on determiners during online processing. However, as with any null result, this outcome must be interpreted with caution. Recall that the time-course data in Figure 1 showed no visible effect of trial type in the familiar-noun condition for the L2 group, while for the novel-noun condition in Figure 2, the effect appeared at a later time point in the L2 than in the L1 group. Since possible between-group differences on same- and different-gender trials were of central interest in this research, we next performed planned pairwise comparisons between RTs on same- and different-gender trials for each group and noun type. For the familiar-noun condition, these comparisons indicated a significant difference between same- and different-gender trials for the L1 ($t(18) = 2.8$, $p = .01$, $d = .93$), but not the L2 group

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

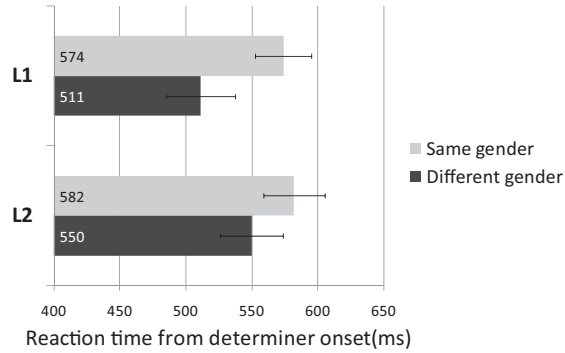

Figure 3 Mean RTs on shifts from distracter to target picture by L1 and L2 participants on same- and different-gender trials in the familiar-noun condition
Note: Error bars indicate standard errors of the means.

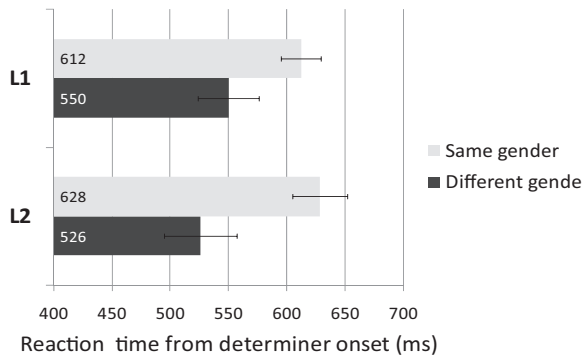

Figure 4 Mean RTs on shifts from distracter to target picture by L1 and L2 participants on same- and different-gender trials in the novel-noun condition
Note: Error bars indicate standard errors of the means.

($t(17) = 1.2$, $p = .26$, $d = .39$). To further explore the potential impact of proficiency on the diminished effect in the L2 group, RTs for the subgroup of L2 learners with a perfect score (80) on the oral proficiency measure (n = 11) were analysed separately. Mean RTs in this subgroup were not different for same- (548 ms) compared to different-gender (556 ms) trials, indicating the absence of a significant effect in the L2 group was not related to within-group differences in proficiency. For the novel-noun condition, pairwise comparisons showed similar differences in RT between same- and different-gender trials in both groups (L1: $t(18) = 2.1$, $p = .05$, $d = .68$; L2: $t(15) = 2.4$, $p = .03$, $d = .86$), with marginally significant p-values in both groups after adjusting for multiple comparisons.

In sum, Experiment 3 confirmed findings by Lew-Williams and Fernald (2007, 2010), demonstrating that native speakers of Spanish use gender-marking on the determiner as a predictive cue in online language processing. However, unlike Lew-Williams and Fernald, who included a group of L2 learners of Spanish with intermediate levels of

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

proficiency, we tested a group of highly proficient L2 learners. We found that their overall performance did not differ from that of native speakers, at least as indicated by the absence of a main effect for group (and any interactions with group). However, planned pairwise comparisons within each group indicated that the difference in mean RT on same- vs. different-gender trials in the familiar-noun condition was significant in the L1 group but not in the L2 group. This suggests that any processing advantage L2 learners gain as a result of gender-marking on determiners preceding familiar nouns is weaker and/or less consistent than what can be observed in native speakers. Interestingly, this qualification does not appear to apply to the novel-noun condition, where marginally significant effects and similar effect sizes were found in both groups, a finding we consider in the discussion.

V Discussion

The goal of the present study was to assess facility with grammatical gender across a variety of experimental tasks, including expressive measures as well as on- and offline receptive measures, in the same group of highly proficient English-speaking learners of Spanish. A research design of this type was necessary in order to tease apart two dimensions often confounded in previous work, namely production vs. comprehension on the one hand, and offline vs. online language use on the other. The key question we addressed was whether any differences between native and non-native speakers would pattern along one dimension or the other. In other words, are advanced learners' difficulties with grammatical gender confined to language production, as argued by researchers who adopt accounts in the spirit of the Missing Surface Inflection Hypothesis (Alarcón, 2011; Montrul et al., 2008; White et al., 2004), or are these difficulties better characterized as problems with the retrieval of gender information in real-time language use, spanning both expressive and receptive domains?

Consistent with both hypotheses, the L2 group performed at ceiling in offline comprehension (Experiment 1), replicating findings by White et al. (2004) and Montrul et al. (2008). Also consistent with both hypotheses, we observed significant differences between the L1 and the L2 group in elicited production (Experiment 2). These differences, however, were confined to errors of gender assignment, which were more frequent in the L2 group. Errors of gender agreement, on the other hand, were rare in both groups, suggesting that persistent difficulty with grammatical gender experienced by highly proficient L2 learners primarily affects lexical, rather than syntactic aspects of gender. Finally, evidence from the online processing of gender-marked determiners was obtained (Experiment 3) in order to distinguish between the two hypotheses. Results, however, were not entirely consistent with either hypothesis. The absence of significant between-group differences and interactions would suggest that L2 learners and native speakers did not differ in their online processing of gender-marking on determiners, as predicted by accounts in the spirit of the MSIH. But, critically, planned comparisons within each group revealed that L2 participants did not process familiar determiner–noun pairs as efficiently as L1 participants, suggesting a weakness in the use of grammatical gender cues in their online processing of familiar nouns. This weakness is unexpected under a MSIH account, yet consistent with the findings of studies showing compromised

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

processing of grammatical gender in real time among non-native speakers. Interestingly, a more robust effect was obtained for the L2 group in the novel-noun condition, indicating that use of gender cues in online processing is not beyond the abilities of L2 learners in principle. We return to differences between the familiar- and the novel-noun condition below, after considering in more detail a potential source of the L1–L2 difference observed in elicited production.

Findings from Experiment 2 suggested that lexical, rather than syntactic, properties of gender may be the primary source of difficulty for advanced L2 learners. The fact that assignment errors were more than 10 times as frequent as agreement errors in this study is striking, but reflective of a similar asymmetry observed in a related study by Alarcón (2011), who also observed more assignment than agreement errors among advanced L2 learners of Spanish. While we are not aware of any theoretical account that would make predictions regarding the relative occurrence of assignment vs. agreement errors, the observed asymmetry may nevertheless be considered surprising (by at least one reviewer, as well as ourselves), given that much of the linguistically oriented literature on the L2 acquisition of grammatical gender has focused primarily on gender agreement, sometimes excluding responses involving potential assignment errors from analysis altogether (e.g. McCarthy, 2008: 472, 474–75). The observation that L2 learners with advanced to near-native proficiency appear to experience more persistent problems with gender assignment than agreement suggests that the lexical representation of grammatical gender, and its acquisition in L1 development, should be reconsidered in some detail.

In traditional descriptions of lexical representations of gender, noun class membership must be stored as an inherent property of each noun. One way of formalizing this is to assume that 'all nouns of a given grammatical gender are linked to one gender node specifying that grammatical gender' (Schriefers and Jescheniak 1999: 577). Of relevance here is how such links or associations between nouns and gender nodes are established in native lexicons during the course of L1 acquisition. Given that phonological and semantic cues alone are insufficient for establishing membership in gender classes in Spanish, learners must rely extensively on co-occurrence relations between nouns and gender-marked modifiers, most importantly determiners, to detect a noun's gender. The computation of co-occurrence relations, and transitional probabilities more generally, has been shown to be a key mechanism in infants' early language learning (Saffran et al., 1996). One indication of children's processing of co-occurrence relations in the case of determiner–noun sequences is their occasional failure to segment these sequences and treat them as unanalysed chunks (e.g. Carroll, 1989; Chevrot et al., 2008). As backward transitional probabilities between determiners and nouns are typically high (see Pelucchi et al., 2009), such misanalyses are not surprising if the computation of co-occurrence relations plays a role in children's early language learning. As a result of this distributional learning, tight associations are formed between frequently co-occurring elements such as determiners and nouns in early L1 lexicons. While at early stages these tight associations between gender-marked determiners and nouns are likely to be lexically specific, it seems reasonable to assume that with increased vocabulary size they will give rise to associations between nouns and more abstract gender classes, as instantiated for example through the gender nodes posited by Schriefers and Jescheniak (1999). Importantly, given the strong initial associations between nouns and their gender-marked

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

modifiers as a result of early distributional learning, associations between nouns and gender nodes in the more mature L1 lexicon can be expected to remain strong, making gender-marking a powerful, well-practised cue for processing in a native language.

We argue that the lexical representation of grammatical gender in the native lexicon is crucially shaped by early distributional learning. To what extent are these same learning mechanisms involved in L2 acquisition? Experimental studies have shown that adults are able to use distributional information to learn properties of an unfamiliar artificial language in laboratory settings (Saffran et al., 1996), although properties of the L1 may impact their ability to detect certain statistical regularities (Finn and Hudson Kam, 2008). Importantly, however, computing distributional information is not the only, and almost certainly not the most efficient way to learn words in a real-world language learning context. Learners beyond early childhood approach the task of language learning with conceptual and linguistic knowledge already instantiated through their L1. This allows them to take advantage of a number of cues that are not available to infants, including parallels between L1 and L2, metalinguistic information, and information specific to written language, such as gaps between words. Given the richness of these information sources, L2 learners are unlikely to rely on the computation of co-occurring elements – such as determiners and nouns – to the same extent as infant L1 learners. As a consequence, the tight associations between determiners and nouns that emerge in the early L1 lexicon are unlikely to arise in developing L2 lexicons. There is good reason to assume, given advanced L2 learners' good performance on offline tasks manipulating gender, that representations of nouns including information on gender-class membership can be established in the L2 lexicon in some way. Crucially, however, the associations between nouns and gender class information are unlikely to attain the same strength in L2 as in L1 lexicons, as a result of L2 learners' reliance on cues other than co-occurrence relations during word learning.

In consideration of these fundamental differences between L1 and L2 learning, we propose that while all learners eventually establish abstract gender categories, a perhaps inevitable difference between native and non-native speakers lies in the strength of associations between nouns and what can be instantiated as gender nodes in the mental lexicon. This difference between L1 and L2 is expected to manifest itself in slower retrieval of gender information in real-time L2 use, leading to occasional errors of gender assignment in production (as observed in Experiment 2), and less effective use of gender cues in online processing (Experiment 3). L2 learners' sensitivity to violations of gender agreement in studies using mismatch paradigms (see Section III above) is compatible with this account: L2 learners do possess the relevant knowledge to detect these violations, although the activation of this knowledge may be more effortful.

The proposed account also yields a possible explanation for the somewhat surprising finding from Experiment 3, where L2 learners appeared to use gender-marking on the determiner as a predictive cue with newly learned novel nouns, but not with familiar nouns. Note that the novel nouns in Experiment 3 had to be learned solely on the basis of exposure to these nouns, preceded by gender-marked (indefinite) determiners, in spoken language during the teaching trials. This learning scenario simulated the word learning context of the infant L1 learner somewhat more closely than the learning context typically encountered by the literate adult L2 learner. Specifically, it enhanced the role of

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

co-occurrence relations and reduced the role of top-down instruction and written cues. We thus speculate that participants in this experiment relied more on distributional cues, leading to strong noun-gender associations in the newly created lexical representations for the novel words. In contrast, they had presumably learned about the gender of familiar nouns with the aid of many non-distributional cues. The increased strength of associations for novel words then allowed L2 learners, we hypothesize, to take advantage of the gender cue on the determiner on different-gender trials in the novel-noun condition.

Independent evidence in support of the proposal that different learning conditions affect learners' ability to associate a noun with its gender class comes from an artificial language learning experiment by Arnon and Ramscar (2009, 2012). In this study, adult learners were better at learning grammatical gender in an artificial language when they were exposed to determiner–noun sequences first, and only then to isolated nouns paired with their referents, as compared to the opposite scenario, where they encountered isolated nouns first, followed by determiner–noun sequences. When learners were initially presented with what Arnon and Ramscar call 'less segmented' input (determiner–noun sequences), they were forced to use distributional information to identify word boundaries before they could map a noun to a referent. By contrast, when they were presented with isolated nouns first, the information on the subsequently encountered determiner–noun sequences added no further cues to facilitate noun-referent mappings, thus leading learners to pay less attention to the determiners in this condition, and consequently fail to learn the gender-class information encoded on them. As these authors pointed out, the isolated-noun-first condition in their experiment can be considered somewhat similar to the language learning situation encountered by adults, who can draw on numerous cues to word boundaries based on their L1 and world knowledge, whereas the opposite scenario – encountering less segmented input first – is more akin to the learning scenario encountered by the infant L1 learner. Crucially, Arnon and Ramscar found better learning of gender-class information in the latter case, thus simulating the findings of a large body of research on the acquisition of grammatical gender in L1 vs. L2 acquisition, and lending further support to our proposal linking these L1–L2 differences to fundamental differences between the cues available to and used by L1 vs. L2 learners.

To sum up: in a research design that experimentally crossed production/comprehension and online/offline factors, we found that highly proficient L2 learners of Spanish performed at ceiling in offline comprehension, continued to commit errors in elicited production (albeit almost exclusively of a lexical nature), and exhibited weaker use of gender-cues in the online processing of familiar (though not novel) nouns than native speakers. These results suggest that persistent difficulty with grammatical gender even in highly proficient L2 learners may not be limited to the realm of language production, but could affect both expressive and receptive use of language in real time. We have presented a proposal linking persistent differences between highly proficient L2 learners and native speakers in the realm of grammatical gender to differences in the strength of associations between nouns and what can be instantiated as gender nodes in L1 vs. L2 lexicons, differences that we argue may be the result of fundamental differences between how infants and adults approach the task of word learning. Future research is needed to test whether our proposal captures relevant differences between L1 and L2 learning. Specifically, systematic manipulations of the cues available during word learning in

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

natural L2 learning contexts are needed to further examine the validity and generalizability of this proposal.

Acknowledgments

Many thanks to all our participants, as well as to Narges Afshordi, Ricardo Bion, Melissa Guevara, Nereyda Hurtado, Virginia Marchman, Christine Potter, Adriana Weisleder, and other staff and students at the Center for Infant Studies at Stanford University, USA. We are grateful for the many helpful comments from the audience at the 35th[h] Boston University Conference on Language Development and the anonymous *Second Language Research* reviewers. This research was supported by the National Institutes of Health (F32 DC010129-01 to Theres Grüter, DC 008838 to Anne Fernald).

Notes

1. Our design does not include an offline measure of production, as we could not conceive of an appropriate task that does not rely heavily on metalinguistic knowledge.
2. As noted by a reviewer, these remaining differences between our highly proficient L2 learners and native speakers are reminiscent of findings by Abrahamsson and Hyltenstam (2009), who concluded that 'absolute nativelikeness in late learners, in principle, does not occur' (p. 294).
3. We thank Silvina Montrul for sharing the stimuli used in Montrul et al. (2008). The reader is referred to the original article for further detail and illustrations of the materials.
4. For the complete list of nouns used in Montrul et al.'s experiment, see Montrul et al. (2008: Appendix B). With a few exceptions due to difficulty with imageability, the same nouns were used here. A more detailed report of the findings from this task, taking into consideration item-level properties, such as transparency of noun-endings and frequency, is currently in preparation. For the purpose of addressing the research questions in the present article, we report only overall accuracy for each group here.
5. In the case of *radio*, there is dialectal variation with regard to gender assignment. While most varieties, including Iberian Spanish, classify radio as feminine, others (e.g. some Mexican varieties) prefer masculine (*el radio*). In the case of *modelo* ('model'), the visual stimuli in the experiment depicted two female fashion models. Most responses to this item consisted of a noun-drop construction (e.g. *la rubia* 'the blonde one'), for which it was difficult to determine whether the choice of gender forms was guided by the grammatical gender of the preceding noun (*modelo*-FEM) or the semantic gender associated with the female referent.
6. Strictly speaking, it cannot be excluded that the noun was classified correctly, and agreement failed both with the determiner and the adjective. Yet if agreement failure results in the insertion of the default form of the modifier (generally agreed to be the masculine in Spanish; Harris, 1991), we would expect, under this scenario, almost exclusively errors of the type Det-MASC N-FEM Adj-MASC, but not Det-FEM N-MASC Adj-FEM. This was not the case in our data, where we found an even distribution of this error type across the two gender classes, providing some evidence against this alternative possibility.
7. A reviewer was concerned that the low number of fillers would not be sufficient to distract participants from the property under investigation. After the completion of Experiment 3, the first task in the test battery, participants were asked whether they could guess what the experiment was designed to investigate. Only one out of all 38 participants indicated that grammatical gender might be at issue, suggesting that the purpose of the experiment was generally not transparent. This finding is consistent with observations from previous experiments using these same stimuli, based on which it was decided not to increase the number of fillers here (Lew-Williams and Fernald, 2007, 2010).

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

References

Abrahamsson N and Hyltenstam K (2009) Age of onset and nativelikeness in a second language: Listener perception vs. linguistic scrutiny. *Language Learning* 59: 249–306.

Alarcón IV (2011) Spanish gender agreement under complete and incomplete acquisition: Early and late bilinguals' linguistic behavior within the noun phrase. *Bilingualism: Language and Cognition* 14: 332–50.

Allopenna PD, Magnuson JS, and Tanenhaus MK (1998) Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language* 38: 419–39.

Arnon I and Ramscar M (2009) Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. In: Taatgen NA and van Rijn H (eds) *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, 2112–17. Available online at http://csjarchive.cogsci.rpi.edu/Proceedings/2009/index.html (February 2012).

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

Bates E, Devescovi A, Hernandez A, and Pizzamiglio L (1996) Gender priming in Italian. *Perception and Psychophysics* 58: 992–1004.

Bernstein J (1993) Topics in the syntax of nominal structure across Romance. Unpublished PhD dissertation, City University of New York, NY, USA.

Carroll S (1989) Second-language acquisition and the computational paradigm. *Language Learning* 39: 535–94.

Carstens V (2000) Concord in minimalist theory. *Linguistic Inquiry* 31: 319–55.

Chevrot J-P, Dugua C, and Fayol M (2008) Liaison acquisition, word segmentation and construction in French: A usage-based account. *Journal of Child Language* 36: 557–96.

Chomsky N (1995) *The minimalist program*. Cambridge, MA: MIT Press.

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

Dahan D, Swingley D, Tanenhaus MK, and Magnuson JS (2000) Linguistic gender and spoken-word recognition in French. *Journal of Memory and Language* 42: 465–80.

Dewaele J-M and Véronique D (2001) Gender assignment and gender agreement in advanced French interlanguage: A cross-sectional study. *Bilingualism: Language and Cognition* 4: 275–97.

Fernald A, Zangl R, Portillo AL, and Marchman VA (2008) Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. In: Sekerina I, Fernández EM, and Clahsen H (eds) *Developmental psycholinguistics: On-line methods in children's language processing*. Amsterdam: John Benjamins, 97–135.

Finn AS and Hudson Kam CL (2008) The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition* 108: 477–99.

Foote R (2011) Integrated knowledge of agreement in early and late English–Spanish bilinguals. *Applied Psycholinguistics* 32: 187–220.

Foucart A and Frenck-Mestre C (2011) Grammatical gender processing in L2: Electrophysiological evidence of the effect of L1–L2 syntactic similarity. *Bilingualism: Language and Cognition* 14: 379–99.

Franceschina F (2005) Fossilized second language grammars: The acquisition of grammatical gender. Amsterdam: John Benjamins.

Gillon Dowens M, Vergara M, Barber HA, and Carreiras M (2010) Morphosyntactic processing in late second-language learners. *Journal of Cognitive Neuroscience* 22: 1870–87.

Grosjean F, Dommergues JY, Cornu E, Guillelmon D, and Besson C (1994) The gender marking effect in spoken word recognition. Perception and *Psychophysics* 56: 590–98.

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

Guillelmon D and Grosjean F (2001) The gender marking effect in spoken word recognition: The case of bilinguals. *Memory and Cognition* 29: 503–11.

Harris J (1991) The exponence of gender in Spanish. *Linguistic Inquiry* 22: 27–62.

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*. London: Continuum International Publishing, 63–78.

Haznedar B and Schwartz BD (1997) Are there optional infinitives in child L2 acquisition? In: Hughes E, Hughes M, and Greenhill A (eds) *Proceedings of the 21st BUCLD*. Somerville, MA: Cascadilla, 257–68.

Keating GD (2009) Sensitivity to violations of gender agreement in native and nonnative Spanish: An eye-movement investigation. *Language Learning* 59: 503–35.

Keating GD (2010) The effects of linear distance and working memory on the processing of gender agreement in Spanish. In: VanPatten B and Jegerski J (eds) *Research in second language processing and parsing*. Amsterdam: John Benjamins, 113–34.

Lardiere D (1998) Case and tense in the 'fossilized' steady state. *Second Language Research* 14: 1–26.

Lew-Williams C and Fernald A (2007) Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science* 18: 193–98.

Lew-Williams C and Fernald A (2010) Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language* 63: 447–64.

McCarthy C (2008) Morphological variability in the comprehension of agreement: An argument for representation over computation. *Second Language Research* 24: 459–86.

Montrul S, Foote R, and 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: 503–53.

Pearson (2009) *Versant Spanish test: Test description and validation summary*. Palo Alto, CA: Pearson Knowledge Technologies. Retrieved from: https://www.ordinate.com/technology/VersantSpanishTestValidation.pdf (February 2012).

Pelucchi B, Hay JF, and Saffran JR (2009) Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition* 113: 244–47.

Pérez-Pereira M (1991) The acquisition of gender: What Spanish children tell us. *Journal of Child Language* 18: 571–90.

Prévost P and White L (2000) Missing surface inflection or impairment in second language acquisition? Evidence from tense and agreement. *Second Language Research* 16: 103–33.

Ritter E (1993) Where's gender? *Linguistic Inquiry* 24: 795–803.

Sabourin L and Stowe LA (2008) Second language processing: When are first and second languages processed similarly? *Second Language Research* 24: 397–430.

Saffran JR, Aslin RN, and Newport EL (1996) Statistical learning by 8-month-old infants. *Science* 274: 1926–28.

Saffran JR, Newport EL, and Aslin RN (1996) Word segmentation: The role of distributional cues. *Journal of Memory and Language* 35: 606–21.

Sagarra N and Herschensohn J (2010) The role of proficiency and working memory in gender and number agreement processing in L1 and L2 Spanish. *Lingua* 120: 2022–39.

Schriefers H and Jescheniak JD (1999) Representation and processing of grammatical gender in language production: A review. *Journal of Psycholinguistic Research* 28: 575–600.

Snyder W, Senghas A, and Inman K (2001) Agreement morphology and the acquisition of noun-drop in Spanish. *Language Acquisition* 9: 157–73.

Teschner RV and Russell WM (1984) The gender patterns of Spanish nouns: An inverse dictionary-based analysis. *Hispanic Linguistics* 1: 115–32.

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012

Tokowicz N and MacWhinney B (2005) Implicit and explicit measures of sensitivity to violations in second language grammar: An Event-Related Potential investigation. *Studies in Second Language Acquisition* 27: 173–204.

White L (2011) Second language acquisition at the interfaces. *Lingua* 121: 577–90.

White L, Valenzuela E, Kozlowska-MacGregor M, and Leung Y-K (2004) Gender and number agreement in nonnative Spanish. *Applied Psycholinguistics* 25: 105–33.

Wicha NYY, Moreno EM, and Kutas M (2004) Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience* 16: 1272–88.

Downloaded from slr.sagepub.com at UNIV OF WISCONSIN on May 23, 2012