Morphological variability in second language learners: an examination of electrophysiological and production data

Article

Accepted Version

Morphological variability in second language learners: An examination of electrophysiological and production data

José Alemán Bañón[a, b], David Miller[c], and Jason Rothman[c, d]

(First author's given name: José; family name: Alemán Bañón)

a: Centre for Research on Bilingualism, Stockholm University

b: Basque Center on Cognition, Brain and Language (BCBL)

c: University of Reading, Department of Psychology and Clinical Language Sciences

d: UiT The Arctic University of Norway, Department of Language and Linguistics

José Alemán Bañón was supported by a postdoctoral fellowship from the Spanish Ministry of Economy and Competitiveness (FPDI-2013-15813). David Miller was supported by a PhD fellowship from the AThEME project (Advancing the European Multilingual Experience), funded by the European Research Council (ERC). The authors thank Ian Cunnings, Robert Fiorentino, Alison Gabriele, Clara Martin, and Nicola Molinaro for their feedback on different aspects of this project. We also thank Laura Domínguez and Pilar Gray-Carlos for their help in recruiting participants, Vince DeLuca for his assistance in data collection, Steve Politzer-Ahles for his help in preparing the manuscript, and Kora Kirby for her help in creating the materials.

Correspondence concerning this article should be addressed to José Alemán Bañón, Centre for Research on Bilingualism, Stockholm University, Universitetsvägen 10 D, Rum B 355, 106 91 Stockholm, Sweden. Email: jose.aleman.banon@biling.su.se

Abstract

We examined potential sources of morphological variability in adult L1-English L2-Spanish learners, with a focus on L1-L2 similarity, morphological markedness, and knowledge type (receptive vs. expressive). Experiment 1 uses event-related potentials to examine noun-adjective number (present in L1) and gender agreement (absent in L1) in online sentence comprehension (receptive knowledge). For each feature, markedness was manipulated, such that half of the critical noun-adjective combinations were feminine (marked) and the other half, masculine; half were used in the plural (marked) and the other half in the singular. With this set-up, we examined learners' potential overreliance on unmarked forms or "defaults" (singular/masculine). Experiment 2 examines similar dependencies in spoken sentence production (expressive knowledge). Results showed that learners (n=22) performed better with number than gender overall, but their brain responses to both features were qualitatively native-like (i.e., P600), even though gender was probed with nouns that do not provide strong distributional cues to gender. In addition, variability with gender agreement was better accounted for by lexical (as opposed to syntactic) aspects. Learners showed no advantage for comprehension over production. They also showed no systematic evidence of reliance on morphological defaults, although their online processing was sensitive to markedness in a native-like manner. Overall, these results suggest that there is facilitation for properties of the L2 that exist in the L1 and that markedness impacts L2 processing, but in a native-like manner. These results also speak against proposals arguing that adult L2ers have deficits at the level of the morphology or the syntax.

Adult second language (L2) learners often exhibit variability in their use of inflectional morphology, even at high levels of proficiency (e.g., Franceschina, 2005; Gillon-Dowens, Vergara, Barber, and Carreiras, 2010; Grüter, Lew-Williams, and Fernald, 2012; Keating, 2009; Lardiere, 1998, McCarthy, 2008; Rossi, Kroll, and Dussias, 2014; Sabourin and Stowe, 2008; see White, 2007 for theoretical considerations). Morphological variability refers to a learner's inconsistent use of obligatory inflectional morphology, as exemplified in (1), which presents elicited production data from an advanced L1-English L2-Spanish learner (McCarthy, 2008, p. 478):

(1) a. *está poniendo las tijeras en la mochila*
she's putting the scissors in the$_{\text{-FEM}}$ backpack$_{\text{-FEM}}$

b. *la mochila es negro*
the$_{\text{-FEM}}$ backpack$_{\text{-FEM}}$ is black$_{\text{-MASC}}$

In (1a-b), the learner correctly establishes gender agreement between the feminine noun *mochila* "backpack$_{\text{-FEM}}$" and the determiner *la* "the$_{\text{-FEM}}$", but then shows incorrect inflection on the adjective *negro* "black$_{\text{-MASC}}$", which is used in the masculine (and, thus, fails to agree with its controller noun). A wealth of research has examined inflectional variability in L2 learners (e.g., Franceschina, 2005; Grüter et al., 2012; Lemhöfer, Schriefers, and Indefrey, 2014; López Prego and Gabriele, 2014; McCarthy, 2008; Montrul, Foote, and Perpiñán, 2008; Morgan-Short, Sanz, Steinhauer, and Ullman, 2010; Prévost and White, 2000; Renaud, 2011, 2012; Rossi, Kroll, and Dussias, 2014; Tokowicz and MacWhinney, 2005; White, Valenzuela, Kozlowska-MacGregor, and Leug, 2004), and some interesting generalizations have emerged from this literature. For example, inflectional errors tend to exhibit systematicity, with some error types occurring more frequently than others (e.g., Dewaele and Véronique, 2001;

Franceschina, 2001; McCarthy, 2008; Montrul et al., 2008; Renaud, 2011, 2012; Sabourin, 2003; White et al., 2004). To account for this asymmetry, some authors have argued that L2ers resort to the use of morphological "defaults", that is, underspecified forms that learners use in target-like contexts and overextend to incorrect ones (e.g., McCarthy, 2008; Montrul et al., 2008; Prévost and White, 2000; White et al., 2004). With respect to number and gender agreement in Spanish, the properties of interest herein, this would mean that learners incorrectly use singular and masculine forms in plural and feminine contexts, but the reverse pattern rarely occurs. The error in (1b), where the learner incorrectly uses masculine inflection in a feminine context constitutes a good example of potential reliance on default morphology.

In addition, some morphosyntactic properties exhibit greater variability than others, even at the highest proficiency levels. For example, Franceschina (2005), López-Prego and Gabriele (2014), McCarthy (2008), Rossi et al. (2014), and White et al. (2004) all compared syntactic number and gender agreement in L2-Spanish by English-speaking learners at different proficiency levels, and found that number was relatively unproblematic across the proficiency spectrum (see also Gabriele, Fiorentino, and Alemán Bañón, 2013). In contrast, gender agreement showed more variability, among both advanced L2ers (e.g., López Prego and Gabriele, 2014; McCarthy, 2008; Rossi et al., 2014; but see White et al., 2004) and even near-native speakers (e.g., Franceschina, 2005; Grüter et al., 2012). Since grammatical gender is not instantiated by these learners' L1, some authors have claimed that inflectional variability is due to brain maturation effects specifically affecting novel L2 syntactic properties (e.g., Franceschina, 2005; Hawkins and Chan, 1997; Long, 2005; Sabourin, 2003; Tsimpli and Dimitrakopoulou, 2007). Recent proposals for the domain of grammatical gender (Grüter et al., 2012; Hopp, 2013), however, have argued that variability with

grammatical gender is more tied to aspects of lexical gender assignment (i.e., associating nouns to their appropriate gender classes as part of their mental lexical entry). Along these lines, recent studies have shown that even L2ers whose L1 realizes gender exhibit variability with gender inflection due to weak knowledge of lexical gender (e.g., Lemhöfer et al., 2014), even at high levels of proficiency (e.g., Sabourin, 2003; Sabourin and Stowe, 2008; White et al., 2004).

Finally, variability appears to emerge in some tasks more than others. Several studies have shown that learners usually perform better in tasks measuring comprehension (e.g., sentence-picture matching, written recognition task), relative to those examining oral production (e.g., Alarcón, 2011; Grüter et al., 2012; Montrul et al., 2008), and some authors have proposed that inflectional variability might be a production-specific phenomenon (Prévost and White, 2000; Rothman, 2007; White, 2011). However, as pointed out by Grüter et al. (2012), the difference between comprehension and production shows a confound with processing burden in many studies. Indeed, comprehension has often been examined via offline tasks (e.g., McCarthy, 2008; Montrul et al., 2008; White et al., 2014), while the very nature of spoken language production calls for online tasks, where the processing burden is higher, as learners must retrieve and articulate the words in real-time. Therefore, the observed performance differences between comprehension and production may well be related to task type, rather than differences between the receptive and expressive knowledge of morphology.

The present paper is devoted to the study of morphological variability in adult L2 learners, with a focus on the central issues highlighted above. The properties of interest are number and gender agreement in L2-Spanish, with a novel emphasis on markedness relations, since it has been argued that underspecified features (i.e., defaults) correspond to unmarked ones (e.g., Harley and Ritter, 2002). We examine the extent to which L2

inflectional variability can be accounted for by (i) reliance on default morphology; (ii) the properties of the learners' L1; and (iii) the type of knowledge tapped into (receptive vs. expressive) related to methodological design (comprehension vs. production).

Number and Gender Agreement in Spanish

Spanish nouns belong to one of two genders, masculine or feminine. Although neither gender value is associated with a unique marker (Harris, 1991), a clear regularity can be observed: 99.8% of nouns ending in *–o* are masculine and 96.3% of nouns ending in *–a* are feminine (Teschner and Russell, 1984). These transparent nouns make up approximately two thirds of the Spanish lexicon (Harris, 1991), suggesting that the *–o* and *–a* markers provide strong distributional cues to gender. However, the Spanish lexicon includes many nouns ending in vowel *–e* or a consonant, for which gender can be less reliably determined. These less transparent nouns are the focus of the present study.

Several observations suggest that, in Spanish, feminine is marked for gender and masculine is underspecified (Battistella, 1990; Bonet, 1995; Harris, 1991). For example, when a genderless word (e.g., preposition *para* "for") is modified by an agreement-bearing element (e.g., the indefinite adjective *demasiado* "too-many"), the latter must show masculine inflection (*demasiados paras en ese párrafo* "too-many-MASC fors-NO-GENDER in that paragraph") (Harris, 1991). Likewise, when masculine and feminine nouns are conjoined, all agreement targets must also show masculine inflection. This suggests that masculine inflection is underspecified for gender, since it can appear with genderless elements and even feminine ones, but feminine forms are marked since they can only appear with feminine nouns.

The Spanish number system distinguishes between singular and plural. Singular shows zero inflection, while plural is formed by suffixing *–s* or *–es* to the singular form (the root) (e.g., *coche/coches* "car/cars", *árbol/árboles* "tree/trees") (Saporta, 1965). This asymmetry with respect to the presence of overt inflection has been taken as evidence that plural forms are marked, relative to singular (e.g., Battistella, 1990). Additional evidence that singular and plural are asymmetrically represented is that singular has a broader syntactic distribution than plural. For example, the singular dative clitic *le* can be coindexed with a plural phrase (*Julia le$_i$ teme [a las ratas]$_i$* "Julia CL$_{-SG}$ fears rat$_{-PL}$"), but its plural counterpart *les* cannot be coindexed with a singular phrase. This suggests that singular forms are underspecified for number, since they can agree both with singular and plural phrases, but plural forms are marked, since they are restricted to plural elements.

Theories on L2 Morphological Variability

Different L2 theoretical models make competing claims regarding the locus and nature of L2 morphological variability. The "representational accounts" posit that L2 morphological variability stems from a representational deficit at the level of the syntax (e.g., Franceschina, 2005; Hawkins, 2001; Tsimpli and Dimitrakopoulou, 2007). Under these models, only syntactic properties of the L2 that exist in the learners' L1 can be acquired to native-like levels, due to maturation. For novel properties, it is argued that L2ers use compensatory strategies. With respect to the acquisition of grammatical gender by speakers of gender-free languages, one potential strategy would be phonological rhyming between noun endings and inflectional forms (Hawkins, 2001; White et al., 2004). This position is well represented by the Interpretability Hypothesis (Tsimpli and Dimitrakopoulou, 2007), for which it is novel syntactic features (i.e., those

which make no semantic contribution to the interpretation of a lexical item) that become inaccessible in adult L2 acquisition. For syntactic agreement, this would be the case with number and gender information on determiners and adjectives.

In contrast, the "computational accounts" argue that the properties of the learner's L1 do not constrain L2 acquisition, but rather that morphological variability is a corollary of performance limitations (e.g., Haznedar and Schwartz, 1997; Hopp, 2010; Prévost and White, 2000). This is the position adopted by the Missing Surface Inflection Hypothesis "MSIH" (Prévost and White, 2000; Haznedar and Schwartz, 1997) according to which, variability results from the difficulty associated with the retrieval of the appropriate inflectional forms and their mapping onto lexical items, particularly in oral production (White, 2011).

The proponents of the MSIH offer the following analysis for the observation that L2ers often adopt defaults. They assume that features are fully specified in the syntax, but not in the morphology (Halle and Marantz, 1993; Harley and Noyer, 1999). In the morphology, singular and masculine are underspecified, whereas plural and feminine are marked (i.e., fully specified) (Bonet, 1995; Cowper, 2005; Harley, 1994; Harley and Ritter, 2002; Harris, 1991). For agreement to be successful, the features on lexical items must be compatible with those of the syntax. A perfect match is not required, but there can be no feature clash. For cases where the syntax is specified as plural or feminine, the parser will select a plural or feminine form (i.e., fully specified in the morphology), as they provide a perfect match (e.g., *casa roja* "house-$_{\text{FEM}}$ red-$_{\text{FEM}}$"). However, masculine or singular forms do not clash in this context, due to their lack of specification (e.g., *casa rojo* "house-$_{\text{FEM}}$ red-$_{\text{UNDERSPECIFIED}}$"). For cases where the syntax is specified as singular or masculine, only underspecified forms can be inserted, since inflectional forms that are fully specified as masculine or singular are not

available (e.g., *coche rojo* "car-$_{\text{MASC}}$ red-$_{\text{UNDERSPECIFIED}}$"), and the insertion of plural or feminine forms would cause a feature clash (e.g., *coche roja* "car-$_{\text{FEM}}$ red-$_{\text{MASC}}$"). The proponents of the MSIH argue that, although L2ers can acquire the full specifications of features, they have trouble retrieving them in production, due to processing burden. In such cases, L2ers select a "good enough form" (i.e., an underspecified form or default) even if a better candidate is available. This yields the well-attested asymmetric pattern of errors in production, where learners are more likely to underspecify a feature, as in *casa rojo* (house-$_{\text{FEM}}$ red-$_{\text{UNDERSPECIFIED}}$), than to produce a feature clash.

Grüter et al. (2012) agree that inflectional variability with grammatical gender is tied to difficulty with lexical retrieval, but point to gender assignment (i.e., classifying nouns as masculine or feminine) as the source of variability. The authors examined gender agreement in advanced L1-English L2-Spanish learners, and found that they were native-like in offline comprehension, but made errors of gender assignment in production and could not utilize gender predictively in online comprehension. Grüter et al. propose that the links between nouns and their abstract gender classes are weaker in L2ers. Consequently, L2ers have difficulty with the retrieval and use of gender information online. A subsequent study by Hopp (2013) looking at L1-English L2-German learners provides support for this proposal. Hopp found that only those L2ers who showed robust and consistent knowledge of lexical gender (i.e., those who assigned almost all nouns to their appropriate gender values) behaved like German native speakers in their ability to utilize gender information predictively. Taken together, these studies suggest that the quality of the learners' lexical representations for gender accounts for variability with gender agreement. Following Hopp (2013), we will refer to this proposal as the Lexical Gender Learning Hypothesis.

Finally, an alternative account of inflectional variability is provided by McCarthy (2008), who builds on the idea that variability is systematic and consists of the overuse of default morphology. McCarthy distinguishes between two types of errors, *default/underspecification errors* and *feature clash errors*. Default errors are cases where the syntax is fully specified as plural or feminine, but the learner uses an underspecified form on lexical items (i.e., singular, masculine). This is the case in (2a) and (2b) for number and gender, respectively:

(2) a. *las mochilas* *son *negra*
the backpack-FEM-**PL** are black-FEM-**SG**

b. *la mochila* *es *negro*
the backpack-**FEM**-SG is black-**MASC**-SG

Feature clash errors show the opposite pattern; the syntax is fully specified as singular or masculine, but the lexical items are fully specified as plural or feminine. Examples are shown in (3a) and (3b) for number and gender, respectively:

(3) a. *el bolso* *es *negros*
the purse-MASC-**SG** is black-MASC-**PL**

b. *el bolso* *es *negra*
the purse-**MASC**-SG is black-**FEM**-SG

The main tenet of McCarthy's proposal (2008) is that L2ers' errors mainly consist of default errors. Unlike the MSIH, however, McCarthy argues that variability is representational, and that overreliance on default morphology is not specific to production, but also emerges in comprehension. Her proposal also differs from other representational accounts in two ways. First, the deficit is located at the level of the morphology. That is, L2ers are assumed to be able to acquire all syntactic projections of the L2, but not the full specification of features in the morphology. Second, variability

is not restricted to novel properties, but can also emerge for properties that exist in the L1.

The Present Study

The present study investigates the nature of L2 morphological variability and evaluates the above theoretical proposals in a group of adult L1-English L2-Spanish learners of upper-intermediate to advanced proficiency. We examine both number (present in L1) and gender (absent in L1) agreement, in order to examine the role of the learners' L1. For each feature, we examine how markedness impacts agreement. In addition, we examine both comprehension and production. Comprehension in our study was examined via event-related potentials "ERPs", which are brain responses that are time-locked to specific events of interest. ERPs provide high temporal resolution, allowing us to examine the learners' sensitivity to agreement exactly at the time when it is computed. This is important, in light of models which assume that inflectional variability is linked to the learners' inability to rapidly retrieve lexical information in real-time (e.g., Grüter et al., 2012; Hopp, 2013; Prévost and White, 2000). In addition, different processing mechanisms are associated with qualitatively different ERPs. Thus, if learners and native speakers show qualitatively different brain responses to the same property, this might indicate that differences at the level of linguistic representation cause L2ers to recruit different processing mechanisms (e.g., Tsimpli and Dimitrakopoulou, 2007; see also Alemán Bañón et al., 2014).

For example, in native speakers, agreement violations elicit a P600, a positive deflection between ~500-900ms in central-posterior electrodes (e.g., Osterhout and Holcomb, 1992; Hagoort, Brown, and Groothusen, 1993). The P600 has been argued to reflect syntactic integration (e.g., Kaan, Harris, Gibson, and Holcomb, 2000), reanalysis

(e.g., Osterhout and Holcomb, 1992) and repair (e.g., Barber and Carreiras, 2005; see Molinaro, Barber, and Carreiras, 2011 for a review). Importantly, although the P600 does not exclusively index morphosyntactic anomalies (i.e., it has been reported for certain types of semantic violations; see Bornkessel-Schlesewsky & Schlesewsky, 2008; Kim & Osterhout, 2005), it is consistently found for morphosyntactic errors in native speakers. In contrast, lexical semantic processes are typically reflected in the N400 component, a negativity between ~250-500ms that is sensitive to the strength of lexical associations (e.g., Kutas and Hillyard, 1980; see Lau, Phillips, and Poeppel, 2008 and Kutas and Federmeier, 2011 for reviews). Interestingly, a number of studies have found that low-proficiency learners elicit an N400 for morphosyntactic errors for which native speakers show a P600, which has been interpreted as evidence for qualitative differences between L1 and L2 processing at lower levels of proficiency (e.g., Osterhout, McLaughlin, Pitkänen, Frenck-Mestre, and Molinaro, 2006; McLaughlin, Tanner, Pitkänen, Frenck-Mestre, Inoue, and Valentine, 2010; Tanner, McLaughlin, Herschensohn, and Osterhout, 2013). Importantly, in the case of gender agreement, this has even been the case among advanced L2ers (e.g., Foucart and Frenck-Mestre, 2012; Morgan-Short et al., 2010), suggesting that qualitative differences between L1 and L2 processing are not confined to the lower levels of proficiency.

The P600 is sometimes preceded by a Left Anterior Negativity (LAN), a negative deflection between ~300-500ms typically captured by left anterior electrodes (Friederici et al., 1996). Some have proposed that it reflects automatic morphosyntactic processing (see Molinaro et al., 2011), although a problem with such interpretation is that the LAN is absent in many L1 studies on agreement (e.g., Alemán Bañón, Fiorentino, and Gabriele, 2012; Hagoort, 2003; Wicha et al., 2004). Others have argued that the LAN is reminiscent of the N400 and reflects either the semantic integration difficulty caused by

the agreement error (e.g., Guajardo and Wicha, 2014) or individual differences with respect to processing mechanisms (e.g., Tanner and Van Hell, 2014). Importantly, many studies on agreement have reported P600 effects not preceded by a negativity, but not the reverse. This suggests that the P600 is the more reliable index of agreement processing in L1 speakers. This is important, since some studies on L2 processing have interpreted the absence of the LAN for morphosyntactic errors as evidence for processing deficits in adult L2ers (Clahsen and Felser, 2006; Ullman, 2001; Weber-Fox and Neville, 1996). However, the observed variability with respect to LAN elicitation in native speakers indicates that the LAN might not be a reliable metric to examine the nature of L2 processing (see Alemán Bañón, Fiorentino, and Gabriele, 2014; McLaughlin et al., 2010; Tanner et al., 2013).

To our knowledge, this is the first ERP study that examines the unique contribution of markedness to agreement processing in L2ers.

ERP Studies on Number/Gender Agreement and Markedness

Natives.

ERP studies comparing number and gender agreement in native speakers have reported largely similar results for both features (Nevins, Dillon, Malhotra, and Phillips, 2007; Gillon-Dowens, Vergara, Barber, and Carreiras, 2010; Alemán Bañón et al., 2012; cf. Barber and Carreiras, 2005), suggesting that similar processes underlie the two agreement types. With respect to morphological markedness, Deutsch and Bentin (2001) found that gender violations in Hebrew yielded a larger P600 when they were realized on plural (i.e., marked) as opposed to singular verbs, which they relate to plural being more salient. Kaan (2002) reports a larger P600 for subject-verb violations in Dutch when the offending verb was plural (although this effect only emerged when a

singular noun intervened between the agreeing words). Along similar lines, Mehravari, Tanner, Wampler, Valentine, and Osterhout (2015) report a larger P600 for English subject-verb violations that involve overt incorrect inflection relative to violations caused by missing inflection. Finally, a study by Tanner and Bulkes (2015) provides evidence that violations of subject-verb agreement in English yield a larger P600 when the subject NP provides additional plural cues.

Alemán Bañón and Rothman (2016) is one of first studies to have examined the unique contribution of morphological markedness to the native processing of agreement. The study focused on noun-adjective number and gender agreement in Spanish (…*catedral que parecía inmensa*… "cathedral-$_{\text{FEM-SG}}$ that looked huge-$_{\text{FEM-SG}}$). Markedness was examined by manipulating the number and gender specification of the controller nouns, such that half of them were feminine and the other half, masculine; half of the nouns were used in the plural and the other half, in the singular. This design yielded two types of gender errors, which correspond to McCarthy's default errors (feminine noun + masculine adjective) and feature clash errors (masculine noun + feminine adjective), and two types of number errors, default errors (plural noun + singular adjective) and feature clash errors (singular noun + plural adjective). Results from 27 Spanish native speakers revealed that, in the 500-1000ms time window, all four violation types yielded robust P600 effects. Interestingly, the P600 emerged earlier for both types of feature clash errors (i.e., it became significant between 250-450ms). In addition, P600 amplitude was larger for feature clash than default errors, although this effect only emerged for number. In this same time window (500-1000ms), all violation types also yielded a late negativity with an anterior distribution. In studies that involve a grammaticality judgment task this negativity has been argued to reflect the cost of maintaining the violations in working memory (e.g., Alemán Bañón et al., 2012;

Gillon-Dowens et al., 2010; Sabourin and Stowe, 2004; Zawiszewski, Santesteban, and Laka, 2014).

Alemán Bañón and Rothman's results did not reveal a LAN across violation types, a finding that is consistent with many studies on the native processing of agreement. Although feature clash errors were more negative than their grammatical counterparts between 250-450ms, this effect did not exhibit the canonical morphology of the LAN. In the case of gender, the negativity was sustained. In the case of number, it was marginal and did not show a left anterior distribution. It is, therefore, unclear the extent to which markedness impacts the processes reflected by the LAN (see Molinaro et al., 2011 and Tanner and Van Hell, 2014 for discussions on some of the factors which might impact the elicitation of the LAN).

Alemán Bañón and Rothman interpreted these findings as evidence that native speakers are sensitive to markedness asymmetries, such that violations where the mismatching feature is marked (i.e., feminine for gender; plural for number) are detected earlier (as indicated by the earlier onset of the P600) and, at least in the case of number, are more salient or disruptive (as indicated by a larger P600).

L2 Learners.

L2 ERP studies comparing number and gender have shown a quantitative advantage for number, but only in cases where number is present in the L1 and gender is unique to the L2. For example, Gillon-Dowens et al. (2010) and Alemán Bañón et al. (2014) found that advanced L1-English learners of Spanish elicited a larger P600 for number than gender violations in most contexts examined (see also Rossi et al., 2014). This advantage, however, was absent in the study by Gillon-Dowens, Guo, Guo, Barber, and Carreiras (2011), who compared Spanish number and gender agreement in native

speakers of Chinese, a language that does not instantiate number or gender agreement. Crucially, neither Gillon-Dowens et al. (2010) nor Alemán Bañón et al. (2014) controlled for markedness in the way number and gender were compared. While gender violations included both default and feature clash errors, number violations only involved feature clash errors, which are presumably more disruptive in comprehension (e.g., McCarthy, 2008). It is, therefore, possible that the larger P600 for number over gender in these studies was due to differences in markedness, in line with the results by Alemán Bañón and Rothman (2016) and other studies (e.g., Deutsch and Bentin, 2001).

In addition, native-like processing for gender appears to depend on whether the target nouns provide strong distributional cues to gender. When this is the case, learners tend to show native-like processing in terms of ERP responses, even when their L1 is gender-free (e.g., Gillon-Dowens et al., 2010, 2011; Alemán Bañón et al., 2014; Rossi et al., 2014). This has been the case for studies looking at gender agreement in Spanish, all of which have exclusively tested masculine nouns ending in *–o* and feminine nouns ending in *–a*. For example, the studies by Gillon-Dowens et al. (2010, 2011) and Alemán Bañón et al. (2014) report robust P600 effects for gender violations in Spanish across different syntactic domains (within the Determiner Phrase "DP", across the Verb Phrase "VP"). This was also the case for the most proficient L1-English L2-Spanish learners in the study by Rossi et al. (2014), who examined gender agreement on clitic pronouns. A similar pattern of results has emerged in L2 learners of Spanish at lower proficiency levels (Gabriele et al., 2013; Bond et al., 2011; Tokowicz and MacWhinney, 2005), which is surprising, given that mastery of this property often appears restricted to highly proficient L2ers. Most of these studies also tested the L2ers' knowledge of lexical gender offline and reported at-ceiling accuracy rates (e.g., Alemán Bañón et al., 2014: 99%; Bond, 2012: 98%; Gabriele et al., 2013: 99%; Gillon-Dowens et al., 2010:

98%; Gillon-Dowens et al., 2011: 96%). This suggests that, when nouns provide strong distributional cues to gender, learners across the proficiency spectrum can correctly assign it, and resolve agreement online in a native-like manner.

A different picture arises from studies that have examined French (e.g., Foucart and Frenck-Mestre, 2012) and Dutch (e.g., Meulman, Stowe, Sprenger, Bresser, and Schmid, 2014; Sabourin, 2003; Sabourin and Stowe, 2008). Although French nouns provide some morphophonological cues to gender (e.g., ~80%; Lyster, 2006), the masculine and feminine values of the French system are associated with a wider range of word endings than their Spanish counterparts (e.g., Séguin, 1969; Lyster, 2006), making rules for gender assignment more complex. Foucart and Frenck-Mestre (2012) found that advanced L1-English L2-French learners did not consistently show native-like sensitivity to gender violations, despite a low error rate with offline gender assignment. The L2ers elicited a P600 for noun-adjective violations within the DP, but an N400 for adjective-noun violations (a word order that is dispreferred in French), and no effects for violations across a VP. Notice that these results contrast with those by Gillon-Dowens et al. (2010, 2011), Alemán Bañón et al. (2014), and Rossi et al. (2014), where the P600 for gender errors remained robust across a range of different syntactic domains (including clitics, a syntactic category that is absent in English). One possibility is that the lack of strong distributional cues to gender made it difficult for the L2ers in the Foucart and Frenck-Mestre study to retrieve gender information online, at least in contexts that can be considered more taxing (in line with Grüter et al., 2012 and Hopp, 2013).

Sabourin (2003) and Sabourin and Stowe (2008) examined the processing of gender agreement in L2-Dutch by advanced learners whose L1 did (German or Romance) or did not instantiate gender (English). Although the Dutch and German gender systems

comprise different gender values (Dutch: common, neuter; German: masculine, feminine, neuter), they show extensive overlap. That is, most masculine and feminine nouns in German correspond to common nouns in Dutch, and most German neuter nouns are also neuter in Dutch. This is likely to facilitate gender assignment for L1-German learners of Dutch. No such overlap exists between Dutch and Romance. Their results revealed that only the L1-German group showed robust offline knowledge of lexical gender (mean accuracy rate: 93%) and native-like processing for gender violations (i.e., P600). In contrast, both the L1-Romance and L1-English groups scored below 80% accuracy with offline gender assignment, and neither group showed native-like processing for gender violations. This suggests that, when nouns do not provide strong distributional cues to gender, even advanced L2ers show difficulty with both gender assignment and agreement, even if their L1 instantiates gender (see also Meulman et al., 2014, who replicated these findings with a group of advanced L1-Romance L2-Dutch learners).

Lemhöfer et al. (2014) provide further evidence for lexically-based variability with gender agreement in a group of L1-German L2-Dutch learners. The authors examined gender agreement with cognates which exhibit opposite gender values in German and Dutch, and found that the L2ers showed no sensitivity to gender violations when only objective gender assignment was taken into account, that is, when only the native speakers' rules for gender assignment were considered. In contrast, when the learners' idiosyncratic gender assignment was taken into account, they showed a native-like P600.

To summarize, previous L2 studies have shown that, with increased proficiency, learners tend to show native-like processing for both number and gender agreement, although the evidence for gender mainly comes from studies that have examined nouns

with strong distributional cues to gender (e.g., Spanish *–o* and *–a*). In addition, the unique contribution of markedness to agreement processing remains to be investigated, and some previous studies arguing for L1 facilitation effects (Gillon-Dowens et al., 2010; Alemán Bañón et al., 2014) have confounded markedness with L1-L2 similarity. In the present study, we address both issues. First, we systematically manipulate markedness relations for both number and gender agreement. In addition, we examine gender via Spanish nouns that do not provide strong distributional cues to gender. Our design shies away from masculine and feminine nouns showing the *–o* and *–a* markers and, instead, focuses on Spanish nouns ending in vowel *–e* or in a consonant. Crucially, while distributional gender cues in some of these nouns are not entirely absent (e.g., nouns that end in suffix *-ión* tend to be feminine, although there are exceptions, such as *avión* "plane" or *camión* "truck"), such cues are much weaker than those provided by *–o* and *–a*, due to their reduced frequency in the input. In addition, unlike previous L2 ERP studies on Spanish gender agreement, our design involves a wide range of endings for both the masculine and feminine values (e.g., masculine: *traje* "suit", *reloj* "watch", *pastel* "cake", *álbum* "album", *avión* "plane", *ordenador* "computer", *pez* "fish"; feminine: *pared* "wall", *calle* "street", *cárcel* "jail", *reunión* "meeting", *flor* "flower", *ley* "law", *nuez* "walnut"), which is expected to increase the difficulty of online lexical gender retrieval in the L2ers.[1]

Research Questions and Predictions

Our study addresses the following questions:

[1] A complete list of the experimental nouns is provided in the Appendix.

(i) To what extent is variability accounted for by the learners' reliance on default morphology? We address this question by systematically manipulating markedness relations for both number and gender agreement across tasks.

(ii) To what extent is morphological variability determined by the properties of the learners' L1? We examine this question by comparing number (present in the L1) and gender agreement (unique to L2). We also examine the relation between the L2ers' knowledge of lexical gender and their ability to establish gender agreement online, to shed light on the qualitative nature of variability with gender (syntactic vs. lexical).

(iii) Is morphological variability a production-specific phenomenon or does it also emerge in comprehension? We address this question by examining both comprehension and production of agreement morphology (receptive vs. expressive knowledge). By focusing on online comprehension and production, we can better compare the L2ers' productive vs. receptive knowledge while controlling for the online nature of the task (e.g., Grüter et al., 2012).

Predictions according to each model.

Representational accounts predict an advantage for number over gender across measures. This is because number is realized in the learners' L1, but gender is unique to the L2. Importantly, qualitatively native-like processing in the EEG task (i.e., P600) is predicted for number, but not gender, especially for nouns that lack strong distributional cues to gender and do not allow for the use of compensatory strategies (i.e., phonological rhyming between noun and adjective endings). In addition, sensitivity to gender violations is not predicted to differ as a function of error type. This is because variability with gender is assumed to be nonsystematic for L2ers as a group (e.g., Hawkins, 2001), meaning that some learners might use masculine as the default gender

and others, feminine. Such behavior might then yield a null effect of markedness in a group analysis.

Under the computational accounts, an overall advantage for comprehension over production is predicted, due to the difficulty associated with lexical retrieval in spoken production. As for the number vs. gender comparison, it is possible that there will be no differences given the L2ers' proficiency, even if the target nouns do not show canonical gender marking (e.g., White et al., 2004). In the ERP data, learners are predicted to be able to show native-like brain responses for the two features, although a quantitative advantage for number is still possible due to L1 bootstrapping (e.g., Alemán Bañón et al., 2014; Gillon-Dowens et al., 2010; Rossi et al., 2014). As for markedness, it is possible that L2ers will make more default than feature clash errors in production. However, they are not necessarily predicted to show greater sensitivity to feature clash than default errors in comprehension, as reliance on default morphology is assumed to be caused by the difficulty associated with lexical retrieval in production.

Under the Lexical Gender Learning Hypothesis, the L2ers' knowledge of lexical gender should positively correlate with their sensitivity to gender across measures (e.g., mean accuracy detecting gender errors in comprehension, P600 amplitude to gender errors, and mean accuracy with gender in production). In addition, the L2ers mean accuracy with gender in production and comprehension should correlate, since robust knowledge of lexical gender should result in target-like performance across tasks (e.g., Hopp, 2013).

Finally, McCarthy's proposal (2008) predicts an effect of markedness across features and tasks, consistent with the notion that L2ers have a general deficit at the level of the morphology which causes them to overuse default forms. In comprehension, L2ers are predicted to be more accurate with the detection of feature clash than default errors, for

both number and gender (although McCarthy's study found that number was relatively unproblematic). In production, L2ers are predicted to make more default than feature clash errors, for both number and gender.

In the ERP data, there are different ways in which markedness could impact processing. One possibility is that only feature clash errors will yield a P600. Default errors in comprehension might not be sufficiently disruptive to yield a P600.This is because, under McCarthy's account, underspecified forms in contexts where the syntax is fully specified (i.e., default errors) are allowed by the learner's grammar. Such a pattern of results would be nonnative-like. It is also possible that both error types will yield a P600, but that P600 amplitude will be larger for more disruptive errors (i.e., feature clash). Such a pattern would suggest that agreement processing in the L2 is sensitive to morphological markedness, but would not be consistent with McCarthy's proposal, since the native speakers in Alemán Bañón and Rothman (2016) showed a similar pattern of results (for number). Finally, it is also possible that the P600 will emerge earlier for feature clash than default errors. Such a pattern, which also emerged in Alemán Bañón and Rothman (2016), would also be consistent with the idea that feature clash errors are more disruptive, but would not be indicative of a representational deficit.

Experiment 1: Comprehension

Participants.

Twenty-two English-speaking learners of Spanish (12 females; mean age: 25; SD: 7.5) participated in the study. None of them were significantly exposed to Spanish before age eight (mean age of acquisition: 14; range: 8-23) and, therefore, they can be considered late learners. Proficiency in L2 Spanish was measured with a 50-item test

that includes the cloze section from the Diploma de Español como Lengua Extranjera "DELE" and the reading section from the MLA Cooperative Foreign Language Test (e.g., White et al., 2004; McCarthy, 2008; Grüter et al., 2012). Sixteen learners scored within the advanced range (43-50), and six of them, within the intermediate range (33-38). The mean score for the group was 43 (SD: 5).

All of the learners were native speakers of English and none were significantly exposed to languages with grammatical gender before they started learning Spanish.[2] They were all university students or post-graduates, and most of them had Spanish as one of their academic concentrations. On average, they reported having received 7.3 years of instruction in Spanish (SD: 2.7) and having lived in a Spanish-speaking country for 15 months (range: 0-48 months, with only four learners having lived in a Spanish-speaking environment for less than eight months).

The control group included 27 native speakers of Castilian Spanish, reported in Alemán Bañón and Rothman (2016). Since our L2 group can be best characterized as being of intermediate to advanced proficiency, their data will be analyzed independently and their results will be compared to those of native speakers to identify potential qualitative differences. All 49 participants had normal or corrected-to-normal vision and indicated no history of neurological disabilities. They were all right-handed, as assessed by the Edinburgh Handedness Questionnaire (Oldfield, 1971). All of the participants were tested in the UK and compensated for their time.

Stimuli.

[2] One of the learners was minimally exposed to Irish during childhood. Another learner indicated being a heritage speaker of Japanese, a language which has word classes, but not gender agreement.

The agreement dependency of interest is that between the head noun of a relative clause and a predicative adjective, which was located across a Complementizer Phrase (CP). An example is provided in (1a). The rationale for examining nonlocal agreement is that L2ers' sensitivity to morphosyntactic dependencies has been found to decrease in nonlocal contexts, due to increased complexity (e.g., Keating, 2009; Gillon-Dowens et al., 2010; Foucart and Frenck-Mestre, 2012). Thus, we assumed that learners' reliance on default morphology would be more likely to emerge when the dependency involved elements from different phrases.

Markedness was examined by manipulating the number and gender specification of the controller nouns, such that half of them were masculine (1a, 1b) and the other half, feminine (1c-d). In addition, half of the trigger nouns were used in the singular (1a, 1c) and the other half, in the plural (1b, 1d).

(1)

Masculine Singular Noun

a. Andrés alquiló un coche que parecía barato durante la excursión.
Andrés rented a car-MASC-SG CP[that looked cheap-MASC-SG] during the excursion

Masculine Plural Noun

b. Andrés alquiló unos coches que parecían baratos durante la excursión.
Andrés rented some car-MASC-PL that looked cheap-MASC-PL during the excursion

Feminine Singular Noun

c. Andrés alquiló una habitación que parecía espaciosa la semana pasada.
Andrés rented a room-FEM-SG that looked spacious--FEM-SG the week past

Feminine Plural Noun

d. Andrés alquiló unas habitaciones que parecían espaciosas la semana pasada.
Andrés rented some room-FEM-PL that looked spacious--FEM-PL the week past

The agreement by markedness by feature manipulation yielded a total of 12 experimental conditions, which are shown in Table 1. We designed 20 items for each of these conditions (240 sentences total). To achieve the 40 items per condition recommended by Molinaro et al. (2011), we collapsed across gender when examining

number and vice versa (see Morgan-Short et al., 2010 for a similar approach). That is, items examining the singular vs. plural asymmetry encompassed both masculine and feminine nouns (equally distributed across the singular and plural conditions). Likewise, items examining the masculine vs. feminine asymmetry included both singular and plural nouns (equally distributed across the masculine and feminine conditions).

<Insert Table 1 here>

These materials were interspersed with 160 sentences (80 ungrammatical) from a separate study that does not manipulate number and gender and does not include any adjectives, plus 80 grammatical fillers which involve predicative adjectives modifying personal pronouns (e.g., *Nosotros somos muy simpáticos y ellos también* "We are very friendly and so are they"). There was an equal amount of grammatical and ungrammatical sentences in the overall design, in order to prevent an excessive number of ungrammatical sentences from attenuating the P600 (Coulson, King, and Kutas, 1998; Hahne and Friederici, 1999). These materials were counterbalanced across 6 experimental lists, such that a given learner would see 20 items per each of the 12 conditions, but no participant saw the same sentence twice. Each list also included one version of each sentence from a separate study, and all of the grammatical fillers.

Item controls.

None of the critical nouns exhibited the *–o*/*–a* markers strongly associated with masculine and feminine gender. Instead, we selected masculine and feminine nouns that show a wide range of endings. The log count for all nouns and adjectives was obtained from the EsPal database (EsPal Written Corpus, 2012; Duchon, Perea, Sebastián Gallés,

Martí, and Carreiras, 2013). The masculine and feminine nouns were matched with respect to both frequency, $t(118) = -1.471$, $p > .1$, and length: $t(118) = -1.512$, $p > .1$. The masculine and feminine forms of the adjectives were also matched for frequency, $t(238) = 1.6$, $p > .1$, and their length was the same. With respect to the singular-plural comparison, it was not possible to control the nouns or the adjectives for either frequency or length. Plural items were longer and less frequent than their singular counterparts.

The critical adjectives were never sentence-final, to avoid semantic wrap-up effects. In addition, their position within the sentence was held constant across conditions (e.g., Van Petten and Kutas, 1990). Each critical adjective was used twice, once with a masculine noun (e.g., *bosque…oscuro* "forest…dark") and once with a feminine one (e.g., *catedral…oscura* "cathedral…dark"). Each critical noun was also used twice. Since the testing involved two sessions (see *Procedure*), the experimental lists were designed such that learners would only see one version of each critical adjective per session, to minimize repetition effects.

Procedure.

The testing involved two sessions (Alemán Bañón et al., 2012, 2014; O'Rourke and Van Petten, 2011), separated by a minimum of three days and a maximum of two weeks. Each session lasted for approximately 3 hours (EEG recording: 1 hour). During the first session, participants gave informed consent, filled out a background questionnaire and the handedness inventory. Then, they completed the first EEG recording and took the proficiency test. The second session started with the second EEG recording. Then, participants took the elicited production task (Experiment 2) and a Gender Assignment Task.

For the EEG recordings, participants were instructed to silently read a series of Spanish sentences and decide if they were good or bad (e.g., Alemán Bañón et al., 2012, 2014; Gillon-Dowens et al., 2010, 2011; Kaan, 2002; Nevins et al., 2007). Each session began with a practice set that included eight sentences, half of which were ungrammatical. None of the ungrammatical practice trials involved agreement errors. To ensure that participants understood the task, they received feedback for the first three trials. Immediately after the practice, the experiment began. Each experimental session was divided into six blocks of 40 sentences, separated by five short breaks. Within each block, sentences from all experimental conditions (plus distractors) were randomly intermixed. No feedback was provided for the experimental items. The presentation of the sentences was carried out using *Paradigm* by Perception Research Systems Inc. (Tagliaferri, 2005).

The trial structure was as follows: first, a fixation cross appeared in the center of the monitor for 500ms. Then, the sentence was presented one word at a time using the RSVP (Rapid Serial Visual Presentation) method. Each word was presented for 450ms and followed by 300ms pauses (e.g., Alemán Bañón et al., 2012, 2014). At the end of each sentence, there was a 1000ms pause, followed by the prompts for the grammaticality judgment: the words *Bien* "good" for grammatical sentences and *Mal* "bad" for ungrammatical ones. Participants were asked to respond with their left hand (middle and index fingers, respectively) and to favor accuracy over speed. The prompts remained on the screen until the participant pressed one of the two buttons on the computer mouse. Following the behavioral response, there was an inter-trial interval ranging between 500-1000ms, pseudo-randomly varied at 50ms increments.

The purpose of the Gender Assignment Task was to measure the participants' knowledge of lexical gender. Participants were presented with all 120 critical nouns

from the comprehension task and instructed to select the appropriate gender-marked determiner from among two options (*el* "the$_{\text{-MASC}}$" vs. *la* "the$_{\text{-FEM}}$").

EEG recording and analysis.

The continuous EEG was recorded from 64 sintered Ag/AgCl electrodes attached to an elastic cap (Easycap, BrainProducts, GmbH, Germany) and placed according to the 10% System (midline: FPz, Fz, FCz, Cz, CPz, Pz, POz, Oz; hemispheres: FP1/2, AF3/4, AF7/8, F1/2, F3/4, F5/6, F7/8, FC1/2, FC3/4, FC5/6, FT7/8, FT9/10, C1/2, C3/4, C5/6, T7/8, CP1/2, CP3/4, CP5/6, TP7/8, TP9/10, P1/2, P3/4, P5/6, P7/8, PO3/4, PO7/8, O1/2). The recording was referenced online to FCz and re-referenced offline to average mastoids. An additional external electrode (IO) was placed on the outer canthus of the right eye to monitor eye movements. Electrodes FP1 and FP2 (above each eyebrow) were used to monitor blinks. Impedances were kept below 10kΩs for all electrodes. The recordings were amplified by a BrainAmp MR Plus amplifier (BrainProducts, GmbH, Germany) with a bandpass filter of .016 to 200Hz, and digitized at a sampling rate of 1kHz.

The raw EEG was segmented into epochs relative to the critical word (-300ms to 1200ms). Trials with artifacts (blinks, horizontal eye movements, excessive muscle artifact, and excessive alpha waves) were manually rejected from analysis, as were trials that were incorrectly judged in the behavioral task. This resulted in the exclusion of approximately 15% of the data.[3] Data were filtered offline with a 30Hz low-pass filter,

[3] After this exclusion, the number of trials per condition did not reliably differ across the gender conditions (all *p* values > .05) (conditions 1 and 4 grammatical: 34/40; conditions 7 and 10 grammatical: 35/40; conditions 3 and 6 gender default error: 33/40; conditions 9 and 12 gender feature clash: 32/40). The number of trials per condition was numerically similar across the number conditions (conditions 1

baseline-corrected relative to the 300ms pre-stimulus baseline, and averaged per condition and per participant.

Upon visual inspection of the waveforms and previous reports, ERPs were quantified via mean amplitudes in two time windows of interest: the 250-450ms time window, which includes the LAN/N400, and the 450-900ms time window, which includes the P600. Nine regions of interest (ROI) were computed for statistical analysis, by averaging together the mean amplitudes of the relevant electrodes (Left Anterior: F1, F3, F5, FC1, FC3, FC5; Right Anterior: F2, F4, F6, FC2, FC4, FC6; Left Medial: C1, C3, C5, CP1, CP3, CP5; Right Medial: C2, C4, C6, CP2, CP4, CP6; Left Posterior: P1, P3, P5, P7, PO3, PO7; Right Posterior: P2, P4, P6, P8, PO4, PO8; Midline Anterior: FZ, FCz; Midline Medial: Cz, CPz; Midline Posterior: Pz, POz). To ensure that the signal to noise ratio was similar in the ROIs being compared, analyses were carried out separately for the hemispheres and the midline, which comprise different numbers of electrodes. Mean amplitudes were submitted to a repeated-measures ANOVA with Markedness (marked, underspecified), Agreement (grammatical, ungrammatical), Hemisphere (left, right) and Anterior-Posterior (anterior, central, posterior) as repeated factors. For the analyses conducted on the midline, the only topographical factor in the ANOVA was Anterior-Posterior. These analyses were carried out separately for number and gender. Additional analyses were conducted on ERP effect size to directly compare the two features (see *Number versus Gender*, p. 36). We consider *p* values below .05 as significant and those between .05 and .1 as marginal. A false discovery rate correction was applied for post-hoc tests (Benjamini and Hochberg, 1995). The Geisser and

---

and 7 grammatical: 35/40; conditions 4 and 10 grammatical: 33/40; conditions 5 and 11 number default error: 35/40; conditions 2 and 8 number feature clash: 36/40), although in this case there were more items with a singular than a plural noun, due to the L2ers' higher accuracy with the former.

Greenhouse correction was applied for violations of sphericity. Degrees of freedom and *p*-values are reported after correction (Field, 2005).

Results.

Results for the native speaker controls are reported in Alemán Bañón and Rothman (2016) and a detailed summary is provided on pages 14-15. Recall that, in native speakers, all violation types yielded a P600 (500-1000ms), which emerged earlier for both types of feature clash errors (between 250-450ms). In addition, in the case of number, P600 amplitude was larger for feature clash than default errors. Here we report results for the L2 learners.

Behavioral results: Grammaticality Judgment Task.

Table 2 summarizes the L2 learners' accuracy rates for the critical conditions in the Grammaticality Judgment Task. In terms of mean accuracy rates, learners performed at 85% or above in all conditions (range: 85-97), suggesting that they understood the task well and were able to tease apart grammatical and ungrammatical sentences. To examine whether learners were more accurate with some error types than others, d-prime scores (a measure of sensitivity to signals that reflect standardized differences in acceptance rates for ungrammatical versus grammatical sentences) were entered into a two-way repeated-measures ANOVA with Markedness (marked, underspecified) and Feature (number, gender) as repeated factors. Results revealed a main effect of Feature, $F(1, 21) = 44.026$, $MSE = .125$, $p < .001$, driven by the fact that learners were more accurate detecting number than gender errors overall, and a Markedness by Feature interaction, $F(1, 21) = 12.009$, $MSE = 0.58$, $p < .01$. This interaction was driven by the fact that, for number, learners were more accurate with feature clash than default errors;

however, the opposite was true for gender. Pairwise comparisons revealed that this asymmetry was not significant for gender and was only marginal for number, $F(1, 21) = 4.987$, $MSE = .112$, $p = .074$.[4]

<Insert Table 2>

Behavioral results: Gender Assignment Task.

Learners showed a mean accuracy score of 93% in the Gender Assignment Task (range: 78-100). This suggests that, as a group, the L2ers knew the lexical gender of the critical nouns, although there was some variability. A paired samples t-test revealed no accuracy differences between masculine and feminine nouns, $t(21) = \pm 1.105$, $p > 1$, suggesting that the learners' accuracy with gender assignment was balanced across the two gender values (mean accuracy with masculine nouns: 94%; feminine: 92%).

ERP results.

Visual inspection of the grand average ERPs reveals that both number and gender agreement violations yielded more positive waveforms than grammatical sentences between approximately 450-900ms, in central-posterior electrodes. This pattern is consistent with the P600 and is similar to that of the L1-Spanish controls in Alemán Bañón and Rothman (2016). Figures 1-2 show the ERP waveforms for the number conditions (Figure 1: feature clash errors; Figure 2: default errors), and Figures 3-4 for the gender conditions (Figure 3: feature clash errors; Figure 4: default errors). Figures

[4] A similar pattern of results emerged when analyses were conducted on the mean accuracy rates for the ungrammatical conditions (e.g., López Prego and Gabriele, 2013). A similar pattern also emerged when analyses were restricted to the 16 L2ers who scored within the advanced range in the proficiency test.

5-6 show topographic plots of the violation effects for number and gender, respectively. Overall, effects appear more robust for number than gender errors (compare Figures 1-2 to Figures 3-4, and Figure 5 to Figure 6), a difference that did not emerge in the L1-Spanish controls. With respect to the markedness manipulation, the positivity seems equally robust for both types of gender errors (see Figures 3 and 4). In contrast, for number, it appears slightly larger for feature clash than default errors (see Figures 1 and 2), similar to the native controls in Alemán Bañón and Rothman (2016). In the same time window associated with the P600, all violation types also show a late anterior negativity with a left-hemisphere bias (see Figures 5 and 6), similar to the L1-Spanish controls in Alemán Bañón and Rothman (2016).

Preceding the P600, between approximately 250-450ms, number feature clash errors also appear more negative than grammatical sentences (see Figures 1 and 5). This effect shows an anterior distribution, with a left-hemisphere bias. A similar negativity emerged in the L1-Spanish controls, although in the learners it appears sustained, not restricted to the 250-450ms window. No negativities are apparent for all other violation types in this time window. The following statistical analyses were conducted.

<Insert Figure 1>

<Insert Figure 2>

<Insert Figure 3>

<Insert Figure 4>

<Insert Figure 5>

<Insert Figure 6>

Gender: 450-900ms (P600 time window), hemispheres.

The omnibus ANOVA revealed a marginal main effect of Agreement, $F(1, 21) = 3.121$, $MSE = 3.28$, $p = .092$, driven by the fact that gender violations yielded more positive waveforms than grammatical sentences. The main effect of Agreement was qualified by an interaction with Anterior-Posterior, $F(1.46, 30.67) = 4.645$, $MSE = 1.652$, $p < .05$, and by an interaction with Hemisphere, $F(1, 21) = 14.524$, $MSE = .963$, $p = .001$. In addition, the three-way interaction between Agreement, Anterior-Posterior, and Hemisphere was significant, $F(1.36, 28.61) = 5.561$, $MSE = .371$, $p < .05$. Due to the presence of this three-way interaction, follow-ups were conducted in the different ROIs, to better understand the scalp distribution of the Agreement effects. These tests showed that gender violations were more positive than grammatical sentences in Right Posterior, $F(1, 21) = 9.636$, $MSE = 1.353$, $p < .01$, Left Posterior, $F(1, 21) = 5.557$, $MSE = .868$, $p < .05$, and Right Medial, $F(1, 21) = 5.351$, $MSE = 1.616$, $p < .05$ (see Figures 3-4 and 6). In addition, violations were more negative than grammatical sentences in Left Anterior, $F(1, 21) = 6.469$, $MSE = 1.227$, $p < .05$.

Gender: 450-900ms (P600 time window), midline.

Analyses revealed a marginal main effect of Markedness, $F(1, 21) = 3.567$, $MSE = 2.018$, $p = .073$, driven by the fact that sentences with a feminine noun were more positive than sentences with a masculine noun overall, possibly due to baselines differences between the masculine and feminine noun conditions (e.g., *un uniforme que parecía* ADJECTIVE vs. *una catedral que parecía* ADJECTIVE). Analyses also revealed an Agreement by Anterior-Posterior interaction, $F(1.28 , 26.98) = 8.36$, $MSE = 1.5$, $p < .01$, and a marginal Markedness by Agreement by Anterior-Posterior interaction, $F(1.24, 26.01) = 3.389$, $MSE = 1.03$, $p = .069$. Since an interaction that involves Markedness and Agreement is theoretically relevant, we examined the Markedness by

Agreement interaction at each level of the Anterior-Posterior dimension, but it was not significant in any of the regions. The follow-up tests did reveal a marginal main effect of Agreement in Midline Posterior, $F(1, 21) = 6.688$, $MSE = 2.705$, $p = .051$, driven by the fact that gender violations were more positive than their grammatical counterparts (see Figures 3-4 and 6).

Gender: 250-450 (N400 time window)

The omnibus ANOVA revealed no significant effects in the hemispheres. In the midline, however, it showed a significant Markedness by Agreement interaction, $F(1, 21) = 5.408$, $MSE = 1.361$, $p < .05$. Seemingly, the interaction was driven by the fact that default errors yielded more negative waveforms than grammatical sentences, but feature clash errors were more positive than their grammatical counterparts, although none of these differences were significant. The Markedness by Agreement interaction was qualified by a marginal interaction with Anterior-Posterior, $F(1.17, 24.52) = 3.836$, $MSE = .834$, $p = .056$. We, therefore, conducted follow-up tests to examine the nature of the Markedness by Agreement interaction in the different ROIs. Only in Midline Posterior was the interaction significant, $F(1, 21) = 9.657$, $MSE = .916$, $p < .05$, driven by the fact that feature clash errors elicited more positive waveforms than their grammatical counterparts, $F(1, 21) = 6.732$, $MSE = 1.468$, $p < .05$ (signaling the beginning of the P600; see Figure 6), but default errors did not differ from grammatical sentences.

Number: 450-900ms (P600 time window), hemispheres.

The omnibus ANOVA revealed a main effect of Agreement, $F(1, 21) = 9.054$, $MSE = 2.413$, $p < .01$, driven by the fact that number violations yielded more positive

waveforms than grammatical sentences overall. The main effect of Agreement was qualified by an interaction with Anterior-Posterior, $F(1.34, 28.21) = 6.711$, $MSE = 2.327$, $p < .01$, and an interaction with Hemisphere, $F(1, 21) = 7.187$, $MSE = .803$, $p < .05$. In addition, the Agreement by Anterior-Posterior by Hemisphere interaction was significant, $F(1.51, 31.73) = 13.185$, $MSE = .282$, $p < .001$. Follow-up tests conducted within each ROI revealed that the main effect of Agreement was significant in Right Posterior, $F(1, 21) = 9.177$, $MSE = 1.054$, $p < .01$, Left Posterior, $F(1, 21) = 7.106$, $MSE = 1.506$, $p < .05$, Right Medial, $F(1, 21) = 13.204$, $MSE = 1.232$, $p < .01$, and Left Medial, $F(1, 21) = 10.847$, $MSE = .579$, $p < .01$, driven by the fact that number violations overall were more positive than grammatical sentences (see Figures 1-2 and 5). In addition, violations were more negative than grammatical sentences in Left Anterior, $F(1, 21) = 7.082$, $MSE = 1.267$, $p < .05$.

The omnibus ANOVA also showed a marginal Markedness by Agreement by Anterior-Posterior interaction, $F(2, 42) = 3.042$, $MSE = .789$, $p = .058$. Follow-up tests revealed that the Markedness by Agreement interaction was marginal in Right Posterior, $F(1, 21) = 3.09$, $MSE = .52$, $p = .093$, driven by the fact that feature clash errors yielded a larger P600 than default errors (similar to the Spanish controls, where the effect was significant) (see Figures 1-2 and 5).

Number: 450-900ms (P600 time window), midline.

Analyses revealed a main effect of Agreement, $F(1, 21) = 23.844$, $MSE = 3.181$, $p < .001$, driven by the fact that number errors yielded more positive waveforms than grammatical sentences. This effect was modified by an interaction with Anterior-Posterior, $F(1.39 , 29.37) = 8.36$, $MSE = 1.88$, $p = .01$, driven by the fact that the main effect of Agreement was restricted to Midline Posterior, $F(1, 21) = 17.977$,

MSE = 2.387, $p < .001$, and Midline Medial, $F(1, 21) = 28.566$, *MSE* = 1.65, $p < .001$ (see Figures 1-2 and 5).

Number: 250-450ms (N400 time window), hemispheres.

The omnibus ANOVA revealed a main effect of Markedness, $F(1, 21) = 5.473$, *MSE* = 1.276, $p < .05$, driven by the fact that sentences with a plural noun were more negative than sentences with a singular noun overall, possibly due to baselines differences between the singular and plural noun conditions (e.g., *un coche que parecía* ADJECTIVE vs. *unos coches que parecían* ADJECTIVE). The ANOVA also revealed an Agreement by Anterior-Posterior by Hemisphere interaction, $F(1.48, 31.11) = 5.738$, *MSE* = .184, $p < .05$. Follow-up tests were conducted within each ROI to better understand the nature of the three-way interaction. These tests revealed that the main effect of Agreement was not significant in any of the regions after correcting for Type I error. Before applying the correction, the main effect of Agreement was significant in Left Anterior, $F(1, 21) = 6.758$, *MSE* = .912, $p < .05$, driven by the fact that violations were more negative than grammatical sentences. As can be seen in Figures 1 and 5, this effect, which is mainly driven by feature clash errors, is not restricted to the 250-450ms time window, but overlaps with the late anterior negativity shown by all violation types.

The omnibus ANOVA also revealed a Markedness by Agreement by Anterior-Posterior interaction, $F(1.54, 32.39) = 5.585$, *MSE* = .956, $p < .05$. Follow-up tests at each level of Anterior-Posterior showed that number violations were numerically more positive than grammatical sentences in posterior regions, the effect being larger for feature clash errors, relative to default errors. However, these differences did not reach significance. In addition, violations were numerically more

negative than grammatical sentences in anterior regions, mainly for feature clash errors, but these differences also failed to reach significance.

Number: 250-450ms (N400 time window), midline.

The analyses conducted on the midline revealed a marginal main effect of Markedness, $F(1, 21) = 4.125$, $MSE = 1.927$, $p = .055$, driven by the fact that sentences with a plural noun were more negative than sentences with a singular one overall, and a marginal Markedness by Agreement by Anterior-Posterior interaction, $F(1.37, 28.84) = 3.503$, $MSE = .514$, $p = .059$. This interaction seems driven by the fact that feature clash errors were more negative than grammatical sentences in Midline Anterior and Midline Medial, but default errors barely differed from grammatical sentences. Follow-up tests showed that this interaction was not significant in any of the regions.

Number versus gender.

Further analyses were carried out to directly compare the magnitude of the Agreement effects for number and gender. This comparison was carried out in a region including ten central-posterior electrodes (CP3/4, CP1/2, CPz, P3/4, P1/2, Pz), corresponding to the area where P600 effects emerged for both number and gender violations. The analysis was limited to the 450-900ms time window, corresponding to the latency of the P600 effects for the two features. P600 magnitude was calculated by subtracting the grammatical condition from the ungrammatical condition, separately for each feature and for each markedness condition. Effect sizes were then entered into a repeated-measures ANOVA with Feature (number, gender) and Markedness (marked noun, underspecified) as within-subjects factors.

The only significant result shown by the omnibus ANOVA was a main effect of Feature, $F(1, 22) = 8.142$, $MSE = .977$, $p = .01$, driven by the fact that number violations were more positive than gender violations overall.[5]

Correlational analyses.

Further analyses were carried out to examine the relation between the L2ers' knowledge of lexical gender and their overall sensitivity to gender errors, in terms of both behavioral accuracy and P600 magnitude. Behavioral accuracy was operationalized as mean d-prime scores for the gender conditions (collapsing across the two types of gender errors, which did not significantly differ). P600 magnitude was calculated (following the procedure described in *Number vs. Gender*, p. 36) for a 14-electrode region comprising all electrodes in Right Posterior, Midline Posterior, and Left Posterior, which are the regions where the P600 emerged for gender.

A hierarchical regression model was used to examine the extent to which the learners' knowledge of lexical gender (Gender Assignment Task Score) predicted their behavioral sensitivity to gender agreement (D-prime Score), over and above the effects of L2 proficiency (Proficiency Test Score), a variable that has been shown to correlate with knowledge of lexical gender (e.g., Hopp, 2013).[6] In the first step, Proficiency Test

[5] A similar pattern also emerged when analyses were restricted to the 16 advanced L2ers.

[6] An analysis of standardized residuals showed that the data contained no outliers (Standardized Residual Minimum = -1.07, Maximum = 1.01). In addition, the data met the assumption of no perfect multicollinearity (Tolerance = .62, *VIF* = 1.6) and the assumption of independent errors (Durbin-Watson value = 1.3). The histogram of standardized residuals suggested that the data contained approximately normally distributed errors. This was also the case for the P-P plot of standardized residuals, which showed points very close to the regression line. The scatterplot of standardized predicted values showed that the assumptions of homogeneity of variance and linearity were also met. Nevertheless, one reviewer

Score accounted for a significant amount of the variance in D-prime Score, $\beta = .51$, $F(1, 20) = 7.229$, $p < .05$, $R^2 = .265$. When Gender Assignment Task Score was included in the second step, the model also explained a significant proportion of the variance in D-prime Score, $F(2, 19) = 7.449$, $p < .01$, $R^2 = .439$, $R^2_{adjusted} = .38$, and the R^2 *change* was significant ($p < .05$). However, only Gender Assignment Task Score remained a significant predictor (Gender Assignment Task Score: $\beta = .53$, $t(21) = 2.43$, $p < .05$; Proficiency Test Score: $\beta = 1.9$, $t(21) = .882$, $p > .1$) (see Figure 7, plot A).[7]

<Insert Figure 7>

Another hierarchical regression model was used to examine whether the L2ers' knowledge of lexical gender (Gender Assignment Task Score) predicted their brain sensitivity to each type of gender error (P600 Size), over and above the effects of L2 Proficiency, but no significant results emerged at any steps of the regression (see Figure 7, plots B and C).

Interim discussion of Experiment 1.

Here, we briefly discuss the most relevant findings of Experiment 1. We will interpret these findings (and those from Experiment 2) in light of current L2 theoretical models in the *General Discussion* (pp. 49-62). Learners were very accurate with both

pointed out that the distribution of Gender Assignment Task Scores seemed negatively skewed. We thus applied a reverse score transformation to this variable, which corrected the skewness. Crucially, the correlation between D-prime Score and Gender Assignment Task Score remained significant ($p = .01$).

[7] A similar pattern of results emerged when analyses were conducted on mean accuracy rates for the gender violation conditions.

number and gender in the Grammaticality Judgment Task, although they performed better with number (e.g., Gillon-Dowens et al., 2010). Importantly, however, their brain responses were qualitatively native-like for both number and gender (i.e., a P600), even though we probed gender with nouns that do not provide strong distributional cues to gender (unlike previous ERP studies examining gender in L2 Spanish). Here again, however, the L2ers showed a quantitative advantage for number over gender (i.e., a larger P600). In addition, the L2ers' offline knowledge of lexical gender predicted their sensitivity to gender agreement in online comprehension (as measured by d-prime scores), even after controlling for proficiency. Surprisingly, however, knowledge of lexical gender did not predict the magnitude of the P600 to gender violations (e.g., Meulman et al., 2016), which might be due to individual differences with respect to processing strategy (e.g., Tanner et al., 2014; see also Tanner and Van Hell, 2014). We come back to this point in the *General Discussion*.

Interestingly, the L2ers showed no evidence of reliance on default morphology for either number or gender agreement in the judgment task. The ERP data, however, suggest that markedness modulates online processing. Similar to the Spanish native speakers in Alemán Bañón and Rothman (2016), the P600 for gender violations emerged earlier for feature clash than default errors (although in the native speaker group this effect also emerged for number). Likewise, the P600 for number violations was found to be marginally larger for feature clash than default errors. Both findings are consistent with the possibility that feature clash errors are more disruptive than default errors in online comprehension. Importantly, however, the native speakers' ERP responses, though more complex (i.e., the P600 emerged earlier for feature clash number errors too), went in the same direction (e.g., López Prego and Gabriele, 2014).

Two findings from the EEG task merit some discussion. First, similar to the native controls in Alemán Bañón and Rothman (2016), both number and gender violations yielded a late anterior negativity (with a left hemisphere bias) relative to grammatical sentences in the P600 time window. In line with previous studies (Gillon-Dowens et al., 2010; Sabourin and Stowe, 2004; Zawiszewski, Santesteban, and Laka, 2014), this late negativity might reflect the cost of keeping the ungrammaticalities in working memory for the purposes of providing the grammaticality judgment, especially since the learners' mean accuracy in the ungrammatical conditions was high (suggesting that they successfully maintained their judgments in working memory).[8]

In addition, similar to the native controls in Alemán Bañón and Rothman (2016), number violations showed a trend towards a left anterior negativity in the time window associated with the LAN (250-450ms). In the learners, this effect did not remain significant after correcting for Type I error. As discussed above, it is possible that number errors that are realized on plural (i.e., marked) elements modulate the processes reflected by this component. This would still suggest that adult L2 learners are sensitive to markedness asymmetries in a native-like manner. However, the variability with respect to LAN elicitation in the native speaker literature and the fact that this effect remained numerical preclude us from drawing strong conclusions.

In sum, the results from Experiment 1 indicate that, at the upper levels of proficiency, adult L2 learners' online comprehension is qualitatively native-like, even

[8] An alternative interpretation discussed in Alemán Bañón and Rothman (2016) is that the late negativity reflects a polarity inversion of the P600. Since the P600 showed a slight right-hemisphere bias in both populations, it is possible that the dipole generating the P600 was oriented in such a way that its positive and negative ends were detected by right posterior and left anterior electrodes, respectively (e.g., Barber and Carreiras, 2005).

for properties that are unique to the L2, although these appear to be harder. Along these lines, our results also suggest that difficulty with the online processing of gender agreement (property that is unique to the L2) is more tied to lexical (i.e., assignment), than syntactic (i.e., agreement) aspects. Our findings also suggest that adult L2ers do not systematically resort to the use of morphological defaults, at least in online comprehension. One possibility is that, as suggested by the lexically-based accounts of variability (e.g., Grüter et al., 2012; Hopp, 2013; Prévost and White, 2000), morphological variability is more tied to the difficulty associated with the retrieval of lexical information in spoken production. We examine this question in Experiment 2.

Experiment 2: Elicited Production

The L2ers completed Experiment 2 after the second EEG recording. The experiment involved a spot-the-difference task aimed at eliciting determiner-noun-adjective agreement. For each trial, the learners saw two characters holding items that differed with respect to some visible property, and their task was to describe what was different between the items. For example, one trial depicted a character holding one clean suit and another character holding two dirty suits (Figure 8 shows an example). After the instructions, participants completed two practice trials, for which they received no feedback. The task involved 10 instances of agreement with a masculine noun and 10 with a feminine one; 10 instances of agreement with a singular noun and 10 with a plural one. All of the nouns were selected from the comprehension task and, therefore, they do not provide strong distributional cues to gender. To ensure that learners used these target nouns, they were spelled out in bare form (i.e., without a gender-marked determiner). This prevented learners from substituting masculine for feminine nouns (or the reverse). An additional twenty trials were added to the task as distractors, which

depicted the characters engaged in different actions (running in the morning vs. at night). With these materials, we created two separate lists, which differed with respect to the order of presentation of the trials. Participants were randomly assigned to one of the two lists.

<Insert Figure 8>

For the purposes of analysis, the L2ers' responses were transcribed by a native speaker of Spanish and coded twice, once for accuracy with number and once for accuracy with gender. Responses were coded for accuracy according to syntactic context (determiner-noun, noun-adjective) and feature specification (number: singular vs. plural noun; gender: masculine vs. feminine noun). Responses without a noun (e.g., *lo que tiene Ana* "what Ana is holding") were excluded from analysis, as were responses without an adjective (e.g., *paquete que pesa mucho* "packet that weighs a lot"). Cases where learners used an invariable adjective for gender (e.g., *enorme* "huge") were excluded from the gender analysis, but retained for the number analysis. Since the task was quite constraining, all of these cases were rare.

Results.

Table 3 shows the learners' accuracy with the production of both number and gender agreement, according to syntactic context and feature specification. We point out that, for items depicting plural nouns, the L2ers generally used numerals instead of determiners, or used the noun in the singular and provided correct singular inflection on determiners and adjectives. We therefore exclude this cell (number: DET + marked N) from analysis. While this prevents us from comparing the incidence of default vs.

feature clash number errors in this syntactic context, examination of the rest of the number conditions reveals an otherwise clear picture; the L2ers were highly accurate with number and, at least for noun-adjective agreement, there was no evidence for reliance on defaults. These results are similar to previous studies that have examined the production of number agreement in L1-English L2-Spanish learners at similar levels of proficiency (e.g., Franceschina, 2005; White et al., 2004).

<Insert Table 3 here>

Table 3 also shows that the L2ers were quite accurate with the production of gender agreement, although they showed some variability. A paired-samples t-test confirmed that the L2ers' overall accuracy with number was higher than for gender ($t(21) = 5.435$, $p < .001$). With respect to error type, the learners made more default than feature clash gender errors in both syntactic contexts (see Table 3). In order to compare the likelihood of both types of gender errors, we ran logistic mixed-effects regression with Accuracy as the dependent variable, and Noun Gender (Feminine Noun vs. Masculine Noun) and Proficiency Test Score as fixed effects. The random effect structure included random intercepts for participants and items. For both syntactic contexts, the results of the model showed that the likelihood of the two error types did not reliably differ (determiner-noun, number of observations: 218; estimate = 8.01, SE: 5.07, $z = 1.58$, $p > .1$; noun-adjective, number of observations: 429; estimate = 4.05, SE: 4.41, $z = 1.19$, $p > .1$). In addition, the L2ers' accuracy increased as a function of proficiency (determiner-noun, estimate = .016, SE: .06, $z = 2.81$, $p < .01$; noun-adjective, estimate = .016, SE: .051, $z = 3.08$, $p < .01$), but the interaction between Error Type and Proficiency Test Score was not significant (determiner-noun, estimate = -0.14, SE: .012,

z = -1.19, p > .1; noun-adjective, number of observations: 429; estimate = -0.06, SE: 0.08, z = -0.76, p > .1). These results suggest that variability with gender agreement in spoken production is accounted for by proficiency, but not by reliance on default morphology.

It must be pointed out that the above analysis on gender error type is blind to whether the learners knew the lexical gender of the target nouns (e.g., see also Montrul et al., 2008). This is because there is no unproblematic way of determining an L2er's choice of lexical gender. In previous studies, the gender of the determiner has been taken as an indication of a learner's gender assignment (e.g., Carroll, 1989). This would mean that utterances like *un flor fea* "a$_{\text{-MASC}}$ flower$_{\text{-FEM}}$ ugly$_{\text{-FEM}}$", where the feminine noun *flor* "flower$_{\text{-FEM}}$" shows incorrect masculine inflection on the preceding determiner *un* "an$_{\text{-MASC}}$" but correct feminine inflection on the adjective *fea* "ugly$_{\text{-FEM}}$", should be analyzed as a feature clash error. This is because, based on the L2er's choice of determiner, we would assume that she assigned masculine gender to the noun and then incorrectly provided feminine inflection on the adjective. However, it is equally possible that the agreement failure happened between the determiner and the noun. That is, the difficulty associated with lexical retrieval might have caused the L2er to select the wrong determiner before accessing the target noun and the relevant gender information. Then, once this information is retrieved, the learner correctly establishes agreement on the adjective. Such an error would better qualify as a default error. Thus, our logistic regression analysis tells us whether, upon encountering a given noun, the learners were more likely to supply masculine (i.e., default) inflection on determiners and adjectives, due to either a problem of agreement or assignment.[9]

[9] An alternative approach would be to use the L2ers' gender assignment from the Gender Assignment Task. However, we found a few cases where the L2ers indicated that a given noun was—for example—

We then conducted an analysis on error type, to examine whether variability with gender is better accounted for by difficulty with assignment or agreement. Bearing in mind the caveats highlighted above, we follow previous studies (e.g., Grüter et al., 2012; Montrul et al., 2008) in classifying as errors of assignment cases where a determiner and an adjective show consistent inflection, but both mismatch the gender of the controller noun (e.g., *un flor feo* "a-MASC flower-FEM ugly-MASC", *una paquete pesada* "a-FEM packet-MASC heavy-FEM"). In contrast, gender mismatches between determiners and adjectives (e.g., *un flor fea* "a-MASC flower-FEM ugly-FEM", *las peces de Ana son muertos* "the-FEM fish-MASC of Ana are dead-MASC") can somewhat safely be considered errors of agreement (e.g., Grüter et al., 2012; Montrul et al., 2008), since no matter what lexical gender was assigned to the noun, the lack of consistency between the two agreement targets reflects a problem at the level of syntactic agreement. An examination of all of the L2ers' errors involving both a determiner and an adjective (50 out of 1320 responses) revealed that both error types were infrequent, but there were more than twice as many errors of assignment (a total of 35) than errors of agreement (a total of 15). This low incidence of agreement errors was also observed in the study by Grüter et al. (2012), although they also found a higher incidence of assignment errors than we did.

Correlational analyses.

masculine in the Gender Assignment Task, but then treated the same noun as if it were feminine in the production task (by providing consistent feminine inflection on determiners and adjectives). While this is compatible with the possibility of a production-based agreement error, it is also compatible with the possibility that the L2er's lexical gender assignment was inconsistent across tasks.

We further explored the relation between the L2ers' knowledge of lexical gender and their accuracy with gender agreement in spoken production via correlational analyses. Accuracy was calculated by collapsing across the gender specification of the nouns, since we found no evidence for asymmetries between the two genders. A hierarchical regression model was used to examine the extent to which the learners' knowledge of lexical gender (Gender Assignment Task Score) predicted their ability to produce gender agreement (Mean Accuracy in Production Task), over and above the effects of L2 proficiency (Proficiency Test Score).[10] In the first step, Proficiency Test Score accounted for a significant amount of the variance in Mean Accuracy in Production Task, $\beta = .61$, $F(1, 20) = 11.878$, $p < .01$, $R^2 = .373$. When Gender Assignment Task Score was included in the second step, the model also explained a significant proportion of the variance in Mean Accuracy in Production Task, $F(2, 19) = 8.281$, $p < .01$, $R^2 = .466$, R^2 *adjusted* = .409 (see Figure 9, Plot A), and the R^2 *change* was marginal ($p = .085$). Examination of the standardized coefficients shows that both Gender Assignment Task Score and Proficiency Test Score marginally predicted the L2ers' accuracy in establishing gender agreement in production (Gender Assignment Task Score: $\beta = .386$, $t(21) = 1.82$, $p = .085$; Proficiency Test Score: $\beta = 3.74$, $t(21) = 1.76$, $p = .094$).

<Insert Figure 9>

[10] An analysis of standardized residuals showed that the data contained no outliers (Standardized Residual Minimum = -2.34, Maximum = 1.9). As mentioned in footnote 3, the data met the assumption of no perfect multicollinearity and the assumption of independent errors. The histogram of standardized residuals suggested that the data contained approximately normally distributed errors. This was also the case for the P-P plot of standardized residuals, which showed points very close to the regression line. The scatterplot of standardized predicted values showed that the assumptions of homogeneity of variance and linearity were also met.

Relation between comprehension and production.

Finally, we compared the L2ers' mean accuracy in comprehension and production in order to examine whether there was an overall advantage for comprehension (i.e., receptive knowledge) over production (i.e., expressive knowledge). For this comparison, accuracy was operationalized as mean accuracy rates (comprehension: mean accuracy rates in the violation conditions in the Grammaticality Judgment Task; production: mean accuracy rates). Since the L2ers were at ceiling with number in production, this analysis was limited to gender. In addition, since we found no difference between the two types of gender errors, accuracy was calculated by collapsing across them. A paired-samples t-test revealed no accuracy differences between the two tasks, $t(21) = 1.007$, $p > .1$.

Lastly, we calculated the zero order Pearson correlation between the L2ers' mean accuracy with gender in comprehension and production, to examine whether robust lexical representations for gender translated into target-like performance with gender across tasks (Hopp, 2013). This correlation was positive, strong, and highly significant, $r = .6$, $p < .001$ (see Figure 9, plot B).

Interim Discussion of Experiment 2.

Learners were very accurate with both number and gender agreement, although they performed better with number (i.e., at ceiling). As for gender, two findings are particularly relevant. First, assignment errors were more frequent than agreement errors. That is, learners showed greater difficulty with lexical (as opposed to syntactic) aspects of gender. In addition, correlational analyses showed that the L2ers' accuracy with gender in production increased as a function of their knowledge of lexical gender and

their proficiency, although the individual contribution of each predictor was marginal. Both of these findings suggest difficulty at the level of lexical gender assignment. The L2ers' better performance with number over gender, which we also observed in Experiment 1, provides further support that there is facilitation for properties that exist in the learners' L1 (e.g., Schwartz and Sprouse, 1996).

As was the case in Experiment 1, the L2ers showed no evidence of reliance on default morphology for either number or gender agreement. In the case of number, the learners performed at ceiling (e.g., White et al., 2004). In the case of gender, the learners made more default than feature clash errors, but analyses revealed no reliable tendency to overuse the default gender, either at the level of assignment or agreement.

General Discussion

We investigated the nature of morphological variability in adult L1-English learners of L2-Spanish of intermediate to advanced proficiency. The main aim of the study was to examine specific factors which, according to contrasting L2 theories, account for inflectional variability in adult L2ers, notably (i) morphological markedness, (ii) the properties of the learners' L1, and (iii) the type of knowledge at use (receptive, as in comprehension vs. expressive, as in spoken production).

To this aim, we conducted two experiments with the same group of adult L1-English L2-Spanish learners. Experiment 1 made use of ERP to examine the online comprehension/processing of noun-adjective number and gender agreement. Unlike previous ERP studies on L2 morphosyntactic processing, our design examined the unique contribution of markedness to agreement resolution. We did so by systematically manipulating the markedness of the trigger nouns, such that half of them were marked (number: plural; gender: feminine), and the other half, underspecified. In addition,

unlike all previous ERP studies that have examined gender in L2 Spanish (Alemán Bañón et al., 2014; Gabriele et al., 2013; Gillon-Dowens et al., 2010; Rossi et al., 2014; Tokowicz and MacWhinney, 2005), all of which show qualitatively native-like processing for gender, we focused on Spanish nouns that do not provide strong distributional cues to gender. We expected gender agreement with these nouns to be more challenging for adult L2ers, based on previous behavioral (e.g., Franceschina, 2005; Grüter et al., 2012; Montrul et al., 2008) and ERP studies (e.g., Foucart and Frenck-Mestre, 2012; Meulman et al., 2016; Sabourin, 2003; Sabourin and Stowe, 2008). Experiment 2 used a spot-the-difference task to examine similar dependencies (determiner-noun-adjective number and gender agreement) in elicited spoken production. Here again, we systematically manipulated the markedness of the trigger nouns. The specific research questions that inform the study are repeated below, alongside the predictions by the most relevant L2 theories:

(i) To what extent is variability accounted for by the learners' reliance on default morphology? We addressed this question by comparing instances of agreement where the trigger noun carried marked vs. underspecified feature values. Recall that computational accounts like the MSIH (e.g., Haznedar and Schwartz, 1997; Prévost and White, 2000) predict that adult L2ers might show evidence of reliance on default morphology in spoken production (Experiment 2), as a result of the computational pressure associated with the online retrieval of inflectional forms and other lexical information, such as lexical gender (Grüter et al., 2012; Hopp, 2013). In contrast, McCarthy (2008) predicts overuse of default morphology across tasks (Experiment 1 and 2), given that learners are hypothesized not to be able to acquire the full specification of features at the level of the morphology, due to a representational deficit.

Our results revealed no reliable evidence that L2ers resorted to the use of morphological defaults for either number or gender agreement, either in online comprehension (Experiment 1) or in spoken production (Experiment 2). We begin with the results of the oral production task, where both the MSIH (e.g., Prévost and White, 2000) and McCarthy (2008) predicted a certain reliance on default morphology (i.e., a higher number of default than feature clash errors). In the case of number, the learners performed at ceiling, replicating the results by White et al. (2004), who found ceiling performance with noun-adjective number agreement in spoken production in L1-English L2-Spanish learners of similar proficiency (i.e., intermediate-advanced). However, our results contrast with those by McCarthy (2008), who found that both intermediate and advanced learners used singular agreement (i.e., zero inflection) on adjectives in the context of a plural noun. Overall, these findings suggest that, at the upper levels of proficiency, number agreement in spoken production is relatively unproblematic, even when it is realized in a syntactic context where the learners' L1 does not mark number (i.e., the adjective).

The results of the gender conditions are more complex, given that learners may overuse the default gender both at the level of lexical gender assignment (i.e., assigning masculine gender to nouns whose gender they felt uncertain about; see Grüter et al., 2012 and Montrul et al., 2008) and at the level of agreement, even when lexical gender has been properly assigned. The results of our logistic regression analysis revealed that learners had no tendency to overuse the default gender at either level. These results are at odds with previous L2 studies on gender. For example, Montrul et al., (2008) examined gender agreement in spoken production in a group of adult L1-English L2-Spanish learners, and found a higher error rate with feminine than masculine nouns (i.e., more default errors) both at the level of syntactic gender agreement and lexical

gender assignment (see also Grüter et al., 2012). One important difference between our study and Montrul et al.'s concerns L2 proficiency, which was measured with the same instrument that we used in the present study. While the proficiency range in Montrul et al.'s study was quite wide and included low-proficiency learners (16-50; mean: 36), the learners in our study were of upper-intermediate to advanced proficiency (range: 33-50; mean: 43). It is, therefore, possible that reliance on default morphology in spoken production is more characteristic of an interlanguage stage that L2ers eventually overcome with increased proficiency, which is not consistent with the idea of a representational deficit at the level of the morphology. Our results are also not consistent with those by McCarthy (2008) and White et al. (2004). In both studies, L2 learners showed higher error rates with feminine than masculine nouns (i.e., more default errors), although this asymmetry was especially characteristic of learners at the intermediate level of proficiency, not so much of advanced learners. In our study, we did find that proficiency significantly impacted the learners' accuracy with gender agreement, but we found no reliable interaction between proficiency and error type, suggesting that agreement was largely unaffected by markedness across the proficiency range examined.

Moving on to the comprehension data, the results of the Grammaticality Judgment Task revealed that markedness impacted each feature type differently, as indicated by a significant feature by markedness interaction. For number, learners were more accurate rejecting feature clash than default errors (e.g., in line with McCarthy, 2008), but they showed the reverse pattern for gender (contra McCarthy, 2008 and White et al., 2004). Follow-up tests, however, failed to confirm these feature-value asymmetries for gender, and the effect was only marginal for number, suggesting that agreement resolution for each feature in isolation was somewhat unaffected by markedness. It is noteworthy,

however, that a similar interaction emerged in a related study by López-Prego and Gabriele (2014). The authors used a speeded grammaticality judgment task to investigate how markedness impacted the processing of number and gender agreement in L1-English L2-Spanish learners. Similar to the present study, agreement was examined between nouns and adjectives located across a restrictive relative clause boundary (e.g., *una tela que era fina* "a$_{\text{-FEM}}$ fabric$_{\text{-FEM}}$ that was fine$_{\text{-FEM}}$"). Their results parallel those in the present study. That is, for number violations, intermediate and advanced learners were more accurate rejecting feature clash than default errors. For gender, however, the learners showed the opposite pattern. To account for these effects, López-Prego and Gabriele (2014) highlight the marked status of the trigger nouns in the case of gender default errors. Since default errors in their design involved a feminine (i.e., marked) DP followed by a masculine (i.e., unmarked) adjective (e.g., *una tela que era *fino* "a$_{\text{-FEM}}$ fabric$_{\text{-FEM}}$ that was fine$_{\text{-MASC}}$"), they propose that DPs that are marked for gender might have greater predictive value than DPs that are underspecified for gender (e.g., Nevins et al., 2007; Wagers and McElree, 2011). Under this account, when the parser encounters the marked features of the trigger DP, it can more reliably predict the gender of the upcoming adjective. Such facilitation results in a more accurate detection of default errors (in a judgment task, at least). The significant feature by markedness interaction that we found in the present study is consistent with this proposal, although it remains an open question why DPs that are marked for number (i.e., plural) do not have the same predictive value as those that are marked for gender. Given that number and gender differ with respect to their status in the L1 feature inventory, one possibility is that learners' predictive strategies are more likely to be recruited for novel properties, given their greater computational demands.

With respect to the ERP data, our results revealed that markedness did impact online processing, but in a native-like manner. That is, similar to the Spanish native speakers reported in Alemán Bañón and Rothman (2016), the P600 for gender violations emerged earlier for feature clash than default errors, consistent with the possibility that errors that involve incompatible features at the level of the morphology are more disruptive and easily detectable (although in the native speakers, an earlier P600 also emerged for feature clash number errors). Likewise, the P600 for number violations was found to be marginally larger for feature clash than default errors in the region where the P600 reached its maximum (i.e., Right Posterior), consistent with the possibility that feature clash errors are more salient and disruptive. Importantly, however, the fact that the native speakers' ERP responses went in the same direction suggests that the L2 data cannot be taken as support for a representational deficit at the level of the morphology (contra McCarthy, 2008), but rather as evidence that L2ers are sensitive to markedness asymmetries. Notice also that a similar pattern has been reported in other studies that have examined the role of markedness on agreement in native speakers (e.g., Deutsch and Bentin, 2001; Kaan, 2002). Similar findings are also reported in López Prego and Gabriele's study (2014), who found that, under high processing burden, native speakers were more accurate rejecting feature clash than default errors for both number and gender, suggesting that sensitivity to markedness is not restricted to adult L2ers, but can also characterize native processing under computational burden.[11]

[11] An alternative interpretation for this pattern of results is that default errors were less disruptive because the English equivalent corresponds to a correct structure (e.g., *uniforms that looked dirty*), due to the fact that English does not realize number on adjectives. While this interpretation cannot be completely ruled out, the fact that number default errors yielded a P600 relative to grammatical sentences suggests that the L2ers were not exclusively relying on the properties of English. Notice that the English equivalent of the

One reviewer wondered whether the lack of evidence for the adoption of a default gender in the L2 group might be due to individual differences, with some learners adopting masculine as a default (in line with morphological theory) and others adopting feminine. To evaluate this possibility, we carried out an exploratory analysis on D-prime scores (Grammaticality Judgment Task) and P600 effect size, which revealed that most learners and native speakers were equally accurate with and yielded equally robust P600 effects to the two types of gender errors. Importantly, although some learners showed greater sensitivity to default errors, and others to feature clashes, the same pattern emerged in the L1 group, suggesting that individual differences with respect to the (potential) adoption of a default are not restricted to L2ers. We think that this is also not in line with representational accounts of variability, although we highlight that these analyses are very exploratory, since a larger sample is needed to identify a bimodal population.

(ii) To what extent is morphological variability determined by the properties of the learners' L1? We addressed this question by comparing the learners' performance with number agreement (present in the L1 feature inventory) and gender agreement (unique to their L2). Representational accounts like the Interpretability Hypothesis (e.g., Tsimpli and Dimitrakopoulou, 2007) predict an overall advantage for number over gender. Crucially, native-like processing in terms of brain responses to agreement violations

---

grammatical sentences, where both the noun and the adjective show plural morphology (e.g., *uniformes que parecían sucios* "uniform-$_{PL}$ that looked dirty-$_{PL}$") corresponds to an impossible string in English. Yet, it was number default errors that were more positive than grammatical sentences, and not the other way around. We interpret this as evidence that the L2ers treated singular adjectives in plural contexts as deviant, and plural adjectives in plural contexts (i.e., the configuration that is disallowed in English) as licit.

(i.e., P600) is predicted to be possible for number, but not gender, especially since the nouns we used did not provide strong distributional cues to gender. Under computational accounts like the MSIH (e.g., Haznedar and Schwartz, 1997; Prévost and White, 2000), L2ers are predicted to be able to show native-like processing for both features, at least in comprehension. It is also possible that, at this level of proficiency, there will be no difference between number and gender.

An additional question that we examined concerns the relation between the L2ers' knowledge of lexical gender and their ability to compute gender agreement in online comprehension and production, in order to better adjudicate between proposals which argue for a deficit at the level of syntactic agreement (e.g., Tsimpli and Dimitrakopoulou, 2007) and proposals which argue for problems at the level of lexical assignment and retrieval (Grüter et al., 2012; Hopp, 2013; Prévost and White, 2000).

In both comprehension (Experiment 1) and production (Experiment 2), learners showed high accuracy rates with both number and gender agreement, although they performed better with number than gender (e.g., Franceschina, 2005; Gillon-Dowens et al., 2010). Importantly, however, their brain responses as revealed by the ERP data showed qualitatively native-like processing for the two features (i.e., a P600) (e.g., Alemán Bañón et al., 2014; Gillon-Dowens et al., 2010), although here again they showed a quantitative advantage for number (i.e., a larger P600), a difference that did not arise in the Spanish native speakers reported in Alemán Bañón and Rothman (2016). Although the P600 is not exclusively linked to morphosyntactic processing, it is the component that is most consistently associated with agreement processing in native speakers. Thus, the fact that learners were qualitatively native-like with the processing of gender agreement, the property that is unique to their L2, seems at odds with theoretical accounts which argue that novel syntactic properties cannot be acquired to

native-like levels due to a representational deficit at the level of the syntax (e.g., Tsimpli and Dimitrakopoulou, 2007), especially if we bear in mind that the nouns we used did not allow for the use of phonological rhyming strategies between noun endings and inflectional forms (unlike previous ERP studies which have examined gender agreement in L2 Spanish). Along similar lines, the L2ers' offline knowledge of lexical gender (as measured by the Gender Assignment Task) was found to be a reliable predictor of their accuracy with gender agreement in online comprehension (as measured by D-prime Scores for the gender conditions in Experiment 1), even after controlling for proficiency. Further correlational analyses showed that the L2ers' accuracy with the production of gender agreement (Experiment 2) increased as a function of their knowledge of lexical gender (as measured by the Gender Assignment Task) and their proficiency (e.g., Montrul et al., 2008), although the individual contribution of each predictor remained marginal. In addition, our results revealed a strong positive relation between the learners' accuracy with gender in comprehension and production, consistent with the idea that knowledge of lexical gender determines the learners' performance with gender across tasks (Hopp, 2013). The fact that the L2ers' accuracy with gender agreement across tasks is better accounted for by their overall knowledge of lexical gender is more in line with proposals which attribute inflectional variability to the quality of the L2ers' lexical representations for gender (Grüter et al., 2012; Hopp, 2013), but not with representational accounts of variability, which predict that L2ers can reach target-like knowledge of lexical gender, but still not be able to establish agreement in a native-like manner.

The reader might wonder whether processing data, such as ERP, constitute a valuable metric to test the predictions of representational accounts, which are mainly concerned with representation. Our take on this is that processing data are precisely the

type of evidence that is needed. To give one example, the proponents of the Interpretability Hypothesis have shown that L1-English L2-Spanish learners can achieve very high accuracy rates with gender agreement in Spanish in offline tasks (e.g., Franceschina, 2005). But crucially, they claim that learners achieve these high accuracy rates by using alternative mechanisms. Therefore, the representational accounts clearly posit that what is different is the underlying process through which learners arrive at such performance outcomes. The use of ERPs can shed light on the qualitative nature of those mechanisms (see also Alemán Bañón et al., 2014).

It could be argued that the presence of a gender-marked determiner preceding the critical noun in the comprehension task (which could not be avoided, given that Spanish generally disallows bare nominals) might have facilitated gender resolution in the L2 group. This is because the L2ers could have used the determiner as a cue to assign lexical gender. While this possibility cannot be completely ruled out, the learners' high accuracy in the Gender Assignment Task suggests that the they could successfully assign lexical gender in the absence of a gender-marked determiner. Likewise, the fact that the L2ers' scores in the Gender Assignment Task (where no determiner was available) predicted their accuracy with gender agreement in the comprehension task also indicates that gender agreement was mediated by knowledge of lexical gender. Moreover, as illustrated in example (1) of the introduction, L2 learners often correctly establish gender agreement between the article and the noun, yet continue to make agreement errors down the line (especially when the agreeing words belong to different syntactic phrases, as is the case in the present study). Therefore, the availability of the determiner does not necessarily provide a reliable cue for the learner, unless the underlying representation is established for the property.

One surprising finding from Experiment 1 is that the L2ers' knowledge of lexical gender (as measured by their score in the Gender Assignment Task) predicted their sensitivity to gender agreement in the Grammaticality Judgment Task, but not the size of the P600 to gender violations (e.g., Meulman et al., 2016). One potential explanation for the lack of a relationship between these two measures concerns individual differences with respect to processing strategy. Indeed, cases have been reported where L2ers show an N400 for the same types of agreement errors for which other learners show a P600 (e.g., Tanner et al., 2014; see also Tanner and Van Hell, 2014 for similar findings in native speakers), which has been interpreted as evidence for individual differences with respect to processing strategy (lexically-based vs. rule-based). In our study, most of the L2ers showed a positivity for both types of gender violations, which explains why this effect was the only one to emerge in the group analysis. However, a subset of the learners elicited negative effects for the same gender errors. One possibility is that these negative responders knew the lexical gender of the target nouns and detected the violations, which would explain why their brain was sensitive to the violations (in the form of a negativity) and why they showed high accuracy in the Grammaticality Judgment Task, but relied on a different strategy to establish gender agreement. In turn, this would explain the lack of a significant correlation between the L2ers' score in the Gender Assignment Task and the size of the P600 for gender errors. We checked this possibility by comparing the scores in the Gender Assignment Task of the four L2ers with the largest positivity to the four L2ers with the largest negativity for each gender violation condition, and we found roughly similar scores. In light of this, we calculated the correlation between the L2ers' score in the Gender Assignment Task and the absolute magnitude of their brain responses to the gender violations (i.e.,

regardless of polarity), but these correlations were not significant.[12] It is thus possible that factors other than individual differences account for the lack of a relationship between P600 size and knowledge of lexical gender.

Although our data suggest that native-like processing in the L2 is not constrained by the properties of the L1, the learners still showed a quantitative advantage for number (instantiated in the L1) over gender in the brain data (i.e., P600 magnitude), and also in the more explicit tasks (the Grammaticality Judgment Task and the production task). These findings are consistent with previous ERP studies which have compared the two features in L1-English L2-Spanish learners, such as Gillon-Dowens et al. (2010) and Rossi et al. (2014). They are also consistent with the results by Alemán Bañón et al. (2014), although in their study the quantitative advantage for number only emerged in the L2ers' brain responses, not in the accuracy data. Importantly, this facilitation for number cannot be attributed to markedness differences in the way number and gender were compared, since we systematically manipulated this factor (unlike Alemán-Bañón et al., 2014 and Gillon-Dowens et al., 2010). Overall, our findings are consistent with theoretical models which assume facilitation for properties that exist in the learners' native language (e.g., Schwartz and Sprouse, 1996), but not with representational accounts of variability.

(iii) Is morphological variability a production-specific phenomenon or does it also emerge in comprehension? We addressed this question by comparing the learners' performance in online comprehension and production (receptive vs. expressive knowledge). Computational accounts like the MSIH (e.g., Prévost and White, 2000) predict a general advantage for comprehension (Experiment 1) over spoken production (Experiment 2). Under this account, it is in spoken production where L2ers might show

[12] We thank Darren Tanner for this suggestion.

reliance on default morphology and where variability with gender agreement is more likely to emerge, due to the burden associated with lexical retrieval (Grüter et al., 2012; Hopp, 2013). In contrast, representational accounts, such as the Interpretability Hypothesis (e.g., Tsimpli and Dimitrakopoulou, 2007) and McCarthy (2008), predict variability across the board (albeit for different reasons).

The learners in the present study performed similarly in online comprehension and spoken production, meaning that they showed no advantage in receptive vs. expressive knowledge of inflectional morphology. For number, the learners even displayed a small advantage in production, where they showed virtually no variability, although their scores in the comprehension task (Grammaticality Judgment Task) were also very high. In the case of gender, the learners showed approximately equal accuracy rates across experiments. These results are at odds with previous studies which have reported an advantage for comprehension over production in adult L2ers (e.g., Montrul et al., 2008). In these studies, however, the difference between comprehension and production was confounded with the online nature of the task. That is, comprehension was tested offline, whereas spoken production was examined online. In the present study, both comprehension and production were probed online. One potential explanation for the lack of a comprehension advantage in our study is that, as suggested by Grüter et al. (2012), learners tend to perform better in offline tasks, regardless of the type of knowledge that the task taps into (i.e., receptive vs. expressive). There are, however, certain differences between the comprehension and production tasks in the present study that might account for the lack of an advantage for comprehension. For example, while the nouns and adjectives in the comprehension task were located across a CP (i.e., a nonlocal domain), learners tended to establish agreement locally (i.e., within a Determiner Phrase) in the production task. It is, therefore, possible that the more taxing

syntactic configuration in the comprehension task reduced a potential overall advantage for comprehension over production. In addition, our design does not compare auditory comprehension vs. oral production, but rather compares reading vs. speaking. Thus, these methodological differences complicate to some extent a direct comparison between comprehension and production. That said, we note that previous studies that have also compared reading vs. oral production have reported an advantage for comprehension (e.g., Grüter et al., 2012; Montrul et al., 2008), an effect which we did not find.

Conclusion

The present study finds that, at the upper levels of proficiency, adult L2 learners can process both number and gender agreement in a native-like manner, even when their L1 (i.e., English) is gender-free, and most importantly, even for nouns that do not provide strong distributional cues for gender. Our results also suggest that learners' mastery of syntactic gender agreement depends upon the learners' ability to correctly assign nouns to their appropriate gender classes, in line with recent lexically-based accounts of inflectional variability. Our study also provides evidence that, at least at the upper levels of proficiency, adult L2 learners can acquire the full specification of all features at the level of the morphology and do not need to resort to morphological defaults, either in online comprehension or spoken production. Most importantly, our results show that learners are sensitive to markedness distinctions in a native-like manner.

References

Alarcón, I. V. (2011). Spanish gender agreement under complete and incomplete acquisition: Early and late bilinguals' linguistic behavior within the noun phrase. *Bilingualism: Language and Cognition, 14*(03), 332-350

Alemán Bañón, J., Fiorentino, R., & Gabriele, A. (2012). The processing of number and gender agreement in Spanish: An event-related potential investigation of the effects of structural distance. *Brain Research, 1456*, 49-63.

Alemán Bañón, J., Fiorentino, R., & Gabriele, A. (2014). Morphosyntactic processing in advanced second language (L2) learners: An event-related potential investigation of the effects of L1–L2 similarity and structural distance. *Second Language Research, 30*(3), 275-306.

Alemán Bañón, J. & Rothman, J. (2016). The Role of Morphological Markedness in the Processing of Number and Gender Agreement in Spanish: An Event-Related Potential Investigation. *Language, Cognition, and Neuroscience*. http://dx.doi.org/10.1080/23273798.2016.1218032

Barber, H., & Carreiras, M. (2005). Grammatical gender and number agreement in Spanish: An ERP comparison. *Journal of Cognitive Neuroscience, 17*(1), 137-153.

Battistella, E. L. (1990). *Markedness: The evaluative superstructure of language*. SUNY Press.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Bond, K. (2013). *The role of the L1 and individual differences in L2 sensitivity to morphosyntactic features: An ERP investigation* (doctoral dissertation, University of Kansas).

Bond, K., Gabriele, A., Fiorentino, R., & Bañón, J. A. (2011). Individual differences and the role of the L1 in L2 processing: An ERP investigation. In D. Tanner, & J. Herschensohn (Eds.), *Proceedings of the 11th Generative Approaches to Second Language Acquisition Conference* (pp. 17-29). Somerville, MA: Cascadilla Press.

Bonet, E. (1995). Feature structure of Romance clitics. *Natural Language & Linguistic Theory*, *13*(4), 607-647.

Carroll, S. (1989). Second-Language Acquisition and the Computational Paradigm. *Language Learning*, *39*(4), 535-594.

Carstens, V. (2000). Concord in minimalist theory. *Linguistic Inquiry*, *31*(2), 319-355.

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

Clahsen, H. & Felser, C. (2006). Continuity and shallow structures in language processing. *Applied Psycholinguistics, 27*, 107-26.

Clahsen, H., Felser, C., Neubauer, K., Sato, M., & Silva, R. (2010). Morphological structure in native and nonnative language processing. *Language Learning, 60*, 21-43.

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

Coulson, S., King, J. W., & Kutas, M. (1998). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and Cognitive Processes, 13*(1), 21-58

Cowper, E. (2005). The geometry of interpretable features: Infl in English and Spanish. *Language*, *81*(1), 10-46.

Deutsch, A., & Bentin, S. (2001). Syntactic and semantic factors in processing gender agreement in Hebrew: Evidence from ERPs and eye movements. *Journal of Memory and Language*, *45*(2), 200-224.

Dewaele, J. M., & Véronique, D. (2001). Gender assignment and gender agreement in advanced French interlanguage: A cross-sectional study. *Bilingualism: Language and Cognition*, *4*(03), 275-297.

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods*, *45*(4), 1246-1258.

Field, A. (2005). Discovering statistics with SPSS. London: Sage.

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*(1), 226-248.

Franceschina, F. (2001). Morphological or syntactic deficits in near-native speakers? An assessment of some current proposals. *Second Language Research*, *17*(3), 213-247.

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

Gabriele, A., Fiorentino, R., & Alemán Bañón, J. (2013). Examining second language development using event-related potentials: a cross-sectional study on the

processing of gender and number agreement. *Linguistic Approaches to Bilingualism*, *3*(2), 213-232.

Gillon-Dowens, M., Vergara, M., Barber, H. A., & Carreiras, M. (2010). Morphosyntactic processing in late second-language learners. *Journal of Cognitive Neuroscience*, *22*(8), 1870-1887.

Gillon-Dowens, M., Guo, T., Guo, J., Barber, H., & Carreiras, M. (2011). Gender and number processing in Chinese learners of Spanish–Evidence from Event Related Potentials. *Neuropsychologia*, *49*(7), 1651-1659.

Grüter, T., Lew-Williams, C., & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research*, *28*(2), 191-215.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes*, *8*(4), 439-483.

Hahne, A., & Friederici, A. D. (1999). Electrophysiological evidence for two steps in syntactic analysis: Early automatic and late controlled processes. *Journal of Cognitive Neuroscience*, *11*(2), 194-205.

Halle, M., & Marantz, A. (1993). Distributed morphology and the pieces of inflection. In K. Hale, & J. Keyser (Eds.), *The View from the Building 20: Essays in Linguistics in Honor of Sylvain Bromberger* (pp. 1–52). Cambridge, MA: MIT Press.

Harley, H., & Noyer, R. (1999). Distributed morphology. *Glot International*, *4*(4), 3-9.

Harley, H., & Ritter, E. (2002). Person and number in pronouns: A feature-geometric analysis. *Language, 78*(3), 482-526.

Harris, J. W. (1991). The exponence of gender in Spanish. *Linguistic Inquiry, 22*(1), 27-62.

Hawkins, R. (2001). The theoretical significance of Universal Grammar in SLA. *Second Language Research* 17, 345-67.

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.

Haznedar, B., & Schwartz, B. D. (1997). Are there optional infinitives in child L2 acquisition? In E. Hughes, M. Hughes, & A. Greenhill. (Eds.), *Proceedings of the 21st Annual Boston University Conference on Language Development* (pp. 257–68). Somerville, MA: Cascadilla Press.

Hopp, H. (2010). Ultimate attainment in L2 inflection: Performance similarities between non-native and native speakers. *Lingua, 120*(4), 901-931.

Hopp, H. (2013). Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research, 29*(1), 33-56.

Kaan, E. (2002). Investigating the effects of distance and number interference in processing subject-verb dependencies: An ERP study. *Journal of Psycholinguistic Research, 31*(2), 165-193.

Kaan, E., Harris, A., Gibson, E., & Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes, 15*(2), 159-201.

Keating, G. D. (2009). Sensitivity to Violations of Gender Agreement in Native and Nonnative Spanish: An Eye-Movement Investigation. *Language Learning*, *59*(3), 503-535.

Kutas, M., & Hillyard, S. A. (1980) Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207, 203-05.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annual Review of Psychology*, *62*, 621.

Lardiere, D. (1998). Case and tense in a fossilized steady state. *Second Language Research*, 14, 1-26.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de) constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920-933.

Lemhöfer, K., Schriefers, H., & Indefrey, P. (2014). Idiosyncratic grammars: Syntactic processing in second language comprehension uses subjective feature representations. *Journal of Cognitive Neuroscience*, *26*(7), 1428-1444.

Long, M. (2005). Problems with supposed counter-evidence to the Critical Period Hypothesis. *International Review of Applied Linguistics in Language Teaching*, *43*(4), 287-317.

López Prego, B., & Gabriele, A. (2014). Examining the impact of task demands on morphological variability in native and non-native Spanish. *Linguistic Approaches to Bilingualism*, *4*(2), 192-221.

Lyster, R. (2006). Predictability in French gender attribution: A corpus analysis. *Journal of French Language Studies*, *16*(01), 69-92.

McCarthy, C. (2008). Morphological variability in the comprehension of agreement: An argument for representation over computation. *Second Language Research*, 24(4), 459–486.

McLaughlin, J., Tanner, D., Pitkänen, I., Frenck-Mestre, C., Inoue, K., Valentine, G., & Osterhout, L. (2010). Brain potentials reveal discrete stages of L2 grammatical learning. *Language Learning*, *60*(s2), 123-150.

Meulman, N., Stowe, L. A., Sprenger, S. A., Bresser, M., & Schmid, M. S. (2015). An ERP study on L2 syntax processing: When do learners fail? *Frontiers in Psychology*, *5*, 1072.

Molinaro, N., Barber, H. A., & Carreiras, M. (2011). Grammatical agreement processing in reading: ERP findings and future directions. *Cortex*, *47*(8), 908-930.

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*(3), 503-553.

Nevins, A., Dillon, B., Malhotra, S., & Phillips, C. (2007). The role of feature-number and feature-type in processing Hindi verb agreement violations. *Brain Research*, *1164*, 81-94.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, *9*(1), 97-113.

O'Rourke, P. L., & Van Petten, C. (2011). Morphological agreement at a distance: Dissociation between early and late components of the event-related brain potential. *Brain Research*, *1392*, 62-79.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*(6), 785-806.

Osterhout, L., McLaughlin, J., Pitkanen, I., Frenck-Mestre, C., & Molinaro, N. (2006). Novice learners, longitudinal designs, and event-related potentials: A paradigm for exploring the neurocognition of second-language processing. *Language Learning*, 56, 199-230.

Pesetsky, D., & Torrego, E. (2007). The syntax of valuation and the interpretability of features. In S. Karimi, V. Samiian, & W. Wilkins (Eds.), *Phrasal and clausal architecture: Syntactic derivation and interpretation*, (pp. 262-294). 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*(2), 103-133.

Renaud, C. (2011). Constraints on feature selection in second language acquisition: Processing evidence from the French verbal domain. In L. Plonsky & M. Schierloh (Eds.), *Selected Proceedings of the 2009 Second Language Research Forum* (pp. 129-141). Somerville, MA: Cascadilla Press.

Renaud, C. (2012). An investigation of the role of gender in the resolution of pronouns in L2 French. In A. Biller, E. Chung, & A. Kimball (Eds.), *Proceedings of the 36th Boston University Conference on Language Development* (pp. 512-524). Somerville, MA: Cascadilla Press.

Rossi, E., Kroll, J. F., & Dussias, P. E. (2014). Clitic pronouns reveal the time course of processing gender and number in a second language. *Neuropsychologia*, *62*, 11-25.

Rothman, J. (2007). Sometimes they use it, sometimes they don't: An epistemological discussion of L2 morphological production and its use as a competence measurement. *Applied Linguistics, 28*(4), 609-614.

Sabourin, l. (2003). *Grammatical gender and second language processing: An ERP study* (doctoral dissertation, Rijksuniversiteit Groningen).

Sabourin, L., & Stowe, L. (2004). Memory effects in syntactic ERP tasks. *Brain and Cognition*, *55*(2), 392-395.

Sabourin, L., & Stowe, L. (2008). Second language processing: when are first and second languages processed similarly? *Second Language Research, 24*(3), 397-430.

Saporta, S. (1965). Ordered rules, dialect differences, and historical processes. *Language*, *41*(2), 218-224.

Schwartz, B. D., & Sprouse, R. A. (1996). L2 cognitive states and the full transfer/full access model. *Second Language Research*, *12*(1), 40-72.

Séguin, H. (1969). *Les marques du genre dans le lexique du français écrit contemporain: compilation des cas et essai de classement* (Doctoral dissertation, Université de Montréal).

Tagliaferri, B. (2005). Paradigm. Perception Research Systems Inc. http://www. perceptionresearchsystems.com.

Tanner, D., Inoue, K., and Osterhout, L. (2014). Brain-based individual differences in online L2 grammatical comprehension. *Bilingualism: Language and Cognition, 17*, 277-293.

Tanner, D., McLaughlin, J., Herschensohn, J. & Osterhout, L. (2013). Individual differences reveal stages of L2 grammatical acquisition: ERP evidence. *Bilingualism: Language and Cognition*, 16, 367-382.

Tanner, D., & Van Hell, J. G. (2014). ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia*, *56*, 289-301.

Teschner, R. V., & Russell, W. M. (1984). The gender patterns of Spanish nouns: An inverse dictionary-based analysis. *Hispanic Linguistics*, *1*(1), 115-132.

Tokowicz, N., & MacWhinney, B. (2005). Implicit and explicit measures of sensitivity to violations in second language grammar: An event-related potential investigation. *Studies in Second Language Acquisition*, *27*(02), 173-204.

Tsimpli, I. M., & Dimitrakopoulou, M. (2007). The interpretability hypothesis: Evidence from wh-interrogatives in second language acquisition. *Second Language Research*, *23*(2), 215-242.

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.

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory & Cognition*, *18*(4), 380-393.

Wagers, M., & McElree, B. (2011). Memory for linguistic features and the focus of attention: Evidence for the dynamics of agreement. *Unpublished Manuscript, University of California Santa Cruz and New York University*.

White, L., Valenzuela, E., Kozlowska–Macgregor, M., & Leung, Y. K. I. (2004). Gender and number agreement in nonnative Spanish. *Applied Psycholinguistics*, *25*(01), 105-133.

White, L. (2007). Linguistic theory, Universal Grammar, and second language acquisition. In B. van Patten & J Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 37-55). Lawrence Erlbaum.

White, L. (2011). Second language acquisition at the interfaces. *Lingua*, *121*(4), 577-590.

Zawiszewski, A., Santesteban, M., & Laka, I. (2015). Phi-features reloaded: An event-related potential study on person and number agreement processing. *Applied Psycholinguistics*, 1-26.