Ampersand 14 (2025) 100217

Contents lists available at ScienceDirect

Ampersand

journal homepage: www.elsevier.com/locate/amper

Grammatical gender in adolescent German learners of Spanish as a foreign language: Evidence from production and online perception☆

Johanna Wolf *, Clara Terlaak, Sarah Schimke

Faculty of Languages and Literatures, Ludwig-Maximilians-Universität (LMU) München, 80799 Munich, Germany

A B S T R A C T

The acquisition of grammatical gender is known to be a difficult and error-prone process in second and foreign language learning. The current study sheds new light on this topic by looking at elicited narrative production data as well as at data from an eyetracking-during- reading and a controlled gender production task. Participants were German-speaking adolescent learners of Spanish as a foreign language. Our results suggest that learners are often successful in assigning gender and marking agreement, and that they are sensitive to agreement errors during reading. Moreover, we found a relation between the sensitivity that learners displayed to agreement relations during reading and their accuracy of gender markings in a production task with unrelated and more challenging lexical items. We conclude that sensitivity to gender markings during input processing facilitates the acquisition of the category, and that this sensitivity is high in adolescent learners and for a relative transparent gender system as present in Spanish.

1. Introduction

Although gender is very frequent in input, the acquisition of gender systems in foreign language learning remains a challenge for learners, even when they have reached an advanced level (Binanzer, 2017, 2020; Doleschal, 2004, IX; Menzel, 2004, 55; Kiyko and Kiyko, 2020, 169; Wegera, 1996, 19, Carroll, 1999). Strikingly, even increased exposure to the target system through a stay abroad does not necessarily lead to measurable progress in the acquisition of gender, as has been shown by Klassen et al. (2023) for adult English-speaking learners of Spanish during a stay abroad. In their study, learners showed decreased sensitivity in a self-paced reading task to gender agreement errors after compared to before a stay abroad. The authors argue that this may be due to learners resorting to more shallow input processing, focusing less on morphosyntactic markings and instead directing their attention to semantic-pragmatic cues. This is in line with the well-known shallow processing hypothesis by Clahsen and Felser (2018). Clahsen & Felser have suggested that a shallow processing mode is a characteristic feature of post-puberty L2[1] learners, and reflects impaired grammatical representations (Clahsen and Felser, 2018). Klassen, Ferreira and Schwieter suggest that this more superficial processing mode may be fostered by a study-abroad experience, because learners would experience pressure to achieve communicative goals, rather than to focus on morpho-syntactic markings, in such an environment. If the often-made assumption that superficial processing may hinder acquisition is correct, it is crucial to take both online processing measures and production measures into account to understand the acquisition process.

A number of recent studies on gender acquisition have indeed incorporated neuro- or psycholinguistic measures, with conflicting results. Hopp (2013) found that adult English-speaking L2 learners of German used phonological and morphophonological cues to predict and identify nouns, but other studies found that this was not always the case and that the processing paths of L2 learners differed significantly from the processing behavior of L1 learners (Guillelmon and Grosjean, 2001; Lew-Williams and Fernald, 2010; Grüter et al., 2012). The majority of studies show that learners react sensitively to gender violations (Grüter et al., 2012). However, it is sometimes not clear whether they react to violations in the assignment or in the agreement. In addition, advanced learners or beginners are usually compared to native speakers, heritage speakers or bilinguals, but the progression across different levels of proficiency is generally not traced (Montrul et al., 2008; Alarcón, 2009, 2011; Grüter et al., 2012; Diebowski, 2021), and the focus is rarely on directly relating performance in online processing during comprehension to performance in production. Finally, the majority of studies have

☆ We thank Cecilia Ugartemendía for help with constructing the stimuli, Mohiuddin Forhad for technical support, all student assistants helping with the creation of the MuLeCo corpus and with collecting eye-tracking data, as well as the schools and pupils involved in this study, who generously contributed their time and effort.

* Corresponding author.
E-mail addresses: wolf.johanna@lmu.de (J. Wolf), clara.terlaak@yale.edu (C. Terlaak), sarah.schimke@germanistik.uni-muenchen.de (S. Schimke).

[1] Note that as usual in the literature on L2 acquisition, "L2" refers to any language that is not acquired from birth, be it in instructed or uninstructed contexts, and also independent of the number of languages involved, e.g., "L2 acquisition" also refers to the acquisition of a third or fourth or further language.

https://doi.org/10.1016/j.amper.2025.100217
Received 31 December 2023; Received in revised form 6 January 2025; Accepted 26 January 2025
Available online 2 February 2025

dealt with adult learners (Montrul et al., 2008; Martoccio, 2022), so that we do not know whether results generalize to other populations, in particular, adolescent learners. In particular, the majority of studies on the L2 acquisition of gender in Spanish focuses on differences in proficiency, but does not put the results in relation to the age of the participants, who are generally referred to as adult learners. The terms late and adult learners are used in an almost interchangeable way, and it is usually not discussed whether late learners were adolescents at the onset of acquisition (see for a short discussion of that issue Au et al., 2002; Montrul et al., 2014).

It seems important to focus on processing patterns in adolescent learners specifically, however. Studying adolescents can shed light on the question whether the propensity to rely on shallow processing that has often been documented in adult learners can also be found in adolescent learners. This question is of theoretical interest because different studies assume widely different age thresholds at which input processing and acquisition mechanisms are assumed to change when compared to L1 acquisition (for an overview, see Unsworth, 2005). Moreover, adolescent learners are an important group from an applied perspective, because most adolescents receive some degree of foreign language instruction in school. A better understanding of how input processing and acquisition of grammatical categories are related in this age group could thus yield important information for language teaching in a large group of learners.

Against this background, the current study aims at looking at both online processing during comprehension and production data from a group of learners that has not been studied using online processing methods so far, namely, adolescent German-speaking learners of Spanish. As will be shown in the following, both German and Spanish have grammatical gender, which should facilitate acquisition in principle compared to English-speaking learners that many studies have focused on so far. As these adolescents are acquiring Spanish as a foreign language in school, with no or very limited natural exposure, the amount of input these learners have received is undoubtedly small compared to adult learners in immersion settings, however. It is thus an open question whether they will prove to be sensitive to agreement markings in the input. To assess potentially subtle differences in sensitivity, in our experiment, we will manipulate the salience of morphosyntactic markings in the input by comparing a condition where agreement is marked only on a determiner (DET + N) to a condition where agreement is marked both on the determiner and on an adjective (DET + N + ADJ). Using eyetracking-during reading, we ask whether there is evidence of online sensitivity to agreement violations on determiners and adjectives, and, crucially, whether any online sensitivity is related to successful production of gender as measured in a separate production task, that probes both gender assignment and agreement with relatively difficult lexical items. If we find a relation between the performance in these two tasks, this would support the assumption that attention to agreement markers, as operationalized by increased reading times of passages containing agreement errors, is beneficial for acquiring and correctly producing gender agreement.

In the following, we will first present the relevant properties of the Spanish and German gender systems, before we turn to a review of the empirical literature on processing strategies, as well as prior evidence from L1 and L2 acquisition.

1.1. Grammatical gender in Spanish and German

One of the most quoted sentences about gender seems to be the first sentence of Greville Corbett's study on gender in 1991 where he describes gender as "the most puzzling of the grammatical categories" and as a topic "that becomes more fascinating the more it is investigated" (Corbett, 1991, 1). Corbett appears to be part of a discursive tradition around the grammatical category of *gender* here: Meillet had noted in 1919 that gender is one of the least logical and unpredictable categories (Meillet [1919 (1982)]) and in 1959 Fodor wrote: "The category of grammatical gender is one of the still unsolved puzzles of linguistic science" (Fodor, 1959, 1). If one looks at the amount of studies[2] dedicated to gender since then in various fields of linguistics this still seems to be true.

The following section will provide an overview of gender as a grammatical category, focusing primarily on Spanish and German. Following Hocketts definition that "genders are classes of nouns reflected in the behavior of associated words" (Hockett, 1958, 231), gender fulfills two main tasks: assignment and agreement. As Schwarze (2000) points out, assignment can be defined as an operation by which nouns are classified in different groups according to certain lexical, inherent and invariable features. In this respect, gender has the function to separate nouns into distinct groups, e.g. masculine, feminine or neuter. Very often, there are no transparent cues to the gender class on the noun itself, such that the gender of a noun often can only be determined based on targets such as articles, adjectives or pronouns agreeing with this noun. Consider for instance noun phrases such as sp. *el/la joven* ("the adolescent") or sp. *el/la dentista* ("the dentist")[3] in Spanish, where gender can be determined only by looking at the article. Note that the reason for the particularity of both examples lies in their morphology: *joven* is a noun derived from an adjective and the suffix *-ista* regularly forms *nomina agentis*, which can be both masculine and feminine in modern Spanish; diachronically, however, derivatives with *-ista* were initially masculine.

Agreement can be defined as an operation when a gender carrying lexical item copies its gender feature onto other lexical items, e.g. adjectives, pronouns, participles etc. In languages with grammatical gender, gender assignment has consequences for the congruence of other syntactic elements. These consequences are far-reaching in German and Spanish, as illustrated in (1) and (2).

(1) La pequeña cabra negra ha llegado.
Die kleine schwarze Ziege ist gekommen.
The little black goat has come.
(2) El pequeño perro negro ha llegado.
Der kleine schwarze Hund ist gekommen.
The little black dog has come.

As these examples show, there are differences between the German and the Spanish gender system: Compared to English and Spanish, German has a more complex gender system, as agreement targets have to be congruent with the noun not only in gender and number, but also in case. Moreover, in German, the gender of the adjective is usually no longer additionally marked if the determiner already indicates the gender in an overt manner, but it has to be marked if the NP is a bare noun; Spanish, however, marks gender in all possible places regardless if the slot of the determiner in the DP is filled – as the examples in 3) show.

(3) der gute Wein – guter Wein
el vino bueno – vino bueno
the good wine – good wine

Finally, German has three noun classes (Masculine, Feminine and Neuter) whereas Spanish as a Romance language only has two noun classes (Masculine and Feminine). Tables 1 and 2 give a brief overview illustrating the two different gender systems, focusing on the differences in case (German) and agreement (Spanish vs German) and exemplifying assignment and agreement with the nouns *the dog* for the class of masculine nouns (ger. *der Hund*/sp. *el perro*), *the cow* for the class of

[2] For a detailed overview of work on acquisition of gender see Diebowski (2021), 1, for work on gender systems of romance languages Schwarze (2000), 40 and Schwarze (2009).

[3] This is the case for only a small number of Spanish nouns, see Diebowski, 2021, 15.

Table 1
Assignment and agreement in the German NP and AP.

assignment and agreement	Masc Sg	Masc Pl	Fem Sg	Fem Pl	Neuter Sg	Neuter Pl
Nominative	der Hund	die Hunde	die Kuh	die Kühe	das Huhn	die Hühner
Genitive	des Hundes	der Hunde	der Kuh	der Kühe	des Huhns	der Hühner
Dative	dem Hund	den Hunden	der Kuh	den Kühen	dem Huhn	den Hühnern
Accusative	den Hund	die Hunde	die Kuh	die Kühe	das Huhn	die Hühner
+Adjectiv Nom	der kleine Hund	die kleinen Hunde	die kleine Kuh	die kleinen Kühe	das kleine Huhn	die kleinen Hühner
+Adjectiv Gen	des kleinen Hundes	der kleinen Hunde	der kleinen Kuh	der kleinen Kühe	des kleinen Huhns	der kleinen Hühner
+Adjectiv Dat	dem kleinen Hund	den kleinen Hunden	der kleinen Kuh	den kleinen Kühen	dem kleinen Huhn	den kleinen Hühnern
+Adjectiv Acc	den kleinen Hund	die kleinen Hunde	die kleine Kuh	die kleinen Kühe	das kleine Huhn	die kleinen Hühner

Table 2
Assignment and agreement in the Spanish NP and AP.

assignment and agreement	Masc Sg	Masc Pl	Fem Sg	Fem Pl
	el perro	los perros	la vaca	las vacas
	el perro	los perros	la vaca	las vacas
	pequeño	pequeños	pequeña	pequeñas

feminine nouns (ger. *die Kuh/la vaca*) and *the hen* for the class of neuter nouns (ger. *das Huhn*):

In sum, while Spanish and German align in being gender languages in principle, the realization of gender in the system is very different in the two languages, both regarding the number of categories, the phonological and morphological transparency of markings, and the interaction with other categories, in particular case. We will turn to data from acquisition in the following.

1.2. Prior evidence from L1 acquisition

In children with L1 Spanish, a relatively early and mostly inconspicuous acquisition of the gender system can generally be observed. Psycholinguistic studies show a few phenomena that appear relevant to our research questions: acquisition is earlier and more stable with animate nouns than with inanimate nouns. In addition to this semantic phenomenon, the morphological nature of the nouns also seems to play a role: Gender acquisition occurs more easily, earlier and is more stable with nouns with overtly marked gender (type fem. *chic-a* "girl", masc. *chic-o* "boy") than with nouns with non-overtly marked gender (type fem. *orden* "order", masc. *guante* "glove"). Young children also seem to favor the masculine gender, in that they produce more masculine nouns overall and are more accurate with masculine than with feminine nouns (Alarcón, 2011, 2013). Overall, it can be stated that the studies on the acquisition of the Spanish gender system in L1 speakers are relatively unanimous and assume that it is acquired at an early age: Children acquire gender and gender agreement at around the age of 3–4 years (or even earlier; see Mariscal, 2009), and their knowledge seems to be very stable (Alarcón, 2011). Furthermore, studies show that Spanish L1 children between the ages of 4 and 11 rely predominantly on morpho-phonological and syntactic information when they have to decide on gender agreement in noun phrases with contradictory items (Lew-Williams and Fernald, 2007; Mariscal, 2009; Alarcón, 2011). There is disagreement regarding the modelling of linguistic knowledge in relation to grammatical categories: While Mariscal (2009), for example, adopts a constructivist approach, Alarcón takes a nativist approach (Alarcón, 2011). Mariscal bases her argument on the high variability in the phase before the gender system is consolidated and shows that children have to acquire several linguistic features and stimuli (meaning they use phonological, morphological, syntactical cues etc. in processing), which then interact with each other and lead to the stabilization of grammatical categories (Mariscal, 2009). However, it should also be mentioned here that processing studies on gender in toddlers are very rare. The first study, looking at children at a mean age of 37.7 months, is Lew-Williams and Fernald in 2007 where children were tested in an eye tracking experiment using a visual world design. However, they did not test opaque words, so that the morphological transparency of the items could not be assumed to be a facilitative factor. This has been investigated subsequently by Arias-Trejo and Alva (2012) in their extended replication study of Lew-Williams and Fernald (2007), in which also opaque items were included. It was found that toddlers aged around 24 months processed the -a/-o pattern, but then included key signals from the determiners and syntax. Arias-Trejo and Alva (2012) were also able to show that toddlers not only used morphological information about gender in articles, but also information contained in adjectives, which is an important information for the test design of the present study.

As for German, Eisenbeiss (2003) assumed that the acquisition of the German gender system proceeds in a similar way as has been shown by Mariscal (2009) for Spanish. Similar to Clahsen (1991), Eisenbeiss (2003), points to a correlation between the acquisition of gender assignment and gender agreement with the acquisition of DP in German and notes an interaction between the linguistic categories of gender, number and case. In her study, she also shows that children in the L1 acquisition of German initially seem to acquire a dichotomy of feminine gender versus non-feminine gender. This effect is also demonstrated for German by Ruberg and Müller, albeit for bilingual children. It is not clear whether the acquisition of neuter or masculine gender occurs later, as the studies are contradictory (Ruberg, 2013). However, neuter forms appear to be more prone to errors than masculine and feminine forms. Overall, similar to the acquisition of L1 Spanish, the acquisition of the gender system (assignment and agreement) for L1 German takes place relatively early and there are few assignment or agreement errors from around the age of 3–4 (Bewer, 2004; Ruberg, 2013, 2021).

However, it is important to take the distinction between lexical (assignment) and syntactic errors (agreement) into account. For instance, in a study by Blom et al. (2008) on the acquisition of Dutch, it was shown that incorrect gender markers are not by themselves an indication of a lack of acquisition of gender agreement, but often only an indication that the gender assignment to the noun has not yet been made correctly (Blom et al., 2008). This distinction is particularly important when analysing production data, as it is not always clear whether the problem is one of assignment or agreement. Blom et al. (2008) showed that although Dutch children exhibit a high rate of gender errors at the age of 3–7 years, gender congruence is consistent in itself, such that errors are exclusively due to incorrect assignment (Blom et al., 2008).[4]

This distinction between two possible ways of acquisition also has consequences for the question of the learnability of gender assignment and agreement and the modelling of the cognitive processes on which they are based - especially with regard to L2 acquisition. Before this is explained for the Spanish gender system, the different strategies assumed for the acquisition of gender assignment and agreement will therefore be examined in more detail below.

[4] The present study also takes this distinction into account in its structure, test design and analysis.

1.3. Acquisition strategies/processes across age

Binanzer (2020) describes two different acquisition tasks with regard to the gender system - with a focus on L1 German, but against the background of the work of Ayoun (2007, 2018), Franceschina (2001, 2003, 2005), Diebowski (2021) etc., her assumptions also appear to be transferable to the acquisition of Romance languages such as Spanish. On the one hand, children have to abstract the multiple phonological, morphological and semantic principles underlying the respective system from the input. On the other hand, a further acquisition task can be described as "the necessity of formally uniform labelling of the category genus in all genus-sensitive linguistic units that refer to the noun (genus congruence)" (Binanzer, 2020, 63).

In order to accomplish this dual task, children must therefore detect and process the linguistic categories in the input, correctly relate form and function to each other and store them in their mental lexicon. There is some controversy in the research literature as to which strategies are effective here. Basically, three basic positions can be identified that speakers use to determine gender and that are orientated towards the input.

a) *lexical route of gender assignment*: gender assignment is done without relying on morphophonological or phonological cues of the noun.
b) *reliable cue hypothesis*: gender assignment uses morphophonological, phonological or semantical patterns when these are disponible.
c) *dual-route access account/postlexical checking hypothesis*: Gender assignment initially takes place without using morphophonological, phonological and semantic information of nouns, but such information can be used at a later processing time if syntactic processing is disturbed or to check the plausibility of the assigned gender.[5]

Theoretically, L1 speakers seem to be able to access all three routes, but it is unclear which they actually use, when and how, and whether certain accesses are dominant in processing (Ruberg, 2013). With regard to L2 learners Weinrich relates gender to a so-called “blind category”, as the assignment is arbitrary, and must be learned as a grammatical fact together with the noun (Weinrich, 2005, 325–326). This refers to a discussion about the mental representation of adnominal gender information which is linked to the assumption of these three routes. This evokes the question whether gender is stored directly with the associated noun as a chunk or whether the assignment happens in each case by rule during production. The assignment would then take place via the semantic, morphological and phonological properties of the noun (Corbett, 1991). Psycholinguistic modelling usually argues for representation as stored information “as an inherent property of nouns in the knowledge of native speakers in his/her language” (Schriefers/Jescheniak, 1999, 577).

Studies with monolingual L1 Spanish children (34–42 months, mean age 37.7 months) indicate that they do use gender-sensitive cues of the article for predictions of the noun’s gender (Lew-Williams and Fernald, 2007). In an eyetracking study, they identified suitable referents more quickly if the article contained informative cues about gender (Lew-Williams and Fernald, 2007). However, the size of the vocabulary and greater grammatical complexity in the syntax also had a positive influence on the children’s speed. Overall, the children were slower than the adult comparison group, which the authors interpreted as a sign that the children were approaching adult speakers in their processing routes. The children therefore appear to rely relatively quickly on morphophonological and phonological patterns that they receive via the article when acquiring gender assignment (Lew-Williams and Fernald, 2007). These results also were reported in Arias-Trejo and Alva (2012); Arias-Trejo et al. (2013). Several studies have also shown that adult speakers react positively to information encoded in the determiner (Lew-Williams and Fernald, 2010). However, as alluded to in the introduction section, this does not always seem to apply to L2 learners, as Guillelmon and Grosjean (2001) found for L1 French speakers (adult) in comparison with advanced learners.

Ayoun (2007, 2018) found that native speakers of L1 French exhibited uncertainties in gender assignment and, in precisely these cases, attempted to obtain more morphophonological cues by, for example, enriching the noun with adjectives or transferring it to a more complex syntactic structure (Ayoun 2018; Scherag et al. (2004) presented similar results for German. However, this is contradicted by Bordag et al. (2006), who showed that adult L1 German speakers showed no response to phonological patterns in gender assignment. In their 2007 study on gender processing in L2 German learners, Bordag and Pechmann found no evidence that gender is permanently stored with the noun, but rather that it has to be processed and recalled anew each time. This may indicate a difference in the transparency of the Spanish and German gender system.

However, the studies seem to be relatively consistent in their findings that there are asymmetries between adult first- and second-language speakers with regard to the efficiency of processing grammatical gender cues. But, even if there were differences between native speakers and learners, the results of Frenck-Mestre et al. (2009), for example, indicate that L2 learners can use such cues and patterns in principle, but may need more transparency with regard to the informative cues. They examined the processing of grammatical gender in adult university students learning French as an L2. The results of Binanzer’s work point in a similar direction: She primarily investigates the acquisition sequences for gender in younger L2 learners (target language German) and demonstrates that learners must first acquire patterns so as then to be able to acquire agreement patterns, the so-called ‘consistent agreement pattern’ (Corbett, 1991), in order to find out which linguistic units are gender-sensitive targets (Corbett, 1991; Binanzer and Wecker, 2020) and how the gender information has to be mapped there (Binanzer and Wecker, 2020). Her findings certainly indicate that L2 learners may also be able to process probabilistically and filter the input for statistically probable patterns, even if they no longer seem to be as sensitive to this as children in first language acquisition. Nevertheless, implicit statistical learning seems to still be possible (Binanzer, 2020; Binanzer and Wecker, 2020). In consequence, she concludes that the most important step is to access the semantic congruent patterns – which could possibly be a Det + ADJ + N/Det + N + ADJ pattern.

Arnon and Ramscar (2009, 2012) already showed clear evidence that an accumulation of cues in the NP also leads to improved processing. They were able to show that the learning environment had an influence on the acquisition of gender in an artificial language: learners learnt more easily when Det + N sequences were shown than when only isolated nouns were presented in the input. This apparently strengthened the implicit learning process. From this it can be tentatively concluded that a further accumulation of gender-informative cues in the NP, e.g. by adding attributive adjectives, could direct the learner’s attention even better to the features to be acquired and trigger implicit learning even more strongly, since the syntactic processing route would then be added to the lexical processing route. This assumption seems to allow to also include the relationship of the relevant gender cues between article and adjective in the analysis and not only the relationships Det - N and Adj - N, as is usually the case.

In sum, studies agree that cues in the input, such as the comparatively regular and transparent -a and -o endings on Spanish agreement targets, are crucial for the acquisition of gender assignment and agreement. Moreover, it is plausible that input sequences that contain more of these cues - such as noun phrases with adjectives that are overtly marked for agreement - play a particularly important role compared to contexts with less salient markings. While it is uncontroversial that children manage to fully acquire the target system in their first language, many details about which types of L2-learners process agreement cues under which conditions, and how this relates to acquisition, are still open,

[5] For a detailed description of these routes see Ruter 2013, 59–73.

however, as will be summarized in the next section.

1.4. L2 acquisition of the Spanish gender system

As already mentioned, several studies indicate that the acquisition of the gender system in Spanish as a target language may represent a learnability problem (Alarcón, 2013; Grüter et al., 2012). Although the phenomenon is frequent in the input, it seems to stabilize relatively late in the learners' interlanguage, if at all (Alarcón, 2013, 205). Importantly, these findings refer to adult L2 learners of Spanish, often with English as an L1. It is unknown whether similar challenges appear for younger learners and for learners from languages with a gender system.

To cite Grüter et al., there seem to be primarily three reasons for the difficulties learners, even advanced ones, may have in acquiring the gender system of the target language:

> An L2 learner's difficulty with grammatical gender in a specific context of language use could thus stem from a number of different sources: (i) difficulty at the level of gender assignment (lexical knowledge); (ii) difficulty at the level of gender agreement (syntactic knowledge); (iii) or difficulty with accessing and/or deploying this lexical and/or syntactic knowledge within the real-time constraints imposed by the specific context of use. (Grüter et al., 2012, 194).

Several reasons have been put forward for these difficulties, most of which are linked to different approaches to modeling acquisition. For example, more nativist approaches primarily cite the age factor as a reason for the persistent problems that learners exhibit when acquiring gender systems. This is linked to the theory that learners have no or less access to the UG and therefore, for example, bootstrapping effects or knowledge transfer of the corresponding linguistic cues can no longer flow so strongly into the acquisition process (Franceschina, 2005). Most explanations for persistent problems with gender assignment ultimately end up in fossilization, arguing with the *Failed Functional Features Hypothesis* (FFH, Hawkins and Chan, 1997), "which predicts that the acquisition of L2 functional features not instantiated in the learner's L1 feature inventory will result in persistent divergence in adult learners" (Franceschina, 2003, 97) and which is based on the distinction between universal and parameterised features. Nevertheless, UG-based approaches do not fundamentally exclude acquisition, and they also assume that learners can reach a similar level to L1 speakers. However, the age factor is usually cited as a major limitation and the associated restriction of transfer possibilities, with the L1 generally being cited as the main source of transfer.

More input-based theories focus more on the nature and quality of the input and examine relevant characteristics such as frequency, salience, patterns and contextualisation and what influences these have on learners in L2 acquisition contexts (Wegener, 1995, 2016). In studies comparing L1, child L2 and L2 acquisition, it has often been found that changes between acquisition types are more gradual than may be assumed by native approaches assuming a critical period (see, for instance, Verhagen and Schimke, 2009). One factor contributing to such gradual differences could be a gradual shift from exclusively implicit processing patterns in children to a stronger reliance on explicit knowledge and explicit processing and learning strategies in older learners (compare Ellis, 2017).

Overall, it can be stated that at least the proficient L2 learners that have mostly been studied so far seem to be sensitive to gender-informative cues. In production data, they achieve almost as good a rate as natives (Grüter et al., 2012). Psycholinguistic studies also show that they react to gender violations: Both when the violation affects congruence in the DP and when the violation affects congruence between attributive adjective and noun (Keating, 2009, 2010). However, the immediate proximity (adjacent adjective) seems to play a role in adjective-noun combinations, as they react less to congruence violations that are further away from the trigger word (Keating, 2010).

Only very rarely the different acquisition models and different data analyses are brought together in order to investigate the acquisition of the gender system in an L2 – especially with a view to a possible better input structuring. One of the few exceptions – to our best knowledge – is the study by Grüter et al., 2012, which combines both elicited data (production) and data from an eyetracking study (comprehension) as well as offline and online data. However, this study only looks at advanced learners and has no comparison group with beginners. The present study therefore ties in with the combination of different types of data (production and perception, offline and online) on the one hand and also attempts to gain insights from looking at beginning learners at different levels of acquisition of the target language (2nd vs. 5th year learners) on the other.

2. The current study

The current study aims at relating performance from different tasks in the same learner population to each other. The targeted population are adolescent instructed learners of Spanish as a foreign language, with German as L1, coming from grades 9 (2 nd year of Spanish instruction) and 12 (5th year of Spanish instruction) in the German school system. Following our broad definition of L2 acquisition, we will refer to these learners as "L2 learners", but note that they have all acquired at least one other foreign language, English, before starting to learn Spanish, and that this group receives input in Spanish almost exclusively in the context of teaching, as is often the case in adolescent foreign language learning.

The study will consider both article and adjective agreement. While both types of agreement can be considered transparent and salient in Spanish when compared to other languages, the two targets arguably differ in the degree of transparency and salience when compared to each other. While articles are obligatory elements of Spanish NP's in almost all contexts, adjectives are optional and may also appear at a distance from the controlling noun. Moreover, while articles always inflect for gender, there are a number of invariable adjectives that do not inflect for gender, making adjective agreement less consistent and thus, arguably, less transparent when compared to article agreement. Prior research has found that L2 readers are less sensitive to agreement errors on targets at a distance from the controller (Foote, 2011; Keating, 2009), and that formal agreement marking may be acquired later on distant compared to more proximal targets at least in L2 German (Binanzer, 2020; Binanzer and Wecker, 2020). On the other hand, given the relative salience and systematicity of agreement markings on adjectives in Spanish, it is also plausible, from a perception perspective, that the presence of an inflecting adjective may help learners to notice patterns of agreement or non-agreement. To gauge the relative difficulty of determiner and adjective agreement in our learner population, we will first analyse an existing corpus of written texts of adolescent L1 German speaking learners of Spanish at a low level of proficiency, to find out whether there is evidence that gender acquisition has started at this low level, and whether there is a difference between articles and adjectives. Second, we will study the online processing and the controlled production of gender in a separate group of low proficient or intermediate learners. In sum, we asked the following overarching research questions.

1) Do beginning learners produce correct gender agreement markings in elicited narrative production data, and do rates of correct agreement differ between articles and adjectives in these data?
2) Are beginning and intermediate learners sensitive to agreement errors during reading, and does this differ depending on whether errors are marked on articles alone or on both articles and adjectives?
3) Is the individual performance during reading in beginning and intermediate learners related to the rate of correct gender assignment, as measured in a controlled production task?

Table 3
Percentage of demonstrative, definite and indefinite articles with correct agreement for all noun phrases occurring without an adjective referring to the head noun.

	NP
number correct	432
number incorrect	13
% article correct	**97.1**

Table 4
Percentage of adjectives with correct agreement for all noun phrases occurring with an adjective referring to the head noun.

	AP
number correct	121
number article incorrect, adjective correct	1
number article correct, adjective incorrect	7
number article and adjective incorrect	1
% adjective correct	**93.8**

2.1. Corpus analysis of elicited narrative productions

In a first step, we analysed a subset of the *MuLeCo* corpus (*Munich Learner Corpus*, de Crignis/Wolf; https://www.muleco.gwi.uni-muenchen.de/) with respect to gender agreement on articles and adjectives. The subset contained data from ten instructed German-speaking learners of Spanish at grade 9. The data of the learner production were collected in 2021 by two students as part of their bachelor's thesis in Romance Studies (corpus Nicolussi/Pilz). The participating schools in Austria were BRG Hallein in the city of Hallein (province of Salzburg) and Borromäum in the city of Salzburg. Both schools qualify for the Matura[6] and offer Spanish as a second (BRG Hallein) or third foreign language (Borromäum) with a teaching load of 3 h per week. All learners included in the presented analysis had L1 German. All had reached level A2 in Spanish. Data stem from written retellings of the *Frog Story* (Mayer, 1969). For each retelling, we extracted all noun phrases produced by the learner, and determined whether there was correct agreement between the article and the noun. Note that if agreement was incorrect, there is no way of knowing whether this is due to incorrect gender assignment to the noun, or incorrect agreement marking on the article or adjective. In the following, we will thus refrain from attempting to dissociate assignment from agreement.

In a next step, we determined whether the learner produced an attributive or predicative adjective referring to the head noun or not, and if so, whether agreement on the adjective was correct or not. The results are summarized in Table 3 exclusively for noun phrases where no adjective was used, and in Table 4 for noun phrases where the head noun had a relation not only with the article, but also with an adjective.[7]

These data suggest that overwhelmingly, learners produced correct agreement markings, with only 22 utterances out of a total of 575 utterances containing a gender agreement error at all. Descriptively, the percentage of errors was higher for adjectives than for articles, but as this generalization hinges on a small number of cases, it is not possible to draw definite conclusions as to whether adjectives and articles differ systematically from each other or not. Note, however, that the error rate of adjectives would be even higher if we had restricted the analysis to cases of inflecting adjectives, which was not the case (e.g., our analysis included invariable adjectives). Note also that elicited production data have other limitations inherent to this data type. In particular, whether nouns bore transparent cues to gender, whether the gender of the L1 and the L2 matched, and how many data points were contributed by individual learners was not controlled, and all of this could distort the picture and the comparison between markings on articles and on adjectives. Moreover, our aim was to relate production and perception data from the same individual learners to each other, which is not possible based on existing corpora. For these reasons, as a second step, we constructed a more controlled perceptive eyetracking-during-reading task and productive gender assignment and agreement task. Both of these tasks were completed by the same participants. We concentrated on a group of learners who were comparable to the learners of the subset of the *MuLeCo* corpus (e.g., learners from grade 9). In addition, we also tested more proficient learners (grade 12), to be able to explore how production and perception change at higher levels of proficiency.

2.2. Experimental tasks

24 German-speaking learners of Spanish participated in the experimental tasks. One learner had to be excluded due to excessive track loss in the eyetracking task, another learner had to be excluded due to reporting Portuguese as their native language. Thus, in the following, data from the remaining 22 learners will be reported.

Participants (19 female, 3 male) were German natives enrolled at a German secondary school. We recruited learners from two different grades, grade nine ($n = 9$, $M = 14.33$ years, $SD = 0.5$) and grade twelve ($n = 13$, $M = 17.31$ years, $SD = 0.48$). Due to the respective grade level and corresponding classroom Spanish lessons, the duration of formal Spanish instruction was very uniform among participants in each subgroup, $M = 1.67$ years for 9th graders and $M = 4$ years for 12th graders. In addition to Spanish, all participants indicated that they had studied other foreign languages including English, French and Latin or a combination of those, one participant had also studied Korean.

All learners first participated in the eyetracking task before completing the production task and finally questionnaires that probed their language learning history, their knowledge of the lexical items used in the experiment, and two debriefing questions that asked whether they noticed anything particular in the experimental items. After having completed all of these questionnaires, learners were informed about the purpose of the experiment. All materials, data and analysis scripts can be found in an openly accessible archive (https://osf.io/ey6jx/?view_only=fcb41b186cbd4037adc2285daa25ab4e).

2.2.1. Eyetracking-during-reading

For the eyetracking-during-reading task, we created 40 experimental passages, which each started with a context sentence describing a particular setting, such as 4).

4) La zona del lago era muy idílica.
'The lake area was very idyllic'

All context sentences contained seven words and ended with a full stop. The subsequent sentence (see 5) always started with a short word (either *allí*, 'there', or the interjection *oh*), followed by a comma, and then the critical subject noun phrase containing a definite article, a substantive and an adjective. For this subject phrase, we manipulated whether the adjective was invariable (5a and b) or inflected for gender (5c and d), and whether there were no agreement errors (5a and c), or either the article (5b) or the article and the adjective (5d) did not agree with the gender of the substantive. This created four experimental conditions, distributed across four experimental lists, such that each participant read each item in one of the four conditions only. Each

[6] The Matura exam is taken at the end of secondary school and gives students access to higher education.

[7] See also Schimke and Wolf (accepted) for an analysis of a larger part of the MuLeCo corpus that yielded robust converging evidence, with a low overall error rate in L2 Spanish, and a clear difference in rates of correctness on articles compared to adjectives. Note also that this more comprehensive study suggests that nouns with transparent markings of agreement are less prone to errors than nouns with less transparent markings, which strengthens the idea of the importance of input properties.

participant was presented with ten items per condition. The head nouns of the subject phrase were twenty masculine and twenty parallel feminine role names. All feminine forms contained the suffix -a (e.g., *amiga*, friend$_{FEM}$, or *escritora*, writer$_{FEM}$), while all masculine forms contained the suffix -o (e.g., *amigo*, friend$_{MASC}$) or -or (e.g., *escritor*, writer$_{MASC}$). We selected nouns varying in length from five to twelve characters and from three to four syllables. As for adjectives, we selected ten invariable adjectives for the invariable conditions and ten inflecting adjectives for the variable conditions. The adjectives selected ranged from four to twelve characters and from two to four syllables in length. The variability in length was taken into account by adding this factor into the statistical analysis, as described below. While we attempted to control for similar frequency for both nouns and adjectives, note that, as already mentioned by Keating (2009, 530), "frequency dictionaries proved unsuitable for selecting words for a study that included L2 learners, particularly beginning and intermediate learners, because they do not account for a classroom learner's exposure to the language." With this notion in mind, we selected the vocabulary in the critical noun phrases as well as in the post-critical region in accordance with the textbook used in the Spanish classes of the participants to ensure their knowledge of the vocabulary. Only very few nouns and adjectives selected did not occur in the textbook, however, these were highly semantically transparent and their meaning could easily be deduced from the students' native language or their other foreign languages, such as *constructor*, builder, or *arrogante*, arrogant. The subject phrase was always followed by a regular verb in the past tense, and the sentence concluded with three further words that could constitute the direct object or an adverbial phrase.

'The cute/sad writer was walking for 2 h over there.'

Finally, the passage ended with a concluding final sentence such as 6).

6) ¡La naturaleza es una inspiración!
'Nature is an inspiration!'

In addition to the experimental items, we created twenty filler passages, each containing three sentences of various syntactic structures. There were no agreement errors in the filler items.

2.2.1.1. Procedure. The experiment was implemented in SR Research Experiment Builder (SR Research, 2020) and displayed on a 17-in. Screen. While viewing was binocular, only the eye movements of the participants' dominant eye were recorded with an EyeLink Portable Duo eye tracker with a tracking rate of 1000 Hz, spatial accuracy was better than 0.5°. Participants sat approximately 60 cm away from the screen and placed their chins on a chin rest.

Prior to the actual experiment, participants were instructed, performed a 9-point-calibration and completed three practice trials. Each trial began with a manual drift correct item that served as a calibration check. Subsequently, the passages appeared in black letters against a light gray background using a monospaced font and a line spacing of 3.0. Participants were instructed to read the sentences silently in their normal reading pace and initiate the next trial by pressing the "Advance" button on a MilliKey MH-5 button box. Additionally, to ensure participants paid attention, in two of the practice trials, ten of the experimental trials and five of the filler trials, the passage was followed by a simple comprehension question asking about the content of one of the sentences not containing the critical noun phrase. Participants pressed YES or NO on the button box to answer. Participants were tested individually in

5a) agreement, adjective invariable

Allí	el escritor triste	caminó	durante dos horas.
There	the$_{MASC}$ writer$_{MASC}$ sad$_{INVARIABLE}$	walked	for two hours.

5b) no agreement, adjective invariable

Allí	la escritor triste	caminó	durante dos horas.
There	the$_{FEM}$ writer$_{MASC}$ sad$_{INVARIABLE}$	walked	for two hours.

5c) agreement, adjective variable

Allí	el escritor lindo	caminó	durante dos horas.
There	the$_{MASC}$ writer$_{MASC}$ cute$_{MASC}$	walked	for two hours.

5d) no agreement, adjective variable

Allí	la escritor linda	caminó	durante dos horas.
There	the$_{FEM}$ writer$_{MASC}$ cute$_{FEM}$	walked	for two hours.

'The cute/sad writer was walking for two hours over there.'

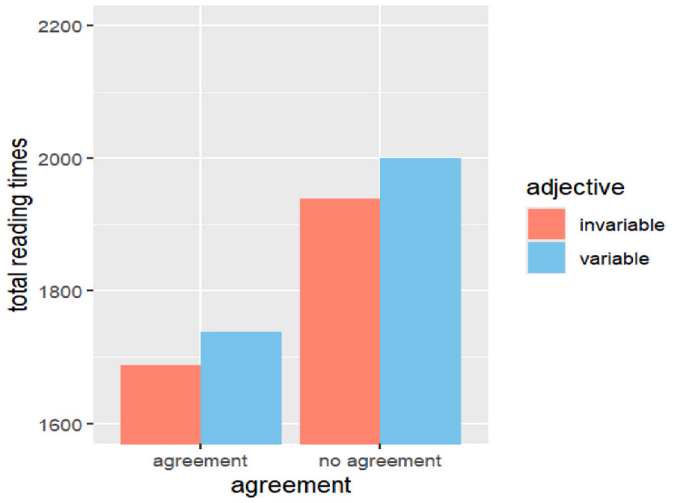

Fig. 1. Mean total reading times in milliseconds in the four experimental conditions for all participants.

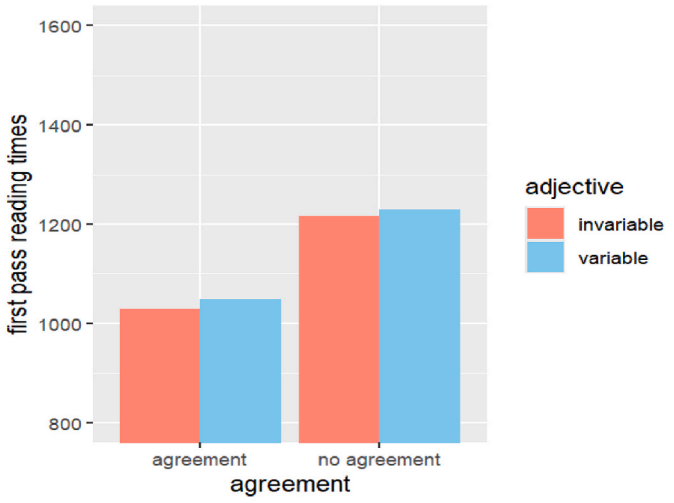

Fig. 2. Mean first pass reading times in milliseconds in the four experimental conditions for all participants.

Table 5
Output for linear mixed effects model for total reading times.

	Estimate	Std. Error	t value	Pr(<|t|)
(Intercept)	1863.588	79.902	21.025	<0.001***
Grade	113.320	78.603	1.442	0.165
Agreement	−123.860	28.988	−4.273	<0.001***
Adjective	−47.512	29.0231	−1.625	0.3104
Length	156.827	31.813	4.930	<0.001***
Grade: Agreement	43.534	28.942	1.504	0.133
Grade: Adjective	−14.732	28.941	−0.509	0.611
Agreement: Adjective	4.729	28.936	0.163	0.870
Grade: Agreement: Adjective	17.434	28.970	0.602	0.547

Note: Formula for the model: lmer(IA_DWELL_TIME ~ grade*agreement*adjective + length + (1 | Participant) + (1 | item)).

a quiet room at their school and each session including instructions, calibration and the reading task lasted about 30 min.

Before analysing the reading times, all fixations above 800 ms were removed, fixations of below 80 ms that were within one degree of visual arc of another fixation were merged, all other fixations below 80 ms were removed as well (compare Cunnings et al., 2015). 0.54% of trials had to be excluded from analysis due to extensive track loss, in 3.37% of trials fixations were manually adjusted to correct erroneous shifts in fixation locations.

We analysed two reading time measures to capture both total and initial processing effort. *Total reading time* is the sum of all fixations on a region over all passes; *first pass reading time* is the summed duration of all fixations on a region from first entering the region until exiting it to the left or right. Our region of interest consisted of the critical subject noun phrase, that is, an article, a noun and an adjective. We computed linear mixed-effects models, which allow for taking both fixed and random effects into account simultaneously, with Grade (9th or 12th grade), Agreement (agreeing or non-agreeing) and Adjective (invariable or inflecting) as well as their interactions as fixed factors and participants and items as random factors. Moreover, we included the length of the critical region in characters (Length) as a fixed factor. The factorial predictors Grade, Agreement and Adjective were sum-coded, and the variable Length was centred and scaled. All predictor variables were sum-coded.

2.2.1.2. Results. Fig. 1 shows the mean total reading times for all participants across all four experimental conditions.

Table 5 summarizes the output of the linear mixed-effects model that was fitted to these data.

There was a main effect of Length (number of characters), indicating that participants generally took longer to read longer compared to shorter passages. Importantly, there also was a main effect of Agreement, indicating that participants fixated noun phrases containing agreement errors significantly longer than those without agreement errors. Remarkably, no significant effect of the factor Adjective was found: Whether the noun phrase contained an invariable or inflecting adjective and thus, whether gender agreement was also marked on the adjective, did not influence reading time significantly.

Besides total reading times, we also analysed first pass reading times. Fig. 2 displays the means of this measure of all participants across the four experimental conditions.

We analysed these data using a linear mixed-effects model as well, Table 6 summarizes its output. Here, too, there was a significant main effect of Agreement, reflecting the increased first pass reading times for noun phrases containing agreement errors. Again, there was no effect of Adjective. The factor Grade showed marginal significance which can be attributed to the fact that the group of participants from 9th grade exhibited somewhat longer first pass reading times than 12th graders. No interaction effects were found. Finally, while the length of the area of interest did influence total reading times, it did not influence initial reading times.

In sum, both groups of learners read non-agreeing noun phrases more slowly than agreeing noun phrases in initial and total reading times. Moreover, our data suggest that whether an adjective was variable or invariable did not enhance effects of agreement. Before proceeding, let us consider two potential caveats in these data.

First, one may wonder whether we can conclude that 9th graders, as a group, have any sensitivity to agreement at all. Even in the absence of a significant interaction with grade, it is plausible that the effects of

Table 6
Output for linear mixed-effects model for first pass reading times.

	Estimate	Std. Error	t value	Pr(<|t|)
(Intercept)	1146.8497	53.2314	21.3007	<0.001***
G rade	95.7577	49.4701	1.936	0.0672.
Agreement	−101.4146	23.88 96	−4.245	<0.001***
Adjective	−11.5649	24 0.1729	−0.478	0.6325
Length	22.9053	31.3843	0.730	0.4688
Grade: Agreement	−23.8524	23.8007	−1.002	0.316 6
Grade: Adjective	−0.6786	23.7998	−0.029	0.9773
Agreement: Adjective	−4.9008	23.7896	−0.20 6	0.8368
G rade: Agreement: Adjective	−2 3.0544	23.8372	−0.976	0.3338

agreement may mainly have been carried by the more proficient, older learners, which makes it difficult to draw any conclusions from the analysis of the total group as to how early sensitivity to agreement starts. To explore this question, we conducted further subset analyses of the agreement effect for both groups separately. We thus ran the same linear mixed effects models as reported above (excluding the factor Grade) separately for subsets of the data that compromised only the 9th or the 12th graders, respectively. In total reading times, the effect of agreement was significant in the group of 12th graders (estimate: −164.01, SE = 35.73, t = −4.590, p < 0.001***), but only marginally significant in the group of 9th graders (estimate = −85.16, SE = 45,73, t = −1.862, p = 0.0635), while in first-pass reading times, there was a significant effect of the Agreement manipulation in both groups (9th graders: estimate = −301.29, SE = 121.56, t = −2.479, p = 0.0137*; 12th graders: estimate: −74.37, SE: 26.37, t: −2.819, p: 0.005**). In all, this suggests that sensitivity to agreement is not restricted to one of the groups, but that there may be subtle differences in reading patterns between proficiency levels. Second, one may wonder whether the absence of any effects of the adjective could be due to the fact that we analysed a relatively large region of interest, consisting of the article, the noun and the adjective. It is possible that small effects specific to the adjective region are masked in the analysis of this larger region. To explore this possibility, we conducted a further analysis of total reading times of the adjective region only. Again, we applied the same linear models as described above, but selected the initial and total reading times on only the adjective region as dependent variables. This analysis returned a marginal effect of Agreement for the group of all participants, but no effect of Adjective for the group of all participants and no effects of Agreement or Adjective for the subgroup of the 9th graders, but it did reveal a main effect of Agreement on the adjective region for 12th graders (estimate = −46.31, SE = 15 0.195, t-value = −3 0.051, p-value = 0.0026**). While it is suggestive that 12th graders seem to show effects on the adjectives that 9th graders do not show, crucially, there was, again, no effect of the factor Adjective and no interaction between Agreement and Adjective for this subgroup. This indicates that even if we consider the adjective region in isolation, the general pattern of the results does not change: learners tend to be sensitive to agreement errors, but it does not make a difference for their reading times whether these errors appear on articles only, or, in addition, on adjectives. Before we turn to interpreting this observation in the general discussion, let us consider our third research question, which deals with a potential relation between perception and production.

2.2.2. Relation between performances in perception and production

To link the perception and production data collected from the same learners, we aimed at establishing an individual single measure for each of the two tasks for each learner. For perception, we calculated an individual difference score by subtracting the average total reading times for grammatical items from the average total reading times for ungrammatical items (similar to Hopp et al., 2024). The greater this score is, the greater the advantage for grammatical items was for this specific learner. If the score is negative, this indicates that the learner read agreeing items on average more slowly than non-agreeing items, indicating no sensitivity to grammaticality.

Second, we designed a controlled production task to obtain a measure of learner's productive knowledge of gender assignment and agreement on an individual level. For this task, we chose twenty Spanish nouns that were not used in the eyetracking task. Ten nouns were masculine and ten feminine, and within each gender class, we chose nouns ranging from those with typical cues to gender (e.g. the feminine nouns *sorpresa - surprise* - and *cama - bed* and the masculine nouns *libro - book* and *helado - ice cream*) to nouns that were atypical in this respect (e.g. the feminine nouns *mano - hand* and *radio - radio* and the masculine nouns *día - day* and *problema - problem*). Half of the nouns of each gender class had the same gender as the German translation equivalent, while the other half differed in gender from the German translation equivalent. Finally, nouns covered a wide range of frequencies. The aim was to construct a task where gender assignment was presumably challenging at least for some items, so that the score would sensitively reflect different degrees of success in learners' prior acquisition of gender assignment and agreement. All nouns were embedded in a cloze task consisting of a short narrative, and learners were instructed to provide inflected article and adjective forms for each noun. The task thus yielded 20 data points for article agreement and 20 data points for adjective agreement per learner. Note that similar to the free production data, if learners produce errors in this task, it is impossible to tell whether these are errors of gender assignment or gender agreement. Nevertheless, we assume gender assignment to be the relatively greater challenge in this task, because on the one hand, the lexical items were deliberately chosen to pose a challenge with respect to gender assignment, and on the other hand, the task drew explicit attention to gender agreement, which should lower error rates for agreement, but not necessarily for assignment. As the main challenge thus presumably lies in assignment, we expect even less differences between the two agreement targets (articles and adjectives) in this task compared to a free production task such as the one used for the *MuLeCo* corpus.

Table 7 displays the total number of correctly and incorrectly agreeing article and adjective forms, as well as the proportion of correct forms, separately for learners from grade nine and grade twelve, as well as for all learners collapsed. Note that five data points are missing because one learner did not provide any answer for these five slots.

These data suggest that as expected, agreement errors occurred on both targets, with almost no difference between correctness on article and adjective agreement. As for variation between learners, we computed an average correctness score per learner (collapsing data from articles and adjectives), and found a wide range of correctness, ranging from 0.55 to 0.98. Moreover, correctness was strongly influenced by

Table 7
Absolute number of agreeing and non-agreeing as well as proportions of agreeing article and adjective forms in learners from both grades.

grade	target	correct	incorrect	proportion correct
9	article	132	46	0,74
	adjective	125	52	0,71
12	article	223	37	0,86
	adjective	224	36	0,86
total	article	355	83	0,81
	adjective	349	88	0,80

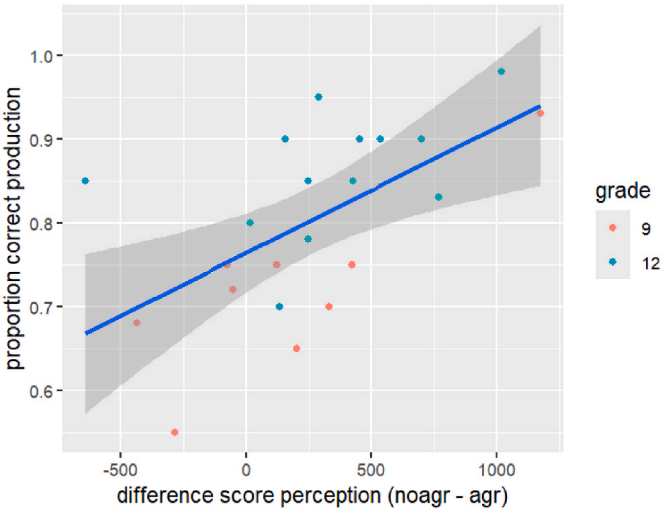

Fig. 3. Relation between difference score in total reading times and performance in the gender production task for learners from both grades.

grade, with only one learner from each grade scoring within the range of the scores of the learners from the other grade, respectively. To sum up, as expected, there was variability in the correctness, which plausibly is mostly due to variability in gender assignment, and individual scores were strongly influenced by the general proficiency level of the learners.

To address research question 3, we correlated the individual correctness score in production with each learner's difference score from the perception task. This correlation as well as data points for the individual learners are displayed in Fig. 5. By conducting this analysis, our aim was to investigate whether learners who show more sensitivity to agreement markings during reading produce a higher number of correct markings in the gender production task. Note that we are thus not comparing perception and production data based on comparable stimuli. Instead, the perception task was designed to probe whether there is any implicit online sensitivity to agreement. Nothing in the task pushes learners to pay attention to this grammatical phenomenon. Our question (in perception) was whether learners nevertheless take gender cues into account implicitly, and the stimuli were designed to make the detection of online sensitivity as easy as possible, by choosing frequent animate lexical items with transparent cues to agreement. For production, however, the task did push learners to pay attention to agreement markings, and even explicitly asked them to do so. The task was thus a deliberately difficult one. Our aim in combining these two different tasks was to probe whether learners who show sensitivity in an implicit perception task (with transparent gender markings) can be shown to be more successful in an explicit production task with deliberately difficult items. This assumption was supported by the statistical analysis which revealed a significant moderate positive correlation between the two variables, $r(20) = 0.59$, $p = 0.003665$**.

A closer look at the data suggests some noticeable patterns. First, most learners have a difference score between -100 and 700 ms, which may be taken as an indication that sensitivity to grammaticality was common on an individual level, and that thus, the group effects reported for the eyetracking analysis above are not due to a small subset of the learners only. Second, an obvious third variable that may explain the relation between the difference score and the gender production score is the grade of the learners, because 9th graders (red dots in Fig. 3) tended to have both a lower difference score and a lower gender production score than 12th graders (green dots in Fig. 3). To explore this potential confound, again, we looked at the correlation for each group separately, and found a robust effect in 9th graders ($r(7) = 0.80$, $p = 0.01027$*) and a moderate positive correlation for 12th graders which, however, did not reach significance ($r(11) = 0.42$, $p = 0.148$). This suggests that while it is plausible that grade is linked to proficiency, and that proficiency as measured in grade certainly contributes to the relation between the reading difference score and the gender agreement score that we observe, this does not fully explain the relation. We will come back to this relation and how it could be further explored in the general discussion.

3. General discussion

We asked whether 1) adolescent instructed learners of Spanish show knowledge of gender agreement both in elicited narrative production and in perception, 2) whether the amount and the stability of knowledge differs between markings on articles and on adjectives, and 3) whether there is a relation between individual performances in perception and production.

Turning first to elicited narrative production data, our results suggest above all that beginning (second year) German-speaking learners have considerable knowledge of gender assignment and agreement in L2 Spanish, as reflected in the overall high rates of correct markings. Descriptively, we found a small difference between agreement marking on articles compared to adjectives, with slightly higher error rates on the later. While this difference was too small to warrant firm conclusions, if confirmed in larger data sets, this pattern could be interpreted to mean that the acquisition of agreement starts with article-noun chunks, and that acquiring markings on more distant targets is an additional and more difficult step (see Schimke & Wolf, accepted, for a larger analysis that confirms this pattern). At the same time, the relatively high rate of correctness on adjectives may also be taken to support the idea that noun phrases that contain markings both on articles and on adjectives may constitute more salient patterns than noun phrases that contain markings only on articles, which is a comparison we looked at in the perception side of our study.

In perception, we found, however, that learners were sensitive to agreement errors that were caused by a mismatch between the article and the noun, but that whether the match or mismatch was additionally marked on an adjective or not did not modulate this pattern. Moreover, in a post-hoc analysis in which we looked at learners from the two grades separately, we found subtle differences between second-year (9th graders) and 5th year (12th graders) learners of Spanish, in that the difference between reading times for agreeing and non-agreeing noun phrases was more strongly reflected in initial reading times in the 9th graders, but was significant and descriptively stronger in total reading times for the 12th graders. Future studies with larger sample sizes should test whether such a difference in the strength of sensitivity in different measures between less and more proficient learners can be replicated and statistically confirmed. Based on the current study, what we can conclude is that there was an effect of agreement in the overall group, and that, importantly, the presence of an inflecting adjective did not prove to make any difference for gender processing, when compared to an adjective that was invariable. One way of interpreting this finding would be to take it as an indication of the importance of article-noun chunks: sensitivity to agreement marking on the article seems to be so great that the additional availability of the adjective form seems to make no further difference. As noted by an anonymous reviewer, it could also play a role that while the article precedes the noun, the adjective followed it in our stimuli, as is the case for most adjectives in Spanish. A potential effect of this word order could be addressed in future studies on different target languages. While it is interesting that there was no effect of the marking of the adjective in our study, this finding should be interpreted with caution, given limitations of our design that could be addressed in further studies. In particular, we did not directly compare the effect of agreement markings on articles and adjectives, in the sense that we did not directly pit these two sources of information against each other. Instead, the information on the adjective was always redundant, in that it provided a further cue to gender to the cue that was already provided by the article in the same stimulus sentence. To directly compare the two targets, future studies would need to dissociate them, for instance by combining agreeing article forms with non-agreeing adjective forms and vice versa. Furthermore, as our study did not include a control group with native speakers or with very proficient learners, we do not know whether the lack of sensitivity to the variability of the adjective reflects an acquisitional stage at which more distant agreement targets receive comparably less attention, or whether this pattern is common to all users of Spanish, also very proficient or native ones.

Finally, turning to the last research question, we found a relation between individual difference scores in reading and individual success of gender assignment in a controlled production task. Even though the production task targeted lexical items that were less frequent and less

transparently marked for gender than the items used in the reading task, learners who showed greater sensitivity during reading also tended to show better performance in production. This would be in line with the idea that being sensitive to agreement patterns during processing for comprehension facilitates entrenchment of knowledge of gender, due to the repeated perception and processing of relevant patterns, and this entrenchment of knowledge may in turn influence performance in production. Of course, the correlation we observed does not necessarily imply such a causal relation, but it does suggest that the two modalities are tightly linked on an individual level. This link could be further explored in future research, to which we turn next.

3.1. Perspectives for future research

Our results suggest several avenues for future research.

First, to deepen our understanding of the role of transparency and salience in the processing and learning of grammatical gender, it would be informative to look at other target languages. As outlined in the introduction, Spanish has a relatively transparent system of grammatical gender assignment and a highly transparent system of agreement that involves comparatively salient markings. Presumably, a high degree of transparency will be beneficial for both a morpho-phonological implicit and a more explicit, item-based strategy of input processing and acquisition. Moreover, it is plausible to assume that the less transparent a system is, the more difficult it will be to acquire in general (Vasić et al., 2012), and based on explicit strategies in particular. Finally, if a system involves markings that have little salience, this may make it particularly hard to rely on the morpho-phonological route in particular for older learners, who may not perceive the relevant markings due to, for instance, blocking effects of the L1 (compare Ellis, 2007) or generally less sensitive perceptual organs. In sum, less transparent systems and systems involving less salient marking may thus pose a greater challenge in particular to non-child learners. For Spanish, however, gender assignment and agreement may be acquirable with relative ease via different routes of acquisition, because there is an abundance of transparent markings in the input, both on articles and adjectives. To further explore this idea, we are currently preparing a parallel study on the processing of French as a foreign language. French and Spanish have a comparable typological distance to German, and are also comparable regarding the conditions under which they are taught in Germany (e.g., it is possible to compare learners of French and learners of Spanish, who learn these two languages under roughly comparable conditions). Crucially, however, the French gender assignment and agreement system is less transparent and involves comparatively less salient markings than the Spanish system. First analyses of learner production data show that L2 French learners are very prone to incorrect gender assignments and alignments even at very advanced levels. The error rate is around 10%, with the adjective phrase appearing to be somewhat more vulnerable than the DP (Wolf, 2024). Conducting a parallel study on French would thus provide another interesting perspective on the role of transparency and salience in foreign language grammar learning and teaching. Concerning the role of the first language, it would be beneficial to replicate the current study with L1 English learners of Spanish, to explore whether the relatively high sensitivity that we documented in this study is restricted to adolescent learners from a first language that is a gender language, or is present in adolescent learners in general.

Second, from a psycholinguistic perspective, a crucial question is how attention to morphosyntactic markings during processing for comprehension on the one hand, and successful acquisition and production of these markings on the other hand, are related to each other. It is plausible to assume that attention to markings is a prerequisite for the acquisition of the underlying system. However, being able to process markings for comprehension has often been found to be less demanding than producing them in spontaneous speech (see, e.g., the missing surface inflection hypothesis by Prévost and White, 2000). In addition, the precise relation between successful processing for comprehension and successful language production may be different in different types of language users, and for different linguistic phenomena (for an overview, see Schimke, 2022). To our knowledge, there is so far no longitudinal evidence on whether, and how fast, processing for comprehension translates into successful production for different learner types and acquisition targets. To gain a first insight into this relation, a given group of learners would need to be studied over larger time spans, with regular tests of their perceptive and productive performance. Given the correlational evidence of the current study that points to a relation between processing and production, it seems worthwhile to pursue such a longitudinal study with respect to gender markings in Spanish as an L2.

Third, a related question from a language teaching/didactics perspective would be whether acquisition and learning can be fostered by increasing the salience of crucial markings during teaching. Teaching materials often rely on teaching gender as a feature of each noun that has to be learnt by heart during vocabulary training, while teaching rules of gender agreement as a separate domain concerning knowledge of grammatical rules. Increasing the salience of agreement patterns in the input, for instance by always presenting nouns in context with articles, and often with adjectives, and possibly by visually highlighting the agreeing suffixes, may be a way of fostering knowledge both of gender assignment and gender agreement in an implicit way. This could be a strategy for addressing the problem of lexical representation that treat gender as an integrated lexical feature rather than isolating it from grammar. By presenting gender agreement as a holistic aspect of language use, rather than a disconnected rule, learners might develop a more robust understanding of how gender functions within the linguistic system. As a possible consequence language instruction should thus incorporate both production and comprehension exercises that emphasize real-time processing of grammatical gender. This approach might help learners develop automaticity in using gender cues during both speaking and listening. In addition, exposure to authentic language use, where gender agreement naturally occurs, can significantly reinforce learning, especially when learners encounter gendered structures across varied materials, such as reading texts and listening exercises. To further enhance this learning process, incorporating activities like reading aloud could be particularly promising. Reading aloud not only engages multiple modalities – visual, auditory, and motor – but also amplifies the phonetic patterns of gender markings, such as the recurring feminine phoneme *-a* in articles, nouns, and adjectives. This phonetic reinforcement may make the patterns more salient and memorable compared to the written mode alone, offering a compelling direction for both language instruction and future research on gender agreement acquisition.

Whether such input enhancement and integration of different modes of input presentation would lead to greater attention to gender markings during processing for comprehension, and whether this would translate into more successful acquisition and learning, could be tested in intervention studies in the classroom.

To conclude, the current study has found that sensitivity to gender markings during input processing is related to the acquisition of the category, as measured in accuracy in a production task, and that this sensitivity is high in adolescent L1 German learners in their L2 Spanish.

To disentangle the role of the adolescent age of our participants from specificities of the source and target language, future research could

conduct similar studies with other L1-L2 combinations. Moreover, we think that future classroom intervention studies with adolescent learners could be beneficial both to further deepen our understanding of the relation between input processing and language acquisition, and to compare the efficiency of different possible input-based intervention methods.

CRediT authorship contribution statement

Johanna Wolf: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Clara Terlaak:** Writing – original draft, Investigation, Formal analysis, Data curation, Conceptualization. **Sarah Schimke:** Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Alarcón, I., 2009. The processing of gender agreement in L1 and L2 Spanish: evidence from reaction time data. Hispania 92 (4), 814–828.

Alarcón, I., 2011. Spanish grammatical gender under complete and incomplete acquisition: early and late bilinguals' linguistic behavior within the noun phrase. Biling. Lang. Cognit. 14 (3), 332–350.

Alarcón, I., 2013. Grammatical gender in second language Spanish. In: Geeslin, K.L. (Ed.), The Handbook of Spanish Second Language Acquisition. John Wiley & Sons, Oxford, pp. 202–218.

Arias-Trejo, N., Alva, E.A., 2012. Early Spanish grammatical gender bootstrapping: learning nouns through adjectives. Dev. Psychol. 49 (7), 1–7.

Arias-Trejo, N., Falcón, A., Alva-Canto, E.A., 2013. The gender puzzle: toddlers' use of articles to access noun information. Psicologica: International Journal of Methodology and Experimental Psychology 34 (1), 1–23.

Arnon, I., Ramscar, M., 2009. Granularity and the acquisition of grammatical gender: how order-of-acquisition affects what gets learned. In: Taatgen, N.A., van Rijn, H. (Eds.), Proceedings of the 31st Annual Conference of the Cognitive Science Society. Cognitive Science Society, Austin, TX, pp. 2112–2117.

Arnon, I., Ramscar, M., 2012. Granularity and the acquisition of grammatical gender: how order-of-acquisition affects what gets learned. Cognition 122 (3), 292–305.

Au, T., Knightly, L., Jun, S., Oh, J., 2002. Overhearing a language during childhood. Psychol. Sci. 13, 238–243.

Ayoun, D., 2007. The second language acquisition of grammatical gender and agreement. French Applied Linguistics 16, 130–170.

Ayoun, D., 2018. Grammatical gender assignment in French: dispelling the native speaker myth. J. Fr. Lang. Stud. 28 (1), 113–148.

Bewer, F., 2004. Der Erwerb des Artikels als Genus-Anzeiger im deutschen Erstspracherwerb. ZAS Papers in Linguistics 33, 87–140.

Binanzer, A., 2017. Genus – Kongruenz und Klassifikation: Evidenzen aus dem Zweitspracherwerb des Deutschen. De Gruyter Mouton, Boston/Berlin.

Binanzer, A., 2020. Genuserwerb als referenzerwerb. In: Ahrenholz, B., Geist, B., Lütke, B. (Eds.), Deutsch als Zweitsprache: Perspektiven auf Schule und Hochschule, Erwerb und Didaktik. Beiträge aus dem Workshop „Deutsch als Zweitsprache, Migration und Mehrsprachigkeit. Fillibach-Verlag, Stuttgart, pp. 15–40, 2017.

Binanzer, A., Wecker, V., 2020. Lernergrammatik und Grammatikunterricht. Überlegungen zu natürlichen Erwerbssequenzen und didaktischer Progression. In: Langlotz, M. (Ed.), Grammatikdidaktik – theoretische und empirische Zugänge zu sprachlicher Heterogenität. Schneider Verlag Hohengehren, Baltmannsweiler, pp. 61–85.

Blom, E., Polišenská, D., Weerman, F., 2008. Articles, adjectives and age of onset: the acquisition of Dutch grammatical gender. Second Lang. Res. 24 (3), 297–331.

Bordag, D., Pechmann, T., 2007. Factors influencing L2 gender processing. Biling. Lang. Cognit. 10 (3), 299–314.

Bordag, D., Opitz, A., Pechmann, T., 2006. Gender processing in first and second languages: the role of noun termination. J. Exp. Psychol. Learn. Mem. Cognit. 32 (5), 1090–1101.

Carroll, S.E., 1999. Input and SLA: adults' sensitivity to different sorts of cues to French gender. Lang. Learn. 49 (1), 37–92.

Clahsen, H., 1991. Child Language and Developmental Dysphasia: Linguistic Studies of the Acquisition of German. Benjamin, Amsterdam.

Clahsen, H., Felser, C., 2018. Some notes on the shallow structure hypothesis. Stud. Sec. Lang. Acquis. 40 (3), 693–706.

Corbett, G.G., 1991. Gender. Cambridge University Press, Cambridge.

Cunnings, I., Patterson, C., Felser, C., 2015. Structural constraints on pronoun binding and coreference: evidence from eye movements during reading. Front. Psychol. 6.

Diebowski, J., 2021. Gender Acquisition in Spanish: Effects of Language and Age, 61. De Gruyter Mouton, Berlin.

Doleschal, U., 2004. Genus als grammatische und textlinguistische Kategorie. Eine kognitivfunktionalistische Untersuchung des Russischen. Lincom Europa.

Eisenbeiss, S., 2003. Merkmalsgesteuerter grammatikerwerb. Eine Untersuchung zum Erwerb der Struktur und Flexion von Nominalphrasen. Heinrich-Heine-Universität Düsseldorf Dissertation, Düsseldorf.

Ellis, N.C., 2007. Blocking and learned attention in Language Acquisition. In: CogSci 2007, Proceedings of the Twenty Ninth Cognitive Science Conference. Nashville, Tennessee, August 1-4, 2007.

Ellis, N.C., 2017. Implicit and explicit knowledge about language. In: Cenoz, J., Gorter, D. (Eds.), Language Awareness and Multilingualism. Encyclopedia of Language and Education. Springer International Publishing Switzerland, pp. 113–124. https://doi.org/10.1007/978-3-319-02325-0_7-1.

Fodor, I., 1959. The origin of grammatical gender. Lingua 8, 186–214.

Foote, R., 2011. Integrated knowledge of agreement in early and late English–Spanish bilinguals. Appl. Psycholinguist. 32 (1), 187–220. https://doi.org/10.1017/S0142716410000342.

Franceschina, F., 2001. Where lies the difference between native and non-native grammars? Evidence from the L2A of Spanish. In: Foster-Cohen, S., Nizegordece, A. (Eds.), EUROSLA Yearbook, 1. John Benjamins, Amsterdam, pp. 143–158.

Franceschina, F., 2003. Parameterized functional features and SLA. In: Liceras, J.M., et al. (Eds.), Proceedings of the 6th Generative Approaches to Second Language Acquisition. Somerville: Cascadilla Proceedings Project. Cascadilla Press, pp. 97–105.

Franceschina, F., 2005. Fossilized Second Language grammars. The Acquisition of Grammatical Gender. John Benjamins, Amsterdam/Philadelphia.

Frenck-Mestre, C., Foucart, A., Carrasco-Ortiz, H., Herschensohn, J., 2009. Processing of grammatical gender in French as a first and second language: evidence from ERPs. EUROSLA Yearb. 9 (1), 76–106.

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

Guillelmon, D., Grosjean, F., 2001. The gender marking effect in spoken word recognition: the case of bilinguals. Mem. Cognit. 29 (3), 503–511.

Hawkins, R., Chan, C., 1997. The partial availability of UG in SLA: the 'failed functional features hypothesis'. Second Lang. Res. 13 (3), 187–226.

Hockett, C.F., 1958. A Course in Modern Linguistics. Macmillan, New York.

Hopp, H., 2013. Grammatical gender in adult L2 acquisition: relations between lexical and syntactic variability. Second Lang. Res. 29 (1), 33–56.

Hopp, H., Schimke, S., Gastmann, F., Öwerdieck, D., Poarch, G., 2024. Processing to learn noncanonical word orders: exploring linguistic and cognitive predictors of reanalysis in early L2 sentence comprehension. Stud. Sec. Lang. Acquis. 46, 686–709. https://doi.org/10.1017/S0272263123000451.

Keating, G.D., 2009. Sensitivity to violations of gender agreement in native and nonnative Spanish: an eye-movement investigation. Lang. Learn. 59 (3), 503–535. https://doi.org/10.1111/j.1467-9922.2009.00516.x.

Keating, G.D., 2010. The effects of linear distance and working memory on the processing of gender agreement in Spanish. In: VanPatten, B., Jegerski, J. (Eds.), Research in Second Language Processing and Parsing. John Benjamins, Amsterdam, pp. 113–134.

Kiyko, S., Kiyko, Y., 2020. Genuszuweisungsstrategien im DaF-Unterricht. Glottotheory 11 (2), 169–207.

Klassen, G., Ferreira, A., Schwieter, J.W., 2023. The role of immersion learning in the acquisition and processing of L2 gender agreement. Appl. Linguist. Rev. 14 (2), 391–413.

Lew-Williams, C., Fernald, A., 2010. Real-time processing of gender-marked articles by native and non-native Spanish speakers. J. Mem. Lang. 63 (4), 447–464.

Lew-Williams, C., Fernald, A., 2007. Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. Psychol. Sci. 18 (3), 193–198.

Mariscal, S., 2009. Early acquisition of gender agreement in the Spanish noun phrase: starting small. J. Child Lang. 36 (1), 143–171.

Martoccio, A., 2022. Gender assignment strategies used by L1 and L2 speakers of Spanish. Int. Rev. Appl. Ling. 60 (2), 229–254.

Meillet, A., 1982. Linguistique historique et linguistique générale. Éditions Slatkine, Paris [1919].

Mayer, M., 1969. Frog, where Are You? ISBN: 978-0-8037-2881-3.

Menzel, B., 2004. Genuszuweisung im DaF-Erwerb. Psycholinguistische Prozesse und didaktische Implikationen. Weißensee, Berlin.

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. Lang. Learn. 58 (3), 503–553.

Montrul, S., Davidson, J., De La Fuente, I., Foote, R., 2014. Early language experience facilitates the processing of gender agreement in Spanish heritage speakers. Biling. Lang. Cognit. 17 (1), 118–138. https://doi.org/10.1017/S1366728913000114.

Ruberg, T., 2021. Gender agreement in SLI: a study on production and inflection of articles in German. Lang. Acquis. 28 (2), 131–152.

Ruberg, T., 2013. Der Genuserwerb ein- und mehrsprachiger Kinder. Dr Kovač, Hamburg.

Scherag, A., Demuth, L., Rösler, F., Neville, H.J., Röder, B., 2004. The effects of late acquisition of L2 and the consequences of immigration on L1 for semantic and morpho-syntactic language aspects. Cognition 93 (3), B97–B108.

Schimke, S., 2022. The psycholinguistics of child second language acquisition. In: Godfroid, A., Hopp, H. (Eds.), The Routledge Handbook of Second Language Acquisition and Psycholinguistics. Routledge, New York.

Schimke, S., Wolf, J., (accepted). Grammatical gender in adolescent learners of French and Spanish: Input properties and acquisition patterns. To appear in: Hartmann, S., Koch, N., Quick, A. (Eds.). Usage-Based Approaches to Multilingualism - Language

acquisition, language contact, language attrition. De Gruyter, Berlin/Boston (Applications of Cognitive Linguistics).
Schriefers, H., Jescheniak, J.-D., 1999. Representation and processing of grammatical gender in language production: a review. J. Psycholinguist. Res. 28 (6), 575–600.
Schwarze, B., 2000. Formale Durchsichtigkeit und semantische Nutzung der Genusklassifikation. Eine kontrastive Darstellung des Spanischen, Französischen und Deutschen. Vox Romanica 59, 40–76.
Schwarze, B., 2009. Genus im Sprachvergleich. Klassifikation und Kongruenz im Spanischen, Französischen und Deutschen. Narr, Tübingen.
Unsworth, S., 2005. Child L2, adult L2, child L1: differences and similarities. A Study on the Acquisition of Direct Object Scrambling in Dutch. Utrecht University. PhD thesis.
Vasić, N., Chondrogianni, V., Marinis, T., Blom, W.B.T., 2012. Processing of gender in Turkish-Dutch and Turkish-Greek child L2 learners. In: Biller, A.K., Chung, E.Y., Kimball, A.E. (Eds.), BUCLD36: Proceedings of the 36th annual Boston University Conference on Language Development. Cascadilla Press, pp. 646–659.
Verhagen, J., Schimke, S., 2009. Differences or fundamental differences? Kommentar zu J. Meisels Artikel ‚Second language acquisition in early childhood'. Z. Sprachwiss. 28 (1), 97–106.
Wegener, H., 2016. Regeln versus Muster: Der Erwerb flektierter Formen in DaZ: der Plural deutscher Substantive. In: Bittner, A., Spieß, C. (Eds.), Formen und Funktionen: Morphosemantik und grammatische Konstruktion. De Gruyter, Boston/Berlin, pp. 193–214.
Wegener, H., 1995. Das Genus im DaZ-Erwerb. Beobachtungen an Kindern aus Polen, Russland und der Türkei. In: Handwerker, B. (Ed.), Fremde Sprache Deutsch. Grammatische Beschreibung, Erwerbsverläufe, Lehrmethodik. Narr, Tübingen, pp. 1–24.
Wegera, K.-P., 1996. Das Genus. Ein Beitrag zur Didaktik des DaF-Unterrichts. Iudicium.
Weinrich, H., 2005. Textgrammatik der deutschen Sprache. Olms, Hildesheim.
Wolf, J., 2023. In chunks we trust…the problem of gender assignment in foreign language learning of French. In: Hennecke, I., Perevozchikova, T., Wiesinger, E. (Eds.), Multiword-Units in Multilingual Speakers. *Special Issue*. International Journal of Bilingualism. https://doi.org/10.1177/13670069241302278.