

Article

Gender Marking and Clitic Pronoun Resolution in Simultaneous Bilingual Children

Vasiliki Koukoulioti ^{1,2,*} , Stavroula Stavrakaki ³ , Maria Vomva ⁴ and Flavia Adani ¹ 

¹ Department of Education and Psychology, Freie Universität Berlin, Schwendenerstr. 33, 14195 Berlin, Germany

² Department of German Language and Literature, Faculty of Philosophy, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

³ Department of Italian Language and Literature, School of Philosophy, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

⁴ Department of Linguistics, University of Potsdam, Campus Golm, 14476 Potsdam, Germany

* Correspondence: vasiliki.koukoulioti@fu-berlin.de or vasilikikou@del.auth.gr

Abstract: The acquisition of clitics still remains a highly controversial issue in Greek acquisition literature despite the bulk of studies performed. Object clitics have been shown to be early acquired by monolingual children in terms of production rates, whereas only highly proficient bilingual children achieve target-like performance. Crucially, errors in gender marking are persistent for monolingual and bilingual children even when adult-like production rates are achieved. This study aims to readdress the acquisition of clitics in an innovative way, by entering the variable of gender in an experimental design targeting to assess production and processing by bilingual and monolingual children. Moreover, we examined the role of language proficiency (in terms of general verbal intelligence and syntactic production abilities). The groups had comparable performance in both tasks (in terms of correct responses and error distribution in production and reaction times in comprehension). However, verbal intelligence had an effect on the performance of the monolingual but not of the bilingual group in the production task, and bilingual children were overall slower in the comprehension task. Syntactic production abilities did not have any effect. We argue that gender marking affects clitic processing, and we discuss the implications of our findings for bilingual acquisition.

Keywords: bilingual acquisition; clitic pronouns; gender; language proficiency; language input; sentence production; sentence processing; Modern Greek



Citation: Koukoulioti, Vasiliki, Stavroula Stavrakaki, Maria Vomva, and Flavia Adani. 2022. Gender Marking and Clitic Pronoun Resolution in Simultaneous Bilingual Children. *Languages* 7: 250. <https://doi.org/10.3390/languages7040250>

Academic Editors: Melita Stavrou and Athina Sioupi

Received: 3 February 2022

Accepted: 9 September 2022

Published: 28 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Clitics, and in particular accusative or direct object clitics, have been an intriguing topic in Modern Greek (MG) linguistics, both in terms of syntactic theory and of processing/acquisition. From a theoretical point of view, their syntactic status is still debated (Mavroggiorgos 2010a; Revithiadou and Spyropoulos 2020). In terms of language development, the data on the acquisition of object clitics in typically developing Greek-speaking children provide clear evidence for an early acquisition in comparison to other languages (Tsakali and Wexler 2004; Varlokosta et al. 2016). However, data from atypical acquisition are inconsistent, with some studies suggesting typical-like acquisition (Manika et al. 2010; Stavrakaki 2001), others indicating a delay (Smith et al. 2008; Tsimpli and Stavrakaki 1999), whereas others indicate individual variability (Stavrakaki and van der Lely 2010). Importantly, complete acquisition has been judged mainly on the basis of a decrease in omission rates in obligatory contexts, with hardly any studies considering substitution errors. This might be misleading, as studies which analysed substitution patterns have attested the production of non-target object clitic forms, and specifically production of object clitics

with incorrect gender marking, both in typical and atypical development (Smith et al. 2008; Varlokosta et al. 2014).

With respect to bilingual acquisition of object clitics in MG, although research on the topic began over 15 years ago, the current picture remains fragmented for various reasons. In terms of methodology, all studies until now focus on omission rates, which might obscure more subtle aspects of the acquisition process, as suggested above for monolingual children. Moreover, some of the available data rely on spontaneous speech (Tsimpli and Mastropavlou 2008; Chondrogianni 2008), which might contain very few obligatory contexts for the production of clitics to begin with, and thus, acquisition patterns are hard to identify. Furthermore, only two studies have examined acquisition patterns of Greek clitics by simultaneous bilingual children, to the best of our knowledge (Andreou et al. 2015; Stavrakaki et al. 2011), whilst most claims about bilingual acquisition of objects clitics stem from data of children who acquire MG relatively late and in the context of formal education (Chondrogianni et al. 2015; Tsimpli and Mastropavlou 2008). Related to this, the children in these studies are older than 6 years old, which obscures earlier stages of the acquisition of MG clitics by bilingual children. Finally, the language pairs studied thus far with respect to the acquisition of the formal properties of object clitics is, to the best of our knowledge, first language (hereafter L1) Turkish and second language (hereafter L2) MG, and for simultaneous bilingual children, French/MG. Given the potential interference of the L1 in L2, the data provide an incomplete picture of possible bilingual acquisition patterns of MG object clitics in interaction with other languages.

This study aims to contribute to the study of the acquisition of object clitics by bilingual children in a novel way, going beyond their pattern of omissions in obligatory contexts. In particular, we capitalized on gender marking, an intrinsic property of clitics and personal pronouns in general, in order to shed light on fine-grained aspects of the acquisition of clitics, and in particular the processing of the reference to its antecedent, as we focus on the error types of the produced clitics and not only on the production/omission rates. Moreover, this study aims to investigate gender marking in a comprehensive way, addressing not only production but also comprehension. Finally, we address the impact of two non-linguistic factors, primarily proficiency and secondarily input. Specifically, we collected baseline measurements of verbal intelligence, as well as information on input and language dominance at the time of testing based on a parental questionnaire.

The paper is organized as follows. In Section 2, we provide an introduction to the linguistic analysis of object clitics and gender. Subsequently, we review previous research on the acquisition of object clitics and gender (Section 3). In Section 4, we present the methods of our study, in Section 5 our results, and in Section 6, we discuss our findings and their implication for bilingual acquisition and the development of language processing. In Section 7, we summarize our conclusions.

2. Theoretical Background

2.1. Clitics in MG

Personal pronouns are substitutes for nouns or noun phrases (NPs) in a sentence, and as they lack intrinsic descriptive content, they are considered as “functors” (Radford 1997). Greek clitics are considered to be the weak/short/non-emphatic forms of the personal pronouns. They are marked for first, second and third person singular and plural and they encode features of direct (accusative case) and indirect object (genitive case) and the possessor (genitive case) and gender (masculine, feminine, neuter) (Alexiadou and Anagnostopoulou 2000). See Table 1 for the full paradigm of clitics in MG:

Table 1. The paradigm of the clitics in MG.

Accusative	Genitive/Dative
me (me)	mu (to me)
se (you)	su (to you)
ton/tin/to (him/her/it)	tu/tis/tu (to him/her/it)
mas (us)	mas (us)
sas (you)	sas (you)
tus/ tis/ ta (them)	tus/ tus/ tus (to them)

Clitics either precede the finite verb (proclitic) or follow non-finite forms and imperatives (enclitic). In this study, we focus on the proclitic accusative object clitics in third person singular (see (1)).

- (1) Tin vlepo
her see_{1stSing}
“I see her”

There is a lively debate in the literature regarding the syntactic status of clitics, whether they are base generated or moved to their surface position and how they relate to the thematic position to which they refer to. More specifically, concerning proclitic accusative clitic pronouns in MG, most accounts agree that they land to their surface position by movement¹, although there are subtle differences due to different theoretical approaches. For instance, it has been argued that clitics adjoin to the left of the inflectional head to which the verb raises (Revithiadou and Spyropoulos 2020; Rivero 1994; Terzi 1996). Alternatively, following Cardinaletti and Starke (1994) who proposed that clitics are deficient (underlying) phrases which have a head status in the surface structure, Tsimpli and Stavrakaki (1999) and Tsimpli (2001) propose that (third) person clitics are clusters of uninterpretable nominal features that appear on the *light v* head of the verb. Mavrogiorgos (2010a) uses the term hybrids to capture the fact that they have properties of both XP/X categories. Their hybrid character is based on the fact that, in order to be assigned case and check their phi features, they have to move to their verbal host, moving first as XPs and end up as Xs (heads). In addition, in the spirit of Sportiche (1996), who assumes that the clitics are base generated in the head of the Clitic Phrase (CIP), where the uninterpretable features of the coreferential object *pro* are checked through the specifier-head agreement, Tsakali and Wexler (2004) propose that in languages with participial agreement (e.g., French, Italian, Catalan) the *pro* also checks the uninterpretable case feature in Agr-O (object agreement) before moving to the specifier of CIP. There is no consensus in the literature as to which features are checked in each position². What is important for the acquisition of clitics is the assumption that in languages with participial agreement, clitics have to check features in two functional projections, whereas in other languages, only in one, in the CIP. This assumption is crucial in order to interpret cross-linguistic differences in the acquisition of clitics (see Section 3 below).

2.2. Gender

As mentioned above, clitics are marked for gender in addition to case and number. NPs in MG are also marked for gender, a grammatical feature that pertains both to the lexicon and the syntax. In particular, gender, if present in one language, has the following properties (Alexiadou et al. 2008; Ralli 2002): (1) it is arbitrary and not predictable from the noun meaning except for human nouns, (2) it is an intrinsic lexical feature and, as such, it has to be learned as a component of the lexical representation, and (3) it is part of the noun stem and not the inflectional affix. Thus, gender differs from features such as number, which are chosen in the frames of an utterance. For some languages, in which gender can be separated from number, such as in Spanish, a syntactic Gender Phrase projecting immediately above the noun phrase has been postulated. On the contrary, in fusional languages such as MG, in which gender marking cannot be teased apart from case

and number in nominal declension, the existence of a separate gender phrase cannot be justified (Alexiadou et al. 2008). Apart from this, although the morphophonological cues of the nouns are consistent to a great extent, there is not a one-to-one correspondence of phonological cues and gender in MG (Alexiadou et al. 2008; Ralli 2002) among others. For instance, the ending *-oç* may be the termination of a masculine, feminine or neuter noun. Anastasiadi-Symeonidi and Cheila-Markopoulou (2003) suggest that gender is predicted on the basis of the notion of prototypicality, which is defined on the basis of noun meaning and suffix (cited from Varlokosta and Nerantzini 2013). Although gender is a lexical property, it is similar to the morphosyntactic categories number and case, as it triggers agreement with other categories, such as adjectives and determiners (Tsimpli 2014).

Concerning gender assignment in clitics, Mavrogiorgos (2010b) suggests that third person direct object clitics are merged in the complement position and have gender features as their basis. Gender combines with number, which has its own projection, and with the feature [def]. Further, he assumes that *V* attracts the clitic to its specifier position, whereas an optional feature at *v* linked to familiarity and old information forms a chain with the clitic. According to Tsakali and Anagnostopoulou (2008), the gender feature of the clitics is checked by means of the specifier-head agreement in the CIP or in the Agr-O phrase, depending on the typological characteristics of a language. In any case, gender is crucial in order to establish a link between the clitic and its antecedent.

3. Psycholinguistic Background

3.1. Acquisition of Clitics

The acquisition of (third person) pronouns is a complex process, which involves the development in two domains, pragmatics/discourse and morphosyntax. The former is necessary in order to pick an entity of the discourse as a referent, and the latter is responsible for the morphological marking and the appropriate placement of the pronominal, in order for it to receive the intended reference (Varlokosta et al. 2016). For this reason, there are several accounts for the acquisition of clitics, which can be distinguished between referential and syntactic-computational (Tsakali 2014). According to the former, on the one hand, omission of clitics is due to the pragmatic immaturity of young children. In particular, Schaeffer (2000) assumes that children cannot always understand that the knowledge of the speaker and of the hearer are distinct entities (*Concept of Non-Shared Knowledge*). Thus, they fail to mark consistently referentiality as the adults. In such cases, the object clitics are omitted (cf. Tedeschi 2008).

The syntactic-computational accounts, on the other hand, attribute clitic omissions to the restrictions of the child computational system in the domain of syntax and, in particular the *Unique Checking Constraint* (Wexler 1998), whose core assumption is that the D-features of a determiner phrase can only check against one functional category in child grammar due to computational restrictions. The consequence of the core assumptions of this constraint, beyond the specific theoretical refinements (Tsakali and Anagnostopoulou 2008), is that in languages in which clitics check their features only in the CIP, clitic pronouns are acquired very early, such as in MG and Spanish (Gavarró et al. 2010; Tsakali and Wexler 2004). On the contrary, in languages in which clitics have to check their features in two phrases, clitic omission is observed, e.g., in Catalan (Gavarró et al. 2010).

Table 2 provides an overview of all developmental papers that are published thus far, whose focus is on the acquisition of clitic pronouns in MG, ranging from monolingual (L1) to bilingual (L2), including typically developing (TD) as well as children with Specific Language Impairment (SLI). In the following, we briefly present the most relevant aspects of this research. First, L1 TD acquire clitics in MG as early as 2 years old in terms of production rates in obligatory contexts (Marinis 2000; Tsakali and Wexler 2004; Varlokosta et al. 2016). Moreover, comprehension of clitics has also been found to be target-like already at three to four years of age (Stavrakaki and van der Lely 2010; Varlokosta 2002).

Table 2. Studies on the monolingual or bilingual acquisition of clitics of typically developing children and children with Specific Language Impairment in MG. TD: typically developing, SLI: Specific Language Impairment, L1: monolingual, L2: bilingual, MG: modern Greek, RTs: reaction times, CA: chronological age. The age range is in brackets if the mean is provided.

Study	Lang. Profile	Participant Number	Language (Pair) Tested	Mean CA or Age Range	Method	Findings
Studies with L1 TD						
Marinis (2000)	L1 TD	Case study (Christofidou Corpus)	MG	1;7–2;8	Spontaneous Speech	adult-like performance from 2 years onwards
Tsakali and Wexler (2004)	L1 TD	5 Children (Stephany Corpus) Experimental Study (Group 1: 15, Group 2: 10)	MG	- Group 1: 2;4–3;0 Group 2: 3;0–3;6	Spontaneous Speech Elicitation task	adult-like performance from 2 years onwards 99.2% correct production (overall in both groups)
Varlokosta et al. (2016)	L1 TD	20	MG	5;0–5;11	Elicitation task	98.4% correct production
Studies with L1 SLI						
Manika et al. (2010)	L1 SLI L1 TD	L1 SLI: 19 L1 TD: 32	MG	SLI: 6;2 (4;10–8;1) TD: 3;10 (3;1–6;0) (vocabulary matched)	Elicitation task	SLI: 95% correct TD: 96% correct
Mastropavlou (2006)	L1 SLI L1 TD 1 L1 TD 2	10 L1 SLI 10 L1 TD 1 10 L1 TD 2	MG	SLI: 4;2–5;9 TD-language matched: 3;0–3;7 TD-age matched: 4;2–6;0	Elicitation task	SLI: 60.1% (22.1% of all responses gender errors) TD language-matched: 83.3% (29.6% of all responses gender errors) TD age-matched: 93.5% (10.3% of all responses gender errors)
Smith et al. (2008)	L1 SLI L1 TD	9 L1 SLI 9 L1 TD language matched 9 L1 TD age matched	MG	SLI: 4;9–6;9 TD language matched: 2;10–4;3 TD age matched: 4;11–5;11	Elicitation task	SLI: 64% (28.8% of all responses gender errors) TD language matched: 94% (18.8% of all responses gender errors) TD age matched: 96%

Table 2. Cont.

Study	Lang. Profile	Participant Number	Language (Pair) Tested	Mean CA or Age Range	Method	Findings
Studies with L1 SLI						
Stavrakaki and van der Lely (2010)	L1 SLI L1 TD	L1 SLI: 9 L1 TD 1: 17 L1 TD 2: 18 L1 TD 3: 12	MG	SLI: 10;6 (7;7–13;5) TD 1: 4;5 (3;11–5;3) TD 2: 5;1 (4;2–6;2) TD 3: 6;2 (4;7–8;3)	Elicitation task Comprehension task	Elicitation task SLI: 65.2% TD 1: 97% TD 2: 97,2% TD 3: 98% Comprehension task SLI: 55% TD 1: 75% TD 2: 87% TD 3: 91.6%
Tsimpli (2001)	L1 SLI L1 TD	L1 SLI: 7 L1 TD: 4 (Stephany Corpus)	MG	3;5–7;00	Spontaneous speech	SLI: 3.8% correct performance TD: adult-like performance from 2 years onwards
Tsimpli and Stavrakaki (1999)	L1 SLI	case study	MG	5;5	Spontaneous speech	3.49% correct production
Varlokosta (2002)	L1 SLI L1 TD	20 TD 4 SLI	MG	SLI: 4;7–8;1 TD: 4;6 (3;6–5;10)	Truth Value Judgment task	SLI: individual variability TD: 88–95% in all constructions except for secondary predicate construction
Varlokosta et al. (2014)	L1 SLI L1 TD	5 L1 SLI 55 L1 TD	MG	SLI: 6;3, (5;11–6;8) TD: 4;7, (3;6–5;11)	2 Elicitation tasks	SLI: Task 1: 73.3% (20% of them gender errors) Task 2: 85% correct (16.7% of them gender errors) TD: Task 1: 92.6% (16.7% of them gender errors) Task 2: 82.6% correct (15% of them gender errors)

Table 2. Cont.

Study	Lang. Profile	Participant Number	Language (Pair) Tested	Mean CA or Age Range	Method	Findings
Studies with L1 SLI and L2 TD						
Chondrogianni et al. (2015)	L1 TD L2 TD L1 SLI	20 L2 TD 31 L1 TD (MG) 16 L1 SLI (MG)	Turkish/MG	L2:7;6 (5;9–8;10) L1 TD: 7;3 (6;0–8;6) SLI L1: 6;8 (5;6–8;4)	Elicitation task Self-paced listening task	Elicitation task: L2 produced fewer clitics than L1 TD and L1 SLI, L2 more omissions than substitutions (with an NP) Self-paced listening task: Critical segment: L1 TD and L2 had longer RTs in ungrammatical sentences, SLI L1 no difference between grammatical and ungrammatical sentences, Post-critical segment: L2 children had longer RT than the L1 children
Tsimpli and Mastropavlou (2008)	L1 SLI L2 TD	SLI: 6 L2 TD 1: 5 L2 TD 2: 5	Turkish/MG	SLI: 4;0–6;2 TD 1: 8;0–9;0 TD 2: 11;0–12;0	Spontaneous speech	SLI: 32–96% correct TD 1: 27% correct TD 2: 56% correct
Studies with L2 TD						
Andreou et al. (2015)	L1 TD L2 TD	38 L2 TD Residents in Greece 39 L2 TD Resident in Germany 20 L1 TD MG	MG/German	Age range of both groups 8;0–12;0,	Spontaneous speech	L1 children used significantly more clitics than the two L2 groups for character maintenance Greek vocabulary and early literacy input predicted production of clitics
Chondrogianni (2007)	L1 TD L2 TD	50 L1 TD MG 66 L2 TD distributed across different proficiency classes (according to a language test)	Turkish/MG	L1: 7;0–12;0 L2: 7;0–12;0 distributed across different proficiency classes	Truth value judgement combined with elicited production task	L1: ceiling performance L2: difference in production rates between the intermediate and the advanced group, gender errors the most prominent error (varying percentages depending on the proficiency level)

Table 2. Cont.

Study	Lang. Profile	Participant Number	Language (Pair) Tested	Mean CA or Age Range	Method	Findings
Studies with L2 TD						
Chondrogianni (2008)	L1 TD L2 TD	18 L1 TD Language matched 50 L1 TD Age-matched 79 L2 TD distributed across different proficiency classes (according to verbal density, lexical diversity and ration of error free utterances)	Turkish/MG	L2: 7;0–12;0 L1 Language matched: 2;8–5;6 L1 Age matched: 7;0–12;0	Spontaneous speech Truth value judgement combined with elicited production task	Spontaneous speech: L2 children fewer clitics than L1 language-matched and age-matched, only the high proficiency had ceiling performance Elicited production task: Ceiling performance for the high proficiency level group L2 significantly fewer clitics than L1 in all levels except for the high proficiency group No task effect
Studies with L2 SLI						
Stavarakaki et al. (2011)	L2 SLI L2 TD	1 SLI 2 TD	French/MG	SLI: 9;0 TD: 4;7–5;11	Elicitation task	SLI: 100% MG, 50% French correct TD: 100% MG, 87.5% French correct

Clitics are expected to be more difficult to acquire by bilingual children, especially because of the syntax–pragmatics interface they involve (Sorace 2004). However, one should keep in mind that bilingual populations can be extremely heterogeneous, ranging from children that are exposed to the second (or third) language(s) from birth (see De Houwer 2009 for an inclusive definition of simultaneous bilingual acquisition, but Yip 2013 for a thorough discussion of the difficulties of formulating an accurate definition) to the ones that are exposed to the second (or more) language(s) later during development. Defining successive or sequential bilingualism is even more complex, as it can ensue at different ages (childhood or adulthood) and under very heterogeneous conditions (natural context or language instruction or a combination of the two) (Li 2013). In this study, we focus on bilingual acquisition during childhood, which takes place either simultaneously when the onset of acquisition is within the critical period, or successively, when the onset falls outside the critical period. This has to be taken into account when comparing data from different studies.

Most of the studies on the acquisition of MG clitics address successively bilingual children, who were exposed to MG in the context of language instruction. A unanimous finding for this population is the high omission rates of clitics in obligatory context. Tsimpli and Mastropavlou (2008) interpret this finding in terms of the interpretability hypothesis (Tsimpli 2003), according to which the acquisition of uninterpretable features is subject to the critical period. Therefore, uninterpretable features of the L2 which are not present in L1 are harder to acquire and/or are not acquired by L2 learners in the same way as the interpretable features.

Studies which focus on the effects of non-linguistic factors, such as proficiency of the speakers, challenge the interpretability hypothesis. The findings in Chondrogianni (2007) and Chondrogianni (2008) suggest a strong proficiency effect on production rates in both spontaneous speech and elicitation tasks. Crucially, the effect was found regardless of the proficiency measurement, which was different in the two studies (language test vs. spontaneous speech indexes).

Another non-linguistic factor, language input, has been addressed by Andreou et al. (2015). They report that when the usage of a clitic is required for character maintenance in storytelling, monolingual children used clitics more frequently than simultaneous bilingual ones. Crucially, the usage of clitics is predicted not only by vocabulary in MG but also early literacy input (i.e., before schooling). Although this study focused on the felicitous usage of clitics in terms of referential strategies and not on their morphosyntactic properties, it highlights that production of clitics can be challenging even for simultaneous bilingual children, at least from a pragmatic point of view. Moreover, it suggests that not only proficiency at the time of testing but also input characteristics and input history play a role in the usage of clitics.

Despite the dissimilarity between L1 and L2 TD children in production both in terms of morphosyntax and pragmatics, there is strong evidence for similarities in processing. Chondrogianni et al. (2015) replicated previous findings in production, namely higher omission rates for L2 TD children than their L1 age-matched peers in the sentence elicitation task. The self-paced listening task they used for testing processing included grammatical and ungrammatical sentences, the latter missing a clitic in an obligatory context. The L1 and L2 group presented a similar pattern, as both groups had longer reaction times (hereafter RT) for ungrammatical sentences at the critical segment, i.e., where the clitic is expected. However, the two groups differed in that L2 children presented longer RT in the post-critical segment than the L1 children. All in all, the findings indicate that high omission rates in L2 production do not imply incomplete acquisition of clitics across modalities.

This insight is corroborated by the comparison of L2 TD children with children with SLI. Concerning the latter, the findings are quite equivocal. A few studies have provided evidence for intact production of clitics in MG at least from the age of 4;10 for L1 but also for L2 children with SLI (Manika et al. 2010; Stavrakaki et al. 2011; Varlokosta et al. 2014), whilst several studies provide evidence for high omission rates in obligatory contexts (Mas-

tropavlou 2006; Smith et al. 2008; Stavrakaki and van der Lely 2010; Tsimpli and Stavrakaki 1999). Studies which compared L2 TD and SLI children indicate divergent patterns. Tsimpli and Mastropavlou (2008) and Chondrogianni et al. (2015) found that (at least some) L1 children with SLI produce more clitics than L2 TD children. Conversely, Chondrogianni et al. (2015) report that L2 TD were sensitive to (un)grammaticality, in contrast to the SLI children. Therefore, processing of clitics seems to be better than production in L2 TD acquisition.

A remarkable finding which applies to L1 TD and L1 SLI but also to L2 TD children is the higher frequency of gender errors in comparison to case and number errors even when production rates are high (Mastropavlou 2006; Smith et al. 2008; Varlokosta et al. 2014 for L1 TD and Chondrogianni 2007 for L2 TD children). Therefore, the complete acquisition of gender marking on object clitics extends far longer than the achievement of adult-like production rates in obligatory contexts. Further evidence that processing of gender during clitic pronoun resolution is challenging as well as that proficiency influences gender processing comes from studies of event related potentials (ERP). In particular, Rossi et al. (2014) have investigated the sensitivity of adult English-speaking learners of Spanish to number and gender violations in clitic processing. They found that the ERP patterns for number violations were similar for the L1 Spanish-speaking group and the learners of Spanish, but only highly proficient learners of Spanish manifested the same ERP pattern with native speakers to violations of gender marking.

In sum, production of clitics in MG is challenging in simultaneous and successive bilingual acquisition. Previous research has focused on omission rates in obligatory contexts, but erroneous productions of clitics have been largely ignored, although gender errors are very frequently attested, and studies from other languages suggest that gender processing during clitic resolution is challenging and mastered at a native-like level only by speakers with high proficiency. Finally, data from comprehension/processing suggest that the difficulty with clitics is modality specific, as the processing patterns of successive bilinguals are similar to monolingual TD children.

3.2. Acquisition of Gender

The age of acquisition of gender varies in different languages depending on the availability of transparent morphophonological cues. Gender is acquired at around 3;6 years of age by monolingual speakers of MG (Marinis 2003; Mastropavlou 2006), although correct gender assignment is not stable across nouns and is affected by the prototypicality of suffixes and the animacy of the noun until the age of 6 (see Varlokosta and Nerantzini 2013 as well as references therein). Concerning acquisition of gender assignment to object clitics in MG, it has been shown on the basis of spontaneous speech, that the most frequent form at the initial stages of acquisition, around the age of 2;0 is the third person neutral (see Stavrakaki and Okalidou 2016 and references therein).

Children learning other languages with few and inconsistent cues, such as Dutch, do not master gender until school age (Tsimpli and Hulk 2013). Focusing on bilingual children acquiring MG, the major questions in previous research have been the crosslinguistic influences, the role of input, the role of age of onset and the correlation with language proficiency. Previous studies yield a complex picture with findings being quite contradictory. Unsworth et al. (2014) provide evidence for input effects (percentage of exposure to MG), with the lexical knowledge predicting performance only for neuter gender. On the contrary, Egger et al. (2018) found that correct gender marking of determiners or adjectives accompanying a noun correlates with lexical knowledge but not with input for simultaneous bilinguals. The authors suggest that input is not crucial for a language with relatively transparent cues such as MG. Other studies advocate for correlations with both lexical knowledge and input (cumulative input and language use at the time of testing) (Kaltsa et al. 2020; Kaltsa et al. 2019; Prentza et al. 2019). Age of onset has not been found to be a predictor for correct gender marking (Kaltsa et al. 2019; Unsworth et al. 2014). Nevertheless, it seems to be relevant at least to some extent, as simultaneous bilingual children (but not

successive bilinguals) performed as the monolinguals in the study of [Unsworth et al. \(2014\)](#). Concerning crosslinguistic influences, the lack of transparent cues in the other language spoken (e.g., Dutch) by the bilingual children does not cause a delay in gender marking in MG ([Egger et al. 2018](#)), whilst the existence thereof in the other language spoken (e.g., German or Albanian) enhances their performance in gender marking in MG ([Kaltsa et al. 2020](#); [Kaltsa et al. 2019](#)).

4. The Present Study

4.1. Aims, Research Questions and Predictions

The aim of this study is to investigate the acquisition of clitics by simultaneous bilingual speakers (henceforth 2L1) of MG by comparing their acquisition pattern to that of monolingual speakers in terms of both production and comprehension. The present study capitalizes on gender as one of the features that can help us to investigate the nature of clitics beyond their overt realization vs. omission. In particular, we manipulated match/mismatch of the gender of the clitic (and its antecedent) with the gender of the subject noun of the sentence in which the clitic appears. The following research questions are addressed in the present study: (1) How does gender marking affect the production of clitics by 2L1 children? (2) How does gender marking affect the processing of clitics by 2L1 children? (3) Does proficiency have an effect on the performance in the production and processing of clitics by 2L1 children? See [Table 3](#) for a summary of the research questions and predictions. By gender marking, we mean the morphological realization of the value of the gender feature on the clitic, which has to be assigned during production and processed during comprehension in order to establish a link between clitic and antecedent.

In order to address the first research question, we manipulated gender match as the independent variable and accuracy as dependent variable (see [Table 3](#), rows referring to the first research question for examples of Gender Match and Gender Mismatch sentences). For answering the second research question, we manipulated both gender match and grammaticality as independent variables and reaction times as dependent variables (see [Table 3](#) for examples of grammatical sentences with gender match and gender mismatch and ungrammatical sentences with gender match). The manipulation of gender match can give us insight into the process of establishing a link between antecedent and clitic when there is potentially interfering information concerning the gender feature. In other words, we capitalize on the role of gender and investigate its contribution to sentence processing by examining whether gender (mis)match drives successful linking between the clitic and its antecedent. Furthermore, the manipulation of grammaticality sheds light into the role of gender in detecting violations concerning the linking between the clitic and its antecedent. Finally, in order to investigate the effect of proficiency, we adopted two independent measures of language proficiency. The first one is a composite measure of verbal intelligence, which refers to grammar, vocabulary and pragmatics, and the second one is sentence repetition, a sensitive measure of syntactic production abilities and of language development and impairment ([Talli and Stavrakaki 2020](#)).

Our study is innovative in the following aspects. First, our participants are substantially younger than in previous studies, so that earlier phenomena in bilingual acquisition can be captured. Second, the typological characteristics of the language pair spoken by our participants can give us insight into possible interference patterns from one language to the other. In particular, the tested language pair is German-MG. The two languages are similar concerning the tripartite gender distinction (masculine, feminine, neutral), the categories on which gender is marked and the lack of one-to-one correspondence of phonological cues with gender categories (masculine, feminine, neutral). MG and German differ concerning clitics, as clitics are available only in MG. Therefore, we can explore whether bilingual children can use a morphosyntactic category which is available in both languages in the same way as monolingual speakers in order to process structures which are existent only in the one language.

Table 3. Research questions, predictions, example sentences (if the prediction concerns difference in the performance between conditions) and the rationale of each prediction. For detailed descriptions of the example sentences the reader is referred to Section 4.3.

Research Question	Prediction	Example Sentence	Rationale/Previous Literature
RQ1: How does gender marking affect the production of clitics by L2 children?	Worse performance in gender mismatch than gender match for both groups	<p>Gender Match: (O gaidaros) ton filaei (The donkey_{MASC}) kisses him</p> <p>Gender Mismatch: (To provato) tin klotsaei (The sheep_{NEU}) is kicking her</p>	Gender of the subject could interfere in the gender assignment on the clitic.
General predictions about production	Prediction 1: Gender errors for both groups	across conditions	Varlokosta et al. (2014); Chondrogianni (2007)
	Prediction 2: L1 and simultaneous 2L1 no omissions	across conditions	Clitics are fully acquired in terms of omission from the age of two (Tsakali and Wexler 2004)
	Prediction 3: 2L1 worse performance than L1 (more omissions)	across conditions	Andreou et al. (2015): older simultaneous L2 children use clitics less frequently than L1 children in spontaneous speech
RQ2: How does gender marking affect the processing of clitics by 2L1 children?	Prediction 1: L1 and 2L1 same pattern: gender match effect (longer RTs for gender match than gender mismatch)	<p>Gender Match: O vatrachos pezei kai o kokoras ton vafei. The frog_{MASC} plays and the rooster_{MASC} paints him.</p> <p>Gender Mismatch: O vatrachos pezi kai i katsika ton vafei. The frog_{MASC} plays and the goat_{FEM} paints him</p>	In gender match, there are two competing NPs to which the clitic might refer to
	Prediction 2: L1 and 2L1 same pattern: Grammaticality effect (longer RTs at the critical segment in ungrammatical sentences)	<p>Grammatical (gender match) O vatrachos pezei kai o kokoras ton vafei. The frog_{MASC} plays and the rooster_{MASC} paints him.</p> <p>Ungrammatical (gender match) O vatrachos pezei kai o kokoras tin vafei. The frog_{MASC} plays and the rooster_{MASC} paints her.</p>	Chondrogianni et al. (2015)
	Prediction 3: L1 and 2L1 same pattern but 2L1 slower at the post-critical segment	across conditions	Chondrogianni et al. (2015)
	Prediction 4: 2L1 less sensitive to gender violations: Interaction Grammaticality by Group	See examples above	Rossi et al. (2014)
RQ3: Does proficiency have an effect on the performance in the production and processing of clitics?	Effect of verbal intelligence and/or sentence repetition	across conditions	Andreou et al. (2015); Chondrogianni (2007, 2008)

Despite these novel aspects, previous findings can help us to formulate predictions. Concerning the effect of gender marking on clitic production (first research question), it is possible that both groups perform worse when they have to produce sentences, in which the gender of the clitic antecedent and of the subject are different (gender mismatch condition), on the grounds that the gender of the subject could interfere in the gender assignment on the clitic. Furthermore, as previous studies attest gender errors to a great extent until the age of 5;7 (Varlokosta et al. 2014) for monolingual children and older successive bilingual children (Chondrogianni 2007), we expect gender errors to be present in both groups. Independently of the manipulation of gender match, we expect that the L1 children perform at ceiling in terms of production rates (even if the clitic form is wrong), as in their age clitics are supposed to be fully acquired in this respect. However, it is possible that 2L1 children omit more clitics in the production task than the L1 children given the vulnerability of object clitics in bilingual acquisition even for older 2L1 children (Andreou et al. 2015). Alternatively, since our 2L1 participants have been exposed to MG from birth, they could behave like the L1 children. The reader is referred to Table 3 for a summary of the predictions, the rationale and examples of the sentences.

With respect to processing (second research question), since 2L1 children have been exposed to MG from birth it is possible that they manifest the same pattern as the L1 group. In particular, it is possible to observe a gender match effect, i.e., longer RTs for both groups in the sentences in which the clitic has the same gender as the subject noun (gender match), since there are two competing NPs to which the clitic might refer to, even if one of the two is structurally inaccessible (ungrammatical gender match sentences, see Table 3, rows referring to the second research question). Moreover, one can expect an effect of ungrammaticality, namely longer RTs at the critical segment in ungrammatical sentences as in Chondrogianni et al. (2015). However, if they have not achieved native-like processing abilities, there are two possibilities: (1) the 2L1 children's processing pattern is similar to that of L1 children, but they are slower than L1 children at the post-critical segment, such as in the study of Chondrogianni et al. (2015) or (2) 2L1 children might be less sensitive to gender violations, as was found in the study of Rossi et al. (2014) for adults, which would mean that there will be no difference between grammatical and ungrammatical sentences for the 2L1 children.

As regards the effect of language proficiency on the production of clitics (third research question), since measures of proficiency have been shown to affect clitic production in L2 (Andreou et al. 2015; Chondrogianni 2007, 2008), we expect an effect of at least one of our language measurements (verbal intelligence or sentence repetition) on the production of clitics for the 2L1 group.

4.2. Participants

Sixteen monolingual typically developing Greek speaking children (hereafter L1) and fourteen simultaneous bilingual typically developing children who acquire German and MG (hereafter 2L1) participated in this study. The L1 children were recruited through kindergartens and the third author's familiar environment in Athens, Greece. The 2L1 children were recruited through the German School of Athens (Deutsche Schule Athen), both from the kindergarten and the preschool departments. Two additional L1 participants were excluded because the testing could not be completed. This results to 30 participants in total. The mean age of the L1 group at the time of testing was 5;6 years (SD = 7.3, range = 6;5 to 4;5), while the 2L1 group had a mean age of 5;3 years (SD = 8.8 range = 7;0 to 4;6). The two groups were matched for age in months ($t(26) = 1.1, p > 0.05$).

The linguistic and cognitive abilities of the participants were assessed by means of verbal and non-verbal intelligence tests and a parental questionnaire, as mentioned above. The tests were used to confirm the typical development of the participants, while the parental questionnaire provided further information regarding the language background of the children, as well as the family risk of language impairment. The Greek version (Vogindroukas et al. 2010) of the Action Picture Test (Renfrew 1997) is a general measure of

verbal intelligence, which taps on lexical, syntactic and pragmatic abilities. In this test, the child is presented with a picture and has to answer a relevant question. The complexity of the pictures and the questions vary. For instance, in one picture, a girl holds a teddy bear and the question is “What is the girl doing?”. Another picture depicts a more complex scene with a woman walking on the pavement, an apple dropping from her shopping bag and a boy trying to catch it. The question is “What is happening here?”. Furthermore, we ran the sentence repetition part of the Diagnostic Verbal IQ Test (DVIQ) (Stavrakaki and Tsimpli 2000) to measure the syntactic production abilities of the participants, who in this case had to repeat sentences after the researcher had produced it. Sentence repetition is considered as a measurement of syntactic production abilities (Klem et al. 2015), and the subtest of DVIQ has been used as a sensitive measure of language development and impairment in MG (see Talli and Stavrakaki 2020 and references therein). Lastly, for the non-verbal intelligence, we used the Raven’s Coloured Progressive Matrices test (Raven et al. 1998).

The exclusion criteria were performance below one standard deviation on Raven’s Coloured Progressive Matrices, a history in language disorders, hearing or neurological impairment and diagnosis of autism. According to the information reported in the parental questionnaire, four L1 children had received speech and language therapy for articulation problems. Two bilingual children had received speech and language therapy, with no further information being provided for the one child, whereas for the other child, the therapy targeted receptive vocabulary and articulation. Since receptive vocabulary and articulation problems do not interfere with the conduction of the tasks and all these children performed within the range in the relevant linguistic abilities tests, these two children were not excluded.

Table 4 presents the mean scores of the tests, their range and standard deviation for each group. In all tests, except for the sentence repetition task, which is not standardized, the standard scores are used. Two-sample equal variance t-tests were run on the scores of the verbal and non-verbal intelligence tests to check the matching across the groups. The groups were matched on verbal intelligence, ($t(28) = 0.56, p = 0.6$), sentence repetition ($t(18) = 0.64, p = 0.5$) and the non-verbal intelligence test ($t(28) = 0.43, p = 0.7$).

Table 4. Performance of the participants on verbal and non-verbal intelligence tests. APT: Action Picture Test, SRT: Sentence Repetition Task, CPM: Colored Progressive Matrices, SD: standard deviation. In all tests, except for the Sentence Repetition Task, which is not standardized, the standard scores are used.

Group	APT Mean (SD) Range	SRT Mean (SD) Range	CPM Mean (SD) Range
L1 (N = 16)	82 (9) 59–96	44.81 (2.78) 39–48	126.2 (8.46) 105–140
2L1 (N = 14)	83.78 (7.18) 71–102	43.71 (5.85) 27–48	125 (7.59) 115–140

The 2L1 children are simultaneous bilinguals (German–MG), as they were exposed to both languages from birth. In order to provide a detailed picture of the individual language profiles, we used a questionnaire, which was an adaptation of the one used by Torregrossa et al. (2021) (see Appendix A for an English translation of the questions on the input). In particular, we computed the following variables (see Table 5): (1) sum of input in both languages and (2) sum of input in both languages for the first 6 years of life (split in intervals of three years). Sum of input was computed by summing the products of the number of hours spent with a specific person (e.g., mother, father, siblings, teachers) by the frequency with which each specific person used each language with the child (e.g., almost never, rarely, half of the time spent, normally, almost always). Each of these frequency levels was assigned a numerical value (1, 2, 3, 4 and 5, respectively). For example, if the mother

spoke with the child almost never in Greek (frequency = 1) and spent with him/her 50 h per week, then the input of the mother would be 50. The same procedure was followed for each person of the family/caregiver or language used in various activities. These indexes do not have a meaning per se (in contrast to hours of exposition, for instance), but they suggest whether the child is exposed to more, less or the same “quantity” (input) of each language on a daily basis. In order to assess the language exposure in the first years of life, we added the frequencies of usage of each language for each person (mother, father, sibling etc.). As can be seen in the table, all children were exposed to both languages from birth, and MG was the dominant language at the time of testing for ten participants. For 2L1_1 and 2L1_9, the dominant language was German and for 2L1_8 and 2L1_14 sum of input in German and in MG was the same.

Table 5. Demographic and language background characteristics of the 2L1 participants (age, birthplace, age of moving to Greece, language input and output in each language and sum of input in the first 6 years).

Code	Age	Birthplace	Age of Moving to Greece (in Months)	Sum of Input MG	Sum of Input German	Sum Input in MG between 0 and 3 Years	Sum Input in German between 0 and 3 Years	Sum Input in MG between 3 and 6 Years	Sum Input in German between 3 and 6 Years
2L1_1	4;10.3	Greece	0	670	800	12	12	12	18
2L1_2	5;5.0	Greece	0	480	96	15	3	17	7
2L1_3	5;0.28	Greece	0	455	325	10	8	10	8
2L1_4	4;8.9	Greece	0	1080	960	16	8	17	13
2L1_5	5;2.0	Greece	0	461	253	13	5	13	5
2L1_6	5;3.5	Aithiopia	5	565	245	14	4	12	6
2L1_7	4;9.10	Greece	0	431	277	13	5	15	9
2L1_8	5;3.14	Greece	0	450	450	9	9	9	9
2L1_9	4;7.27	Cyprus	8	355	455	9	9	12	12
2L1_10	4;8.14	Greece	0	800	670	12	6	15	9
2L1_11	6;5.16	Greece	0	479	163	15	3	17	7
2L1_12	7;0.17	Germany	16	255	135	18	6	22	8
2L1_13	5;8.15	Greece	0	470	418	13	11	16	14
2L1_14	6;3.23	Greece	0	294	294	10	8	11	13

4.3. Design, Materials and Procedure

4.3.1. Production Task

In this task, we manipulated gender match (match vs. mismatch) between two characters participating in one action with the noun phrases (NPs) having either the same (match) or different gender (mismatch). Table 6 presents two examples illustrating the design of the production task.

Table 6. Design of the production task.

Condition	Question						Target Response	
Gender Match	Ti What 'What is the donkey doing to the rooster?'	kani do _{3sing}	o the	gaidaros donkey _{MASC}	ston to the	kokora? rooster _{MASC} ?	Ton _{MASC} Him	filaei. kiss _{3sing} 'He is kissing him.'
Gender Mismatch	Ti What 'What is the sheep doing to the cow?'	kani do _{3sing}	to the	provato sheep _{NEU}	stin to the	agelada? cow _{FEM} ?	Tin _{FEM} Her	klotsai. kicks _{3sing} 'He is kicking her.'

There were 24 experimental items used in the task, 12 for each condition, which together with 5 practice items made 29 items in total. All nouns were animal characters, divided into the three genders, namely four masculine (*gaidaros* = donkey, *kokoras* = rooster, *panthiras* = panther, *vatrachos* = frog), four feminine (*chelona* = turtle, *agelada* = cow, *melisa* = bee, *katsika* = goat) and four neutral (*alogo* = horse, *guruni* = pig, *kuneli* = bunny, *provato* = sheep). All characters were equally combined with each other based on the gender (masc.–

masc., fem.–fem., neut.–neut., masc.–neut., masc.–fem. and neut.–fem.). Every combination appeared twice. The selection of the NPs was based on the syllable length (always three syllables), on the age of acquisition, checked through Greek corpora (CHILDES) and on their appearance in other studies with similar participants, also taking the age into account (Chondrogianni et al. 2015; Varlokosta et al. 2016). Post hoc control with the normative data of Dimitropoulou et al. (2009) showed that all nouns (except for *panther*, for which there is no available data) are acquired before the age of 2;6 (average age). We investigated whether the grammatical gender coincides with the natural gender of the nouns. For details about this procedure, see Appendix B. For all but two of the nouns (*kokoras* = rooster and *katsika* = goat), there was not any overlap between grammatical and natural gender. The verbs that expressed the actions in which the characters were engaged were six highly depictable transitive actions (*filai* = kiss, *klotsai* = kick, *dagkoni* = bite, *agapai* = love, *tsimbai* = pinch, *vrechi* = wet). The verbs were selected again based on Greek corpora (CHILDES) and on the same previous work. There were two lists created with different order of the items.

For the assessment of the children’s production of gender on clitics, we used a picture elicitation task, following Stavrakaki and van der Lely (2010) and Chondrogianni (2008). In this task, children were shown a set of pictures, which depict two animal characters. The experimenter introduced the animals with a sentence of the type “Here there is a X and a Y”. After the introduction children saw a second picture, in which the character X does something to the character Y and are asked the question “What is the X doing to Y?”. This question is expected to elicit a clitic pronoun (see also Table 6) at a typically developing child from age 5 onwards.

4.3.2. Comprehension Task

In the comprehension task, gender match (match vs. mismatch) and grammaticality (grammatical vs. ungrammatical) were manipulated. An online self-paced listening task with picture verification was used for testing comprehension of clitics, namely, in this case, their ability to detect the wrong gender on clitic pronouns. This task has been used widely in the investigation of pronouns resolution (Stewart et al. 2007; Wolf et al. 2004) and according to Marinis (2010), self-paced listening task is child-friendly. The participants define their own rhythm, adapting the task to their needs. Apart from the RTs, the picture verification offers extra information on the concentration of the participants.

The comprehension task consisted of 48 experimental items in total, distributed to four conditions (12 sentences per condition), split into two lists. Table 7 summarizes the design and provides examples of the four conditions of the comprehension task.

Table 7. Example of the conditions in the comprehension task. MASC: masculine, FEM: feminine.

Cond.	Sentence	Factor	Level
1a	<i>O vatrachos pezi ke/ o kokoras/ TON/ vafi/ me ta chromata.</i> The frog _{MASC} plays _{3sing} and/ the rooster _{MASC} / him/ paint _{3sing} / with the colors. The frog plays and the rooster paints him with the colors.	Gender of NPs Grammaticality	Match Grammatical
1b	<i>O vatrachos pezi ke/ i katsika/ TON/ vafi/ me ta chromata.</i> The frog _{MASC} plays _{3sing} and/ the goat _{FEM} / him/ paints _{3sing} / with the colors. The frog plays and the goat paints him with the colors.	Gender of NPs Grammaticality	Mismatch Grammatical
1c	<i>*O vatrachos pezi ke/ o kokoras/ TIN/ vafi/ me ta chromata.</i> The frog _{MASC} plays _{3sing} and/ the rooster _{MASC} / her/ paints _{3sing} / with the colors. The frog plays and the rooster paints her with the colors.	Gender of NPs Grammaticality	Match Ungrammatical
1d	<i>*O vatrachos pezi ke/ i katsika/ TIN/ vafi/ me ta chromata.</i> The frog _{MASC} plays _{3sing} and/ the goat _{FEM} / her/ paints _{3sing} / with the colors. The frog plays and the goat paints her with the colors.	Gender of NPs Grammaticality	Mismatch Ungrammatical

All sentences contained two animal characters NPs, engaged with each other with a certain action. The same 12 characters as in the production task were used. Concerning the actions, there were 12 verbs used in this task (*iremi* = calm, *agkaliazi* = hug, *ksipnai* =

wake up, zografizi = draw, vafi = paint, vriski = find, skepazi = cover, piani = catch, gargalai = tickle, pleni = wash, akui = listen, akoluthi = follow). The verbs were checked for the age of acquisition in Greek corpora (CHILDES), and they had all been used in previous studies (Chondrogianni et al. 2015; Varlokosta et al. 2016). Half of the experimental sentences in each list were ungrammatical and half grammatical, and within grammatical and ungrammatical sentences, in half of the sentences, the genders of the two NPs were the same and in the other half, different. We have striven to make the sentences as interesting for the children as possible. An example of how an experimental trial and a comprehension question looked like is illustrated in the Examples (2) and (3) below. The dashes represent the segments. Figure 1 presents the pictures that accompany the experimental trial (panel a) and the comprehension question (panel b).



Figure 1. Pictures for experimental items. (a) Picture appearing with the experimental trial; (b) Picture appearing with the comprehension question.

(2) Experimental trial.

Introductory part: To guruni echase to pechnidi tu. (one segment)
The pig lost its toy.

Experimental part: To guruni/ klei/ ke/ to alogo/ TO/ agkaliazi/ sfichta.
The pig/ is crying/ and/ the horse/ IT/ is hugging/ tight.

(3) Comprehension question.

Question: Pjos klei? Who is crying?

In the comprehension question in (3), the participants had to press the left or the right arrow, according to the position of the correct character on the screen. There were 16 experimental questions in total, in half of which the left was the correct character, while in the other half the right one. The items were distributed in two lists, whereas the fillers and the practice items were the same in both lists.

We used 24 fillers, which had a similar structure with the experimental items, with the difference that only the verb “is” was used. Therefore, there was not an action expressed but rather a description of the characters, achieved through adjectives (e.g., *thimomenos* = *angry*). This was carried out in order to enrich the stimuli with sentences that contributed to a smooth connectivity among the sentences, and make the visual and auditory stimuli look like a story of animals. They did not contain any kind of pronouns at all, and half of them were grammatical and half ungrammatical. The error was on the gender assignment of the adjective to the NP. Example (4) illustrates an ungrammatical filler.

(4) Introductory part:

O kokoras den thelei na einai defteros. (=The rooster does not want to be second)

*O kokoras einai thymomenos kai o gaidaros einai iremi.

Main part:

The rooster_{MASC} is angry_{MASC} and the donkey_{MASC} is calm_{FEM}.

The rooster is angry and, the donkey is calm.

A female native Greek speaker was recorded reading the sentences in a sound-isolated booth, and the auditory files were recorded and edited with Praat (Boersma and Weenink 2017). To ensure that there was not any kind of disrupted intonation among the different sets of the sentences, the critical segments, namely the clitic pronouns, were recorded separately and were placed in their position at the editing process. That way all the sentences had the same natural intonation.

At the beginning of each trial, the participants saw a picture (see Figure 1a for an example) on the screen, and 0.5 s later, they heard an introductory sentence, which named the two characters and the introductory part (see 2). This occurred to create a story-wise sequence in the action. The introductory sentence was then followed by the experimental part. This sentence was segmented into words or small phrases, consisting always of seven segments. Together with the introductory phrase, which always forms a solid segment, each item consists of eight segments in total. The accusative form of the clitic pronouns in the three genders (*ton* = him, *tin* = her, *to* = it) was the critical segment of the task, the subject of the sentence in which the clitic appears was the precritical segment, the verb was the postcritical segment and the last constituent, an adverbial was the final segment (see example 2 and Table 7 for examples). The participants were instructed to press the space button on the keyboard as fast as they can, in order to continue hearing the segments and to understand exactly what happened. Every three trials, a picture with the two characters of the last trial appeared on the screen (see Figure 1b for an example) and, shortly after the participants heard a comprehension question relevant to the last trial, of the type “Who did the action?”. The participants had to press the left or the right arrow, according to the position of the character on the screen. These questions did not offer any kind of extra valuable measurements regarding the comprehension of the experimental sentences, but they were rather a check of the participation of the children during the experiment. The experimental trials were preceded by 10 practice trials. The task was programmed on the *OpenSesame 3* (Mathôt et al. 2012).

Before testing, we run a piloting testing with three Greek-speaking children living in Berlin and two living in Thessaloniki. All five children seemed to be able to successfully follow the whole process; thus, the material has not been changed. The whole testing included two or three sessions. At the beginning of the first session, the researcher was familiarized with the child by talking or playing together. Then, half of the verbal and non-verbal intelligence tests were run and lastly the production task. At the second session, the remaining intelligence tests together with the comprehension task were administered. The comprehension task always followed the production task, on a 3- to 6-day interval. A third session took place only in five cases, where the participants’ mood or external circumstances impeded the completion of the whole process in the second session. The L1 children were tested either in their school environment or at their homes. Teachers, parents or caregivers were not present, except in two cases, where the presence of the parent was considered necessary, as the child seemed to feel more confident, although no interaction with the parent took place during the testing procedure. The 2L1 children were tested in their school environment, at an adequately designed room, which was normally used for movie projections. In this case, only the researcher and the child were present during the process.

5. Results

5.1. Analyses

The data were analysed using the R software (R Core Team 2020). We fitted generalized linear mixed-effects models for the production data and linear mixed effects model for the RTs of comprehension data using the R package *lme4* (Bates et al. 2015). We used (generalized) linear mixed-effects models because they take into account both the variance between subjects and between items (random intercepts) and the variance between subjects and between items for all factors (random slopes) (Bates et al. 2018 among others). The generalized linear mixed-effects models can handle binary responses (Agresti 2019). Moreover, we used the following packages: *RePsychLing* (Baayen et al. 2015) and *performance* (Lüdtke et al. 2020) for choosing the structure of random effects that is supported by the data, *emmeans* (Lenth 2020) for the post hoc comparisons and *strengjacker* (Lüdtke 2019) for extracting the model parameters in tables. We used the following procedure for analysing the data. First, we fitted a model motivated by the research questions and the predictions (initial model). Then, we applied several tests (e.g., test for singularity) to examine whether

the initial model was supported by the data, and successively, we simplified the structure of the random effects until the model fitted was supported by the data (final model). We will report only the output of the final models. More details for each model fitted are to be found in Appendix C (for production) and E (for comprehension) to which the reader will be referred to.

5.2. Production

Table 8 presents the mean percentage of all response types as well as the standard deviation. As shown in the table, the two groups perform similarly in terms of correct responses, with sentences with gender match being easier than sentences with gender mismatch. The two groups perform similarly also in terms of error types, as the production of wrong gender is the most common error in both groups. However, there are few gender errors in the gender match condition and about 20% in the gender mismatch condition in both groups. Since for 4 out of the 14 2L1 children, the dominant language was not MG, there might be an effect of dominance, as an anonymous reviewer pointed out. In order to examine this, we removed these four children from the 2L1 dataset to see whether the correct responses or some error type change. After the slashes in the panel of the table for the 2L1 children, one can see the results of the subset of the 2L1 children with MG as the dominant language at the time of testing. The differences are minimal except for the percentage correct responses in the mismatch condition, which, surprisingly, is higher when the whole group is taken into account than in the only MG dominant group. We will come back to this later in the analysis.

Table 8. Mean percentage and standard deviation of all response types for each group in each condition. L1: monolingual group, 2L1: bilingual group, 2L1-MG Dominant: the subset of 2L1 children with MG as dominant language, Match: sentences in which the clitic and the subject noun have the same gender, Mismatch: sentences in which the clitic and the subject noun have different gender. Correct: correct responses, Clitic_omission: omission of the clitic, Wrong_clitic: production of a clitic with wrong number or case, Gender_error: production of a clitic with wrong gender, Other errors: any error that did not fall in the above categories.

Group	Gender Match	Response Type	Percentage Correct Response (SD)
L1	Match	Correct	83 (0.10)
		Clitic omission	3 (0.05)
		Wrong clitic	2 (0.05)
		Gender error	8 (0.08)
		Other	4 (0.06)
	Mismatch	Correct	73 (0.10)
		Clitic omission	4 (0.05)
		Wrong clitic	0 (0.00)
		Gender error	23 (0.10)
		Other	0 (0.00)
2L1 (all children)/ 2L1-MG Dominant	Match	Correct	82 (0.13)/82.5 (0.15)
		Clitic omission	8 (0.09)/9 (0.11)
		Wrong clitic	0 (0.00)
		Gender error	8 (0.095)/7.5 (0.092)
		Other	2 (0.04)/1 (0.03)
	Mismatch	Correct	73 (0.19)/69 (0.21)
		Clitic omission	5 (0.11)/7 (0.13)
		Wrong clitic	0 (0.00)
		Gender error	21 (0.17)/23 (0.19)
		Other	1 (0.02)/1 (0.03)

As mentioned above, besides the language experiment, the verbal abilities of the participants were assessed as well. We examined the effect of the two verbal abilities scores,

the Action Picture Test (verbal intelligence) and the sentence repetition part of the DVIQ test (syntactic production abilities), on the performance of the participants in the language task. Figure 2 plots the accuracy scores of each group in the two conditions against the scores in the verbal intelligence test. As can be seen in the plot, the 2L1 group had a smaller range of intelligence scores (71–102) than the L1 group (59–96) but the accuracy scores are more dispersed (see also the SD in Table 8), which indicates more variability in performance. From a visual inspection of the left upper panel, the accuracy in the gender match sentences seems to correlate with the verbal intelligence raw scores for the L1 group, but this is not the case for the gender mismatch sentences (right upper panel). Concerning the 2L1 group (two lower panels), trends for a negative correlation in gender match and a positive correlation in gender mismatch are visible; however, they might not be reliable given the large variability.

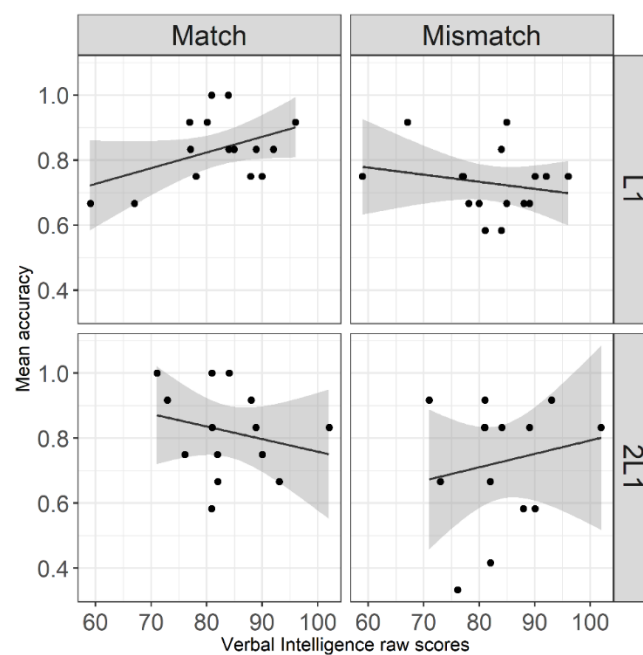


Figure 2. Scatterplot of the accuracy scores and the verbal intelligence scores.

We first analysed the effect of verbal intelligence on accuracy (by means of a generalized linear mixed effects model). The model motivated by our research questions and predictions (see Table 3, first and third research questions for a summary) contained the following predictors: the two-level factor group (L1 vs. 2L1), the two-level factor gender match (match vs. mismatch) and the continuous covariate verbal intelligence which was operationalized by the score in the Action Picture Test (Vogindroukas et al. 2010). The details about contrasts, the model selection procedure, the formulas of the initial and the final model, the one supported by the data, as well as its output are in Section 1 and Table A2 of Appendix C. The model revealed a main effect of gender match ($z = -2.07$, $p < 0.05$) and an interaction gender match by group by verbal intelligence ($z = 2.22$, $p < 0.05$). As can be seen in Figure 2, verbal intelligence does not have an effect on accuracy in neither condition for the 2L1 group, probably due to the large variability in the accuracy scores, whereas for the L1 group, the higher the verbal intelligence, the higher the accuracy, but only in the gender match condition.

A similar analysis was performed for the effect of the syntactic production abilities, which was operationalized by the sentence repetition part of the DVIQ test (Stavrakaki and Tsimpli 2000). Figure 3 plots the accuracy scores of each group in the two conditions against the scores in the sentence repetition task. Both groups had a relatively high score except for a participant with a very low score in the 2L1 group. On the basis of exploratory visual inspection, the performance of the L1 group in the gender match sentences seems

to correlate positively with the scores in the sentence repetition, whereas there seems to be a negative correlation for the gender mismatch. For the 2L1 group, there is a trend for positive correlation in both conditions, but the correlation appears to be weaker in the gender match condition. We analysed the data by means of a generalized linear mixed model. The initial model, based on our research questions and predictions (see Table 3, first and third research questions for a summary), included the two factors gender match and group and the covariate sentence repetition and their interaction as predictors (for details see Section 2 and Table A3 of Appendix C). The analysis showed that there is not any significant effect or interaction, except for the three-way interaction (gender match by group by sentence repetition), which failed to reach significance ($z = 1.93, p = 0.054$). We also ran the same model after removing the outlier, which yields a non-significant three-way interaction ($z = 0.95, p = 0.343$), suggesting that the former quasi-interaction was driven by the outlier (see Section 2 and Table A4 of Appendix C for the formula and the full output of the model).

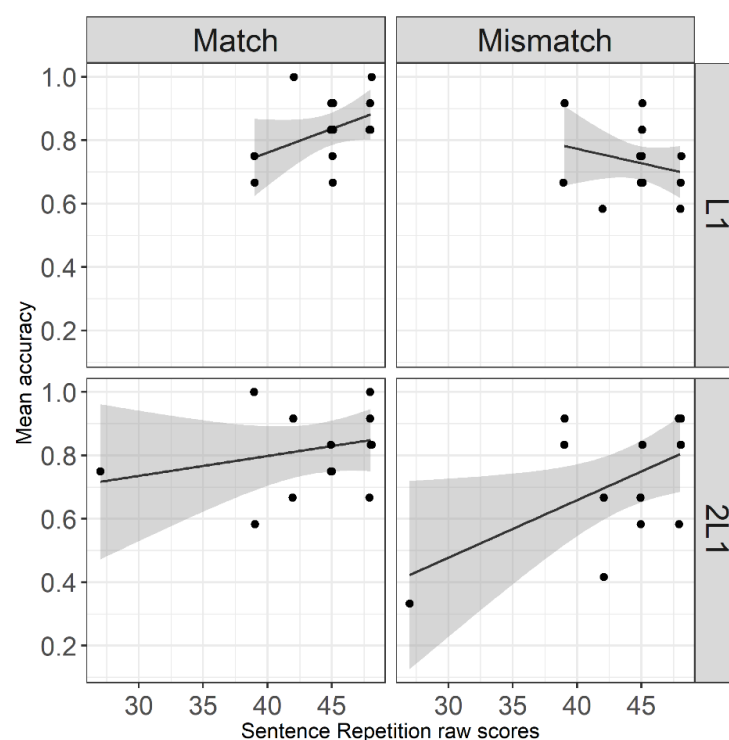


Figure 3. Scatterplot of the accuracy scores with the scores in the sentence repetition test.

The statistical methods we used take into account the individual variability (random intercepts and random slopes for individuals). However, as the bilingual participants were heterogeneous in terms of input, we performed an a posteriori analysis investigating whether their performance is related to some input measure. In this way, we also examined whether there is a language dominance effect, as an anonymous reviewer also suggested. Thus, we performed correlation analyses between the different input measures we collected through parental questionnaire (see Table 5). We used the non-parametric correlation (Spearman's), because the sample size was small. Table 9 presents the correlation coefficients and the p values for the correlations between each input measure and the correct performance in each of the two conditions for the bilingual group.

Table 9. Correlation coefficients and p-values for the correlations between correct performance in each condition and the measurements of sum of input in each language (current exposure), sum of input in each language between 0 and 3 years (total exposure in the first 3 years of life), sum input in each language between 3 and 6 years (total exposure between 3 and 6 years). MG: Modern Greek.

Input Measure	Condition	
	Match	Mismatch
Sum of input MG	$r_s(12) = 0.29, p = 0.31$	$r_s(12) = 0.16, p = 0.58$
Sum of input in MG between 0 and 3 years	$r_s(12) = 0.49, p = 0.072$	$r_s(12) = 0.039, p = 0.9$
Sum input in MG between 3 and 6 years	$r_s(12) = 0.27, p = 0.36$	$r_s(12) = -0.14, p = 0.64$
Sum of input German	$r_s(12) = -0.13, p = 0.65$	$r_s(12) = 0.061, p = 0.84$
Sum of input in German between 0 and 3 years	$r_s(12) = -0.27, p = 0.36$	$r_s(12) = 0.087, p = 0.77$
Sum of input in German between 3 and 6 years	$r_s(12) = -0.25, p = 0.39$	$r_s(12) = 0.15, p = 0.6$

None of the measurements seems to be related to the performance in either condition, as the correlation coefficients were not significant. A reason for this could be the relatively small sample size. Nevertheless, there is a tendency for significance in the correlation between the performance in the sentences with the same gender (match) and the input in MG in the first three years of life.

Although the low correlation coefficients indicate that there is not a relation between current input and performance in the production task, we further investigated the issue of dominance, by reanalysing the whole dataset inserting a categorical variable dominance. All children, except for the two with German as dominant language and the two with the same amount of input in both languages, were considered as MG-dominant (i.e., also the monolingual). We fitted the same models for verbal intelligence and sentence repetition, respectively, adding dominance as a main effect. Neither model yielded a significant effect for language dominance. Second, we analysed the data of the bilingual group, by splitting it into two subgroups, MG-dominant and not MG-dominant, and we analysed the data with identical models, as the ones we reported in the main analyses, with the only difference that the predictor group was substituted by dominance. Again, dominance did not have an effect or participate in any interaction neither for verbal intelligence nor for sentence repetition. See Appendix C, Section 3 and Tables A5–A8.

Finally, we investigated patterns of correct responses and substitutions for each gender combination (Table 10). As shown in the table, both groups had ceiling performance in the match combinations for feminine and masculine gender, whereas the combination neutral–neutral was difficult for both groups. Unfortunately, the substitution patterns are not available, but it is remarkable that gender errors are not the preponderant error type in this condition. Concerning the mismatch combinations, both groups had massive problems with combinations of masculine and neutral (masculine subject noun–neutral clitic and neutral subject noun–masculine clitic). Sentences with a neutral subject noun were generally difficult for the L1 group independently of the gender of the clitic. The same holds for the 2L1 group with the exception that when the subject noun is neutral and the clitic feminine (neutral–feminine), the performance is much better. Sentences with neutral clitics are also problematic for both groups. The performance of the L1 group in sentences with feminine clitics and masculine subject noun (masculine–feminine) is quite good (84%) but worse for the 2L1 group (75%), whereas sentences with feminine subject noun and masculine clitic (feminine–masculine) are relatively easy for both groups. In sum, it seems that neutral gender poses difficulties both in gender match sentences and when it is combined with another gender either as subject or as clitic.

Table 10. Percentage and standard deviations of correct responses and responses with gender errors for each gender combination for the two groups. Feminine–feminine, masculine–masculine, neutral–neutral: both subject noun and clitic have the same gender (feminine, masculine and neutral, respectively), feminine–masculine: subject noun is feminine and clitic masculine, feminine–neutral: subject noun is feminine and clitic neutral, masculine–feminine: subject noun is masculine and clitic feminine, masculine–neutral: subject noun is masculine and clitic neutral, neutral–feminine: subject noun is neutral and clitic feminine, neutral–masculine: subject noun is neutral and clitic masculine.

Group	Gender Combination	Response	Mean Percentage (SD)
L1	feminine–feminine	Correct	0.89 (0.13)
		Gender error	0.05 (0.10)
	feminine–masculine	Correct	0.94 (0.17)
		Gender error	0.06 (0.17)
	feminine–neutral	Correct	0.72 (0.31)
		Gender error	0.28 (0.31)
	masculine–feminine	Correct	0.84 (0.24)
		Gender error	0.09 (0.20)
	masculine–masculine	Correct	0.94 (0.11)
		Gender error	0.02 (0.06)
	masculine–neutral	Correct	0.59 (0.42)
		Gender error	0.31 (0.44)
	neutral–feminine	Correct	0.63 (0.34)
		Gender error	0.34 (0.35)
	neutral–masculine	Correct	0.66 (0.35)
		Gender error	0.31 (0.36)
neutral–neutral	Correct	0.67 (0.28)	
	Gender error	0.17 (0.24)	
2L1	feminine–feminine	Correct	0.89 (0.21)
		Gender error	0.02 (0.07)
	feminine–masculine	Correct	0.89 (0.29)
		Gender error	0.07 (0.27)
	feminine–neutral	Correct	0.68 (0.32)
		Gender error	0.29 (0.32)
	masculine–feminine	Correct	0.75 (0.33)
		Gender error	0.18 (0.32)
	masculine–masculine	Correct	0.91 (0.12)
		Gender error	0.04 (0.09)
	masculine–neutral	Correct	0.54 (0.41)
		Gender error	0.43 (0.39)
	neutral–feminine	Correct	0.89 (0.21)
		Gender error	0.07 (0.18)
	neutral–masculine	Correct	0.61 (0.35)
		Gender error	0.25 (0.38)
neutral–neutral	Correct	0.66 (0.33)	
	Gender error	0.18 (0.23)	

5.3. Comprehension

Before analysing the RTs, all datapoints above 15,000 milliseconds have been removed. Figures 4 and 5 present the RTs per segment for the L1 and the 2L1 group, respectively. For exact numbers, see Tables A9 and A10 in Appendix D. Recall that the precritical segment is the segment before the clitic (subject of the sentence), the critical is the clitic, the postcritical is the segment after the clitic (verb of the sentence), and the final is the segment with an adverbial phrase. The figures suggest that the two groups perform similarly to a great extent. We repeat an example of an experimental item in (5) for convenience:

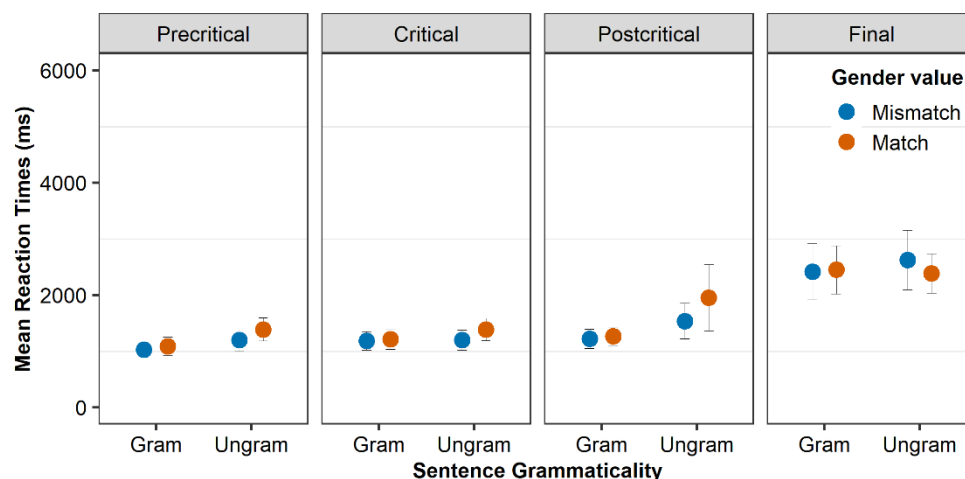


Figure 4. RTs of the L1 group per segment.

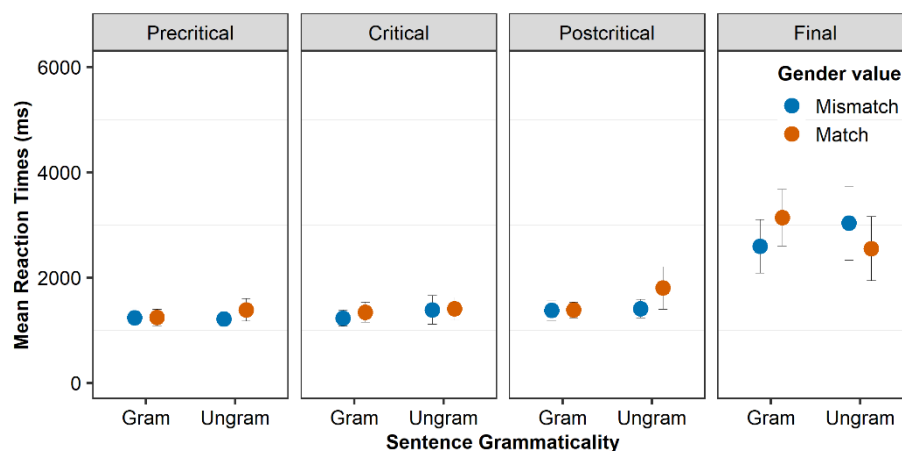


Figure 5. RTs of the 2L1 group per segment.

- (5) *O vatrachos pezi ke* / **(Precritical)** *o kokoras* / **(Critical)** *TON* / **(Postcritical)** *vafi* / **(Final)** *me ta chromata*.
 The frog plays and / **(Precritical)** the rooster / **(Critical)** him / **(Postcritical)** paints / **(Final)** with the colors.

We analysed the effect of gender match, grammaticality, group and verbal intelligence on RTs by means of a linear mixed effects model. The model included the following predictors (according to our research questions and predictions, see Table 3, second and third research questions): the two-level factor group (L1 vs. 2L1), the two-level factor gender match (match vs. mismatch), the two-level factor grammaticality (grammatical vs. ungrammatical), the four-level factor segment (precritical, critical, postcritical and final) and the continuous covariate verbal intelligence, which was operationalized by the score in the Action Picture Test. The rationale for including segment as a factor with four levels that are compared to one another in the order precritical–critical–postcritical–final allows

us to detect potential effects of incremental sentence processing and how these might be modulated depending on the (un)grammaticality of the condition and, crucially, the gender manipulation, in terms of gender match or mismatch between the two potential referents of the nouns. The advantage of such more powerful analysis (as compared to running one model for each segment) is that (a) it avoids running multiple comparisons; (b) the effects related to each comparison are estimated taking into account the variance of the whole dataset, not just a subset of it. The RTs were subject to negative reciprocal transformation, in order to avoid skewness in the distribution. For details on contrasts coding, structure of random effects, the exact formula, the procedure of selecting the best model and full output of the models see Section 1 and Table A11 of Appendix E.

The analysis showed a main effect of group ($t = -2.3, p < 0.05$), a main effect of segment (postcritical to final) ($t = 9.66, p < 0.001$), a main effect of verbal intelligence ($t = -2.06, p < 0.05$), and an interaction grammaticality by gender match by segment (postcritical–final) ($t = -2.25, p < 0.05$). The main effect of group is due to the overall longer RT of the bilingual group in comparison to the monolingual group, and the main effect of verbal intelligence suggests that the higher the verbal intelligence, the shorter the RTs in both groups. We delved into the three-way interaction by conducting post hoc pairwise comparisons with the p-values adjusted with the Tukey method. The tests showed that for all grammatical sentences independently of gender match, there was not any difference between the segments except for the difference between postcritical and final segment (for grammatical gender match: precritical vs. critical: $t(5619) = -1.24, p = 0.92$, critical vs. postcritical: $t(5620) = -1.36, p = 0.87$, postcritical vs. final: $t(5624) = -6.29, p < 0.001$, for grammatical gender mismatch: precritical vs. critical: $t(5619) = 0.22, p = 1$, critical vs. postcritical: $t(5619) = -0.68, p = 1$ and postcritical vs. final: $t(5620) = -4.53, p < 0.001$), a finding which suggests a wrap-up effect. This holds for ungrammatical sentences in which the gender was not matched (precritical vs. critical: $t(5620) = -0.91, p = 0.98$, critical vs. postcritical: $t(5620) = -0.41, p = 1$, postcritical vs. final: $t(5621) = -6.08, p < 0.001$), as well. However, for ungrammatical sentences in which the gender of the clitic and the noun were matched, there were longer reaction times in the final segment, but the difference from the postcritical segment was not significant (precritical vs. critical: $t(5620) = -0.04, p = 1$, critical vs. postcritical: $t(5621) = -0.19, p = 1$, postcritical vs. final: $t(5621) = -2.61, p > 0.05$). The difference in reaction times was the smallest in the comparison between postcritical and final segment for these sentences, as can be seen in Figures 4 and 5 (and Tables A9 and A10 in Appendix D) and by the T-ratios, and this is why the interaction emerged in this specific contrast in the model. Importantly, there was not any difference between the conditions within each segment. We will discuss these findings in the next section.

Furthermore, we tested the effect of the syntactic production abilities, operationalized by the score in the sentence repetition task of the DVIQ test (Stavroulaki and Tsimpli 2000). The model included the following predictors: the two-level factor group (L1 vs. 2L1), the two-level factor gender match (match vs. mismatch), the two-level factor grammaticality (grammatical vs. ungrammatical), the four-level factor segment (precritical, critical, postcritical and final) and the continuous covariate sentence repetition. As in the model with the effect of verbal intelligence, the reaction times were subject to negative reciprocal transformation, and the continuous variable was centred to the mean of both groups. Details about contrast coding, full specification of the initial model, the model selection procedure, the final model and its full output can be found in Section 2 and Table A12 of Appendix E. The analysis showed a main effect of segment (postcritical to final) ($t = 9.65, p < 0.001$) and an interaction grammaticality by gender match by segment (postcritical–final) ($t = -2.25, p < 0.05$). Therefore, there was not any effect of sentence repetition.

Finally, we addressed also for this task, the fact that for 4 out of 14 2L1 children, the dominant language was not MG. Figure 6 presents the RTs of the 2L1 group with MG as the dominant language. Comparing this figure with Figure 5, it is obvious that the pattern is the same with the difference that the error bars of the confidence intervals are larger. We

performed the same checks about a possible effect of dominance on RTs as we did with the accuracy scores in production.

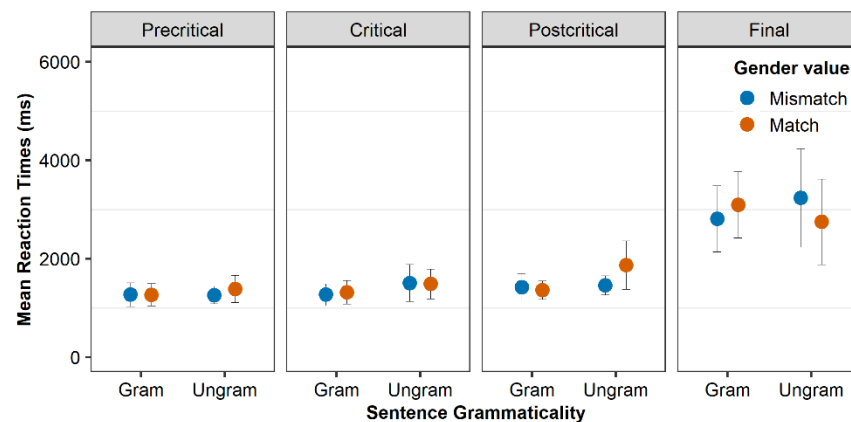


Figure 6. RTs of the 2L1 group excluding the children for which MG was not the dominant language per segment.

The first step was to insert in the models the categorical variable dominance. Recall that by doing this, all children, except for the four 2L1 for whom MG was not the dominant language, i.e., even the monolingual were considered as MG-dominant. We fitted the same models for verbal intelligence and sentence repetition, respectively, adding dominance as a predictor. In neither model was the effect of dominance significant. The second check was to analyse only the bilingual data with identical models, as the ones we reported, substituting the predictor group by dominance. For both covariates (verbal intelligence and sentence repetition), there was not a main effect of dominance, but dominance participated in a significant three-way interaction grammaticality by dominance by segment (precritical–critical). Since this interaction concerns the precritical segment, we do not consider it informative for clitic processing. All models are to be found in Section 3 and Tables A13–A16 of Appendix E.

6. Discussion

This study addressed the effect of gender marking on clitic production and comprehension by simultaneous bilingual children. As regards production, we found that 2L1 children performed similarly to L1 children in terms of correct responses as well as in terms of the distribution of error types. In particular, we found better performance in the elicitation of sentences with nouns of the same gender (gender match), in accordance with our predictions. Besides this, in both groups and in both conditions, the predominant error type was the production of wrong gender. We also found an interaction gender match by group by verbal intelligence, which suggests that verbal intelligence had an effect on the performance of L1 children in the more accurate condition (gender match), contrary to our predictions that such an effect would arise in the 2L1 group. The sentence repetition score did not have an effect and did not interact with any other factor. A post hoc check of the role of input and/or language dominance indicated that the performance of the 2L1 children does not correlate with any input measurement and dominance as a categorical variable did not have an effect on accuracy. Finally, the analysis of the errors suggested that sentences with neutral subject noun and/or neutral clitic are difficult for both groups. Only sentences with neutral subject and feminine clitic were quite easy for the 2L1 group. Moreover, L1 children had substantially higher performance in sentences with masculine subject and feminine clitic than the 2L1 children. In the comprehension task, we found similar RT patterns for the two groups, and in particular, a three-way interaction grammaticality by gender match by segment. However, the two groups differed in terms of overall reaction times, with longer RTs for the 2L1 group, as the group effect suggests. Verbal intelligence and sentence repetition did not have an effect or participate in any

interaction in the comprehension task. Post hoc investigation of an effect of dominance did not reveal any effect or interaction in the segments of interest. In the following, we discuss these findings and their implication for bilingual but also for monolingual acquisition.

Concerning our first research question, in our production task, 2L1 and L1 children performed more accurately in the gender match condition, in which there is not any noun with a competing gender value which can interfere in gender assignment, in accordance with our prediction. This finding indicates the leading role of gender marking in establishing the link between the clitic and its antecedent. However, gender errors are the predominant erroneous response also in this condition, and they outnumber number and case errors (which are reported under the category “wrong clitic” in our study). This provides further evidence that gender marking is more difficult than case and number marking in clitic production, a finding similar to that of [Rossi et al. \(2014\)](#) for clitic processing. Following [Rossi et al. \(2014\)](#) among others, we suggest that gender marking is more demanding, as it is related to the lexical representation of the antecedent, which has to be activated, whereas case and number apply only to morphology.

Remarkably, not only were gender errors the predominant error type, as was predicted on the basis of previous literature, but also the percentage of gender errors in the gender mismatch condition is around 20%, for both groups, which is similar to that found by [Varlokosta et al. \(2014\)](#) for L1 children of this age. Crucially, the percentage of gender errors drops to 8% in the gender match condition. Therefore, the effect of gender match found in the production task can be considered as evidence that clitics may not be difficult per se, rather that the high omission rates in some languages or erroneous responses found in other studies in MG are due to the difficulties with gender marking. Thus, the present study contributes towards identifying the specific source of difficulties with clitics, going beyond more general assumptions, which refer to difficulties with processing discourse related elements ([Avrutin 2006](#)). In particular, [Avrutin \(2006\)](#) has proposed that there are two different domains: the narrow and the discourse syntax, with the former dealing with dependencies within the sentence and the latter pertaining to information structure and being computationally more demanding and vulnerable (for example in cases of language disorders). Our findings indicate that difficulties are not only related to whether an element is discourse linked or not, but also to the quality of the features, for which the element is marked for.

Besides the error pattern, we have formulated two further predictions concerning production and in particular omission rates (see [Table 3](#), general predictions about production). The low omission rates for the L1 children replicate all previous findings, which have indicated very early acquisition of clitics in terms of production rates in obligatory contexts in MG ([Marinis 2000](#); [Tsakali and Wexler 2004](#); [Varlokosta et al. 2016](#)). The low omission rates of the 2L1 group across conditions appear to be in contradiction to previous studies on clitic production by bilingual children acquiring MG. This difference between our findings and previous studies can be explained by the different bilingual profiles of the participants. While the participants in the present study have acquired MG in a natural context, the children tested by [Chondrogianni \(2007\)](#), [Chondrogianni \(2008\)](#) and [Tsimpli and Mastropavlou \(2008\)](#) were successive bilinguals and additionally acquired MG in the context of language instruction.

Nevertheless, we had predicted that since older simultaneous bilingual children present a deviant pattern of attributing reference to clitics in comparison to L1 children ([Andreou et al. 2015](#)), deviant performance in terms of omission rates in younger simultaneous bilingual children is not excluded. In particular, we hypothesized that, since clitics production and processing concern the interface between syntax and pragmatics/discourse, it is possible that difficulties in an earlier stage of acquisition manifest in omissions, but later, omission rates drop, and difficulties manifest in deviant patterns in reference attribution. This prediction was not born out. We suggest that a task tapping on reference strategies for maintenance of a character in discourse, such as the one used in the study of [Andreou et al. \(2015\)](#), requires higher cognitive demands than inserting a clitic in obligatory contexts in a

structured elicitation task. Our results are, thus, in the same line with the ones of [Andreou et al. \(2015\)](#).

Turning to our second research question, which concerns the effect of gender marking on clitics processing, we found an effect of gender match, however, modulated by grammaticality and segment. The post hoc analysis of the interaction indicated that the difference between the postcritical and the final segment in ungrammatical sentences with gender match was not significant, contrary to all other conditions, which implies that in this condition, processing of the clitic was more challenging compared to all other conditions. Although we did not find any difference among the conditions in any segment, our findings point to differences in incremental sentence processing of the sentences. In order to understand the implications of this finding, we first consider our material and in particular an example of ungrammatical gender match (6) and ungrammatical gender mismatch sentences (7):

- | | | | | | | |
|-----|--------------------------|------------|-------------------------------|----------------------|---------|------------------|
| (6) | *O vatrachos | pezi ke/ | o kokoras/ | TIN/ | vafi/ | me ta chromata. |
| | The frog _{MASC} | plays and/ | the rooster _{MASC} / | her _{FEM} / | paints/ | with the colors. |
| (7) | *O vatrachos | pezi ke/ | i katsika/ | TIN/ | vafi/ | me ta chromata. |
| | The frog _{MASC} | plays and/ | the goat _{FEM} / | her _{FEM} / | paints/ | with the colors. |

In (7) (ungrammatical gender mismatch) there is a potential antecedent on the basis of gender cue, even if it is illicit in a strictly structural analysis, i.e., the local noun cannot be interpreted as the antecedent of the clitic. In (6) (ungrammatical gender match), the condition with the longest RTs in the postcritical segment, both preceding nouns have a different gender than the gender of the clitic, and thus, the clitic cannot be integrated/interpreted by any means in the sentence analysis. The interaction gender match by grammaticality by segment arises because the slow-down for the ungrammatical gender match condition already starts in the postcritical segment, whereas, for all other conditions, the slow-down only emerges in the final segment. We suggest that the slow-down in the postcritical segment is more likely to reflect processing difficulties with respect to integrating the clitic in the ungrammatical gender match condition, whereas the later emerging slow-down across conditions in the final segment may rather reflect a wrap-up effect.

This finding corroborates the results of other studies on language processing by children, which point to the crucial role of gender in processing of pronouns. In particular, children have been shown not to always interpret sentences in a strictly structural way, as they consider as candidate antecedent a structurally non-accessible noun. For instance, [Clackson et al. \(2011\)](#) found that 6–9 years old English-speaking children failed to correctly interpret a pronoun more often when there was a structurally inaccessible competitor noun with the same gender as the correct antecedent than when the competitor noun was of another gender. The eye-tracking data in the study of [Clackson et al. \(2011\)](#) also indicated that children were temporarily more distracted than adults in the interpretation of reflexives when a competitor with the same gender as the antecedent was present, although their off-line interpretation was adult-like. Moreover, a similar pattern is also observed in adult language processing. [Patil et al. \(2016\)](#) report the usage of gender as a cue for interpreting reflexive pronouns in English, which led to incorrect interpretation. Similarly, in our study the ungrammatical gender mismatch sentences contained one competitor noun which could be considered as the antecedent on the basis of its gender features, and the ungrammaticality might not have been detected. Unfortunately, we do not have off-line data from all items of our comprehension task, which might shed light on whether the incorrect use of cues in processing is mapped on the final result of the parsing process. It is remarkable, however, that such an interference of gender does not appear in the grammatical sentences in our data. Concerning our research question, whether gender marking affects clitic processing by L2 children, we suggest that this is the case, since gender match affected the processing of ungrammatical sentences, but the question needs further exploration, as the gender match effect was not found in the grammatical sentences.

Concerning the differences between the two groups in the comprehension task, three different scenarios have been put forward (see predictions 2, 3 and 4 for the second research question in Table 3): (1) identical patterns, (2) identical patterns with longer RTs for the postcritical segment for the bilingual group or (3) a grammaticality by group interaction. Our data show that there was a general group effect but not specific to a segment. Crucially, there was not any grammaticality by group interaction, which indicates that there is not a bilingual-specific processing of gender in object clitics, at least for simultaneous bilinguals. Our findings support the prediction that L1 and 2L1 children manifest the same pattern and are in accordance with the findings of Chondrogianni et al. (2015) concerning the similarity of the pattern between L1 and 2L1 and the overall longer RTs for the 2L1 group. Importantly, the pattern similarities arose in both our study and the study of Chondrogianni et al. despite the different language profiles in the two studies (simultaneous vs. successive). The differences in the time windows of the manifestation of the grammaticality effect are probably due to the different focus in each study, as in the study of Chondrogianni et al. (2015), the violation concerned the absence vs. presence of a clitic in an obligatory context (the critical segment), whereas in our material, the clitic was present and its features had to be processed in order for the participants to identify the correct antecedent.

Concerning the third research question, which pertained to the correlation between performance in production and proficiency (as measured with a test of verbal intelligence and a sentence repetition test) we found an interaction of verbal intelligence by gender match but only for the L1 children and no effect of the syntactic production abilities as measured with the sentence repetition test for either group. This is contradictory to studies which found a relation of proficiency in MG with the performance in clitic production for successive but also simultaneous bilingual children (Andreou et al. 2015; Chondrogianni 2007, 2008). Focusing on simultaneous bilingual children, recall that Andreou et al. (2015) found a correlation between vocabulary in MG and the native-like application of reference strategies in discourse. Our verbal intelligence task taps on vocabulary along with grammatical and pragmatic abilities; thus, an effect of verbal intelligence on clitic production is expected. As mentioned already, it is possible that the discrepancy between our study and that of Andreou et al. (2015) is due to the differences in task demands between the two studies. Independently of this, one should account for the effect of verbal intelligence on the performance of the L1 group only in the gender match condition. Our suggestion is that the interaction gender match by group by verbal intelligence reflects the difficulty of the gender mismatch condition even for L1 children with high verbal intelligence. In other words, sentences with gender mismatch are so challenging that high verbal intelligence does not improve performance.

It is remarkable that the sentence repetition task did not have any effect or interacted with any other factor in the production task in neither group, although it is a measure of syntactic production abilities. By contrast, the Verbal Intelligence Test we used subsumes vocabulary, grammatical and pragmatic abilities, as already mentioned. The fact that only the latter modulated the targeted production of clitics (even if this finding applies only to the L1 children) indicates that clitic production is a complex process which pertains to more than one linguistic domain. Hence, the composite score revealed to be more sensitive to explain accuracy than the syntax-only measure.

We would like to add a last note on the absence of an effect of dominance. While significant correlations were not obtained for any measure of input, there is a clear tendency for significance for the correlation between input in the first three years and performance in the match condition. The lack of significance might be due to the small sample size. Even as a tendency, this finding evinces the crucial role of the input in the first three years of life, which is considered to be the critical period for native language learning, and adds to previous studies, which report a correlation with early literacy (Andreou et al. 2015). The tendency for correlation was obtained for the most accurate condition. We point out that this pattern is remarkably similar to the effect of verbal intelligence on the performance of the L1 children in the match condition. Taken together, these results suggest that gender

marking of clitics in sentences with a subject noun with a different gender is challenging even for L1 children with high language abilities and 2L1 children with high input in the first three years. One could speculate that the input in the first three years and the verbal intelligence score are related in some way, and maybe they can be reduced to one latent factor. This is an intriguing question for further research with a larger sample.

Our overarching question was whether there are differences between the L1 and the 2L1 groups. Concerning production, the performance of the two groups is identical in terms of correct responses and error distribution. Nevertheless, we found a crucial difference concerning the contribution of verbal intelligence to the performance of each group. This finding deserves further scrutiny in a larger group of participants, but for the time being, we can argue that there are subtle differences in the production of clitics in MG between 2L1 children and L1 children. This suggestion is corroborated by the comprehension data, which point to an identical processing pattern, with 2L1 children being, nevertheless, slower than the L1 group. Last but not least, it remains an open question how input in the first three years of life and verbal intelligence are related to each other and in which ways they contribute to the production of clitics and gender marking and in general in processing of long-distance dependencies.

7. Conclusions

Summing up, this study contributes to the investigation of bilingual acquisition and in particular to the acquisition of clitics in MG in innovative ways. First, the data suggest that omission of clitics in obligatory contexts might not be due to a general difficulty with these complex elements, whose nature and interpretation are at the interface of morphology, syntax and pragmatics. Rather, we argue that clitic omission as was found in previous studies may be related to gender marking, which is relevant for both production and comprehension. In particular, the present data indicate that difficulties are not only related to whether an element is discourse linked or not or whether it is a phenomenon at the interface of multiple linguistic levels, but also to the quality of the features, for which the element is marked for. Importantly, this also holds for L1 children acquiring MG, a language in which clitics are considered to be acquired very early. Second, the data reveal many similarities but also crucial differences in the performance of simultaneous bilingual children compared to L1 children. Since our 2L1 participants are simultaneous bilinguals, the similarities in the correct performance and error distribution in production as well as the same pattern of RTs in comprehension are far from surprising. Nevertheless, the conclusion that these 2L1 children have acquired native-like and age-appropriate abilities in production is contradicted by the finding that verbal intelligence contributed to the performance of the L1 but not of the 2L1 children in the gender match condition. This finding implies that performance of L1 and 2L1 children is underpinned by different language abilities, a finding which deserves further scrutiny and needs to be explored in conjunction with the effect of input. Similarly, the data suggest that there are subtle differences between L1 and 2L1 concerning processing, as well: although the two groups manifested the same pattern, the 2L1 children were overall slower. All in all, the data suggest that the developmental path of 2L1 children is slightly different from that of L1 children, although their performance is for the most part identical.

Author Contributions: All four authors V.K., S.S., M.V. and F.A. are fully responsible for all parts of the text and the analyses. M.V. collected the data and carried out the transcription and the scoring. All authors have read and agreed to the published version of the manuscript.

Funding: The publication of this article was funded by Freie Universität Berlin.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and was approved by the Ethics Committee at the University of Potsdam (41/2017 on 21 December 2017).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author, V.K., upon reasonable request.

Acknowledgments: We are grateful to all the participants and their families, and to the German School of Athens for their willingness to contribute to this study. Many thanks go to Georgia Athanasopoulou and Yair Haendler for their consultation in data analysis, to Marilena Tsopanidi for the auditory stimuli and to Giacomo Marzona for the construction of the visual stimuli.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Items 12–20

(12) Who is speaking in what language to the child, and what language does the child use when he/she is speaking to that person? Please tick in the appropriate box.

(13) How well do the following persons speak and understand *Greek*? Please tick in the appropriate box.

(14) How well do the following persons speak and understand **German**? Please tick in the appropriate box.

(15) Please think about a typical day of the week in the life of your child. Who is your child spending time with? Please mark below the times spent with each person.

(16) Please think about a typical weekend in the life of your child. Who is your child spending time with? Please mark below the times spent with each person.

(17) How many **hours per week** (approximately) is your child spending time with each of the following activities? Please tell us which **language(s)** your child is speaking during the activities.

(18) Which languages has your child heard and spoken within the **first three years** of age? Please tick in the appropriate box.

(19) Which languages has your child heard and spoken between the ages of three and six? Please tick in the appropriate box.

(20) Which language has your child heard and spoken **from the age of six on**? Please tick in the appropriate box.

Appendix B

Procedure for examining the overlap between grammatical and natural gender:

We investigated whether there is an overlap between grammatical and natural gender by searching: (1) the occurrences of the concordance of each feminine or masculine noun with the adjective *arsenikos* (=masculine) and *thilikos* (=feminine) and (2) whether there exists a separate lemma denoting the feminine animal for the masculine nouns or the masculine animal for the feminine nouns along with its frequency (see Table A1 below). The search was performed in Google search engine on 7/11/2021. For two of these nouns (*kokoras* = rooster and *katsika* = goat), the grammatical gender overlaps with the natural gender, as the natural gender feature is inherent in the meaning of the noun. For the masculine noun *gaidaros* (=donkey), the meaning of a masculine animal was far more frequent, and there is also a frequent, separate lemma with the meaning feminine donkey. However, [Spathas and Sudo \(2020\)](#) suggest that *gaidaros* is a default masculine gender noun, in the sense that it “is not related with a gender-related association” ([Spathas and Sudo 2020](#), p. 34). Otherwise, the concordance of all other nouns with the feminine and masculine adjectives was comparable, which means that for all but two of the nouns, there was not any overlap between grammatical and natural gender.

Table A1. Occurrences of the concordance of each feminine or masculine noun with the adjective arsenikos (=masculine) and thilikos (=feminine) and the frequency of a separate lemma denoting the feminine animal for the masculine nouns or the masculine animal for the feminine nouns along with its frequency.

Noun	Meaning	Frequency with the Adjective Masculine	Frequency with the Adjective Feminine	Lemma with Meaning Masculine Animal	Frequency	Lemma with Meaning Feminine Animal	Frequency
<i>gaidaros</i>	<i>donkey</i>	1.670	846	-	-	<i>gaidura</i>	5.760
<i>kokoras</i>	<i>rooster</i>	1.010	747 (only in the context of Chinese horoscope)	-	-	<i>kota</i>	Not relevant because rooster has a gender-related association
<i>panthiras</i>	<i>panther</i>	218	302	-	-	<i>panthirina</i>	Refers to a specific person
<i>vatrachos</i>	<i>frog</i>	539	638	-	-	<i>vatrachina</i>	7.510
<i>chelona</i>	<i>turtle</i>	3.940	3.890	<i>chelonos</i>	2.390 in children books	-	-
<i>agelada</i>	<i>cow</i>	755 almost always in definition of <i>tavros</i> (=bull)	866	-	-	-	-
<i>melisa</i>	<i>bee</i>	1.740 almost always in definitions	1.250	<i>kifinas</i>	396.000	-	-
<i>katsika</i>	<i>goat</i>	801 almost always in definitions	627	<i>tragos</i>	473.000	-	-

Appendix C

Statistical analysis of the production task

1. Models for accuracy and the effect of verbal intelligence in the production task

The model motivated by our predictions (see model a) contained the following predictors: the two-level factor group (L1 vs. 2L1), the two-level factor gender match (match vs. mismatch) and the continuous variable verbal intelligence. The contrasts of the two categorical factors were set up with sliding contrast coding, which means that the intercept is the grand mean of all conditions. In sliding contrasts coding, the main effects refer to the difference between one level and its adjacent level. In this case, because the factors have only two levels, the main effect of group refers to the difference between the two groups, and the main effect of gender match refers to the difference between sentences with gender match and sentences with gender mismatch. The score in verbal intelligence was centred to the mean of both groups, as they were matched for APT, which means that the main effect refers to the mean value of the groups. Therefore, the fixed effects of the initial model were the two factors gender match and group, which were modelled together with the continuous covariate verbal intelligence and their interaction. In addition to these fixed effects, the following random components were specified in the initial model: an adjustment of each child’s individual average (random intercept for subjects), an adjustment for each child’s individual gender match effect (random slope for subjects) and an adjustment of the average of each item (random intercept for items).

(a) Initial model motivated by the research hypothesis:
`modelAPT_cent <-glmer(accuracy ~ condition*subtype*APT.cent +
 (1+condition | subject) +
 (1 | itemno), data=d_TD,
 family=binomial,
 control=glmerControl(optimizer="bobyqa"))`

Model selection procedure: After running the initial model, we checked for singularity (check was negative) and tried to identify the number of dimensions of the random effects for subjects supported by the data by performing a principal component analysis on the model, following [Bates et al. \(2018\)](#). Our test showed that only one dimension can cumulatively account for 100% of the variance. Subsequently, we dropped the random slope for subjects. This decision was supported by the computation of the correlation between the random intercept and random slope for subjects, which was 1. The formula of the final model is in (b).

(b) Model supported by the data
`modelAPtb_cent <-glmer(accuracy ~ condition*subtype*APT.cent +
 (1 | subject) +
 (1 | itemno), data=d_TD,
 family=binomial,
 control=glmerControl(optimizer="bobyqa"))`

Table A2. Summary of the model for accuracy and the effect of verbal intelligence in the production task.

Predictors	Estimates	std. Error	z Value	p
Intercept	1.465	0.19	7.53	<0.001
Gender Match	−0.751	0.36	−2.07	0.038
Group	−0.055	0.23	−0.24	0.807
Verbal Intelligence	0.004	0.01	0.30	0.763
Gender Match*Group	0.077	0.39	0.20	0.842
Gender Match*Verbal Intelligence	0.003	0.02	0.14	0.890
Group*Verbal Intelligence	−0.015	0.03	−0.57	0.567
Gender Match*Group*Verbal Intelligence	0.102	0.05	2.22	0.026

2. Models for accuracy and the effect of sentence repetition scores in the production task

For the analysis, the score in the sentence repetition task was centred to the mean of both groups, which means that the main effect refers to the mean value of the groups. The contrasts of the categorical variables were coded as sliding contrasts. The initial model (see a) included the two factors gender match and group with the covariate sentence repetition and their interaction as fixed effects. The following random components were specified in the initial model: an adjustment of each child’s individual average (random intercept for subjects), an adjustment for each child’s individual gender match effect (random slope for subjects) and an adjustment of the average of each item (random intercept for items). We ran the same checks as in the model for verbal intelligence and the results showed that the model was supported by the data. The full model output is in Table A3 below.

```
(a) Model of the whole dataset
modelSRT_cent <-glmer(accuracy ~ condition*subtype*SRT.cent +
(1+condition | subject) +
(1 | itemno), data=d_TD,
family=binomial,
control=glmerControl(optimizer="bobyqa"))
```

Table A3. Summary of the final model for accuracy and the effect of sentence repetition in the production task.

Predictors	Estimates	std. Error	z Value	p
Intercept	1.46	0.19	7.77	<0.001
Gender Match	−0.64	0.37	−1.71	0.087
Group	0.01	0.20	0.07	0.946
Sentence Repetition	0.05	0.03	1.79	0.074
Gender Match*Group	0.06	0.40	0.16	0.874
Gender Match*Sentence Repetition	−0.06	0.05	−1.05	0.293
Group*Sentence Repetition	0.04	0.05	0.68	0.499
Gender Match*Group*Sentence Repetition	0.21	0.11	1.93	0.054

```
(b) Model after excluding the outlier
modelSRT_cent_Out <-glmer(accuracy ~ condition*subtype*SRT.cent +
(1+condition | subject) +
(1 | itemno), data=d_TD_Outl,
family=binomial,
control=glmerControl(optimizer="bobyqa"))
```

Table A4. Summary of the model for accuracy and the effect of sentence repetition in the production task excluding the outlier.

Predictors	Estimates	std. Error	z Value	p
Intercept	4.41	0.20	7.54	<0.001
Gender Match	0.55	0.38	−1.57	0.117
Group	1.02	0.21	0.11	0.910
Sentence Repetition	1.05	0.03	1.33	0.182
Gender Match*Group	1.16	0.41	0.37	0.710
Gender Match*Sentence Repetition	0.90	0.07	−1.53	0.126
Group*Sentence Repetition	1.03	0.07	0.42	0.673
Gender Match*Group*Sentence Repetition	1.13	0.13	0.95	0.343

3. Models for the effect of language dominance on accuracy

- (a) Model with the whole dataset and verbal intelligence and dominance as predictors
`modelAPTb_Dom <-glmer(accuracy ~ condition*subtype*APT.cent+Dominance + (1 | subject) + (1 | itemno), data=d_TD, family=binomial, control=glmerControl(optimizer="bobyqa"))`

Table A5. Summary of the model for accuracy and the effect of verbal intelligence and language dominance in the production task for the whole dataset.

Predictors	Estimates	std. Error	z Value	p
Intercept	2.1093	0.3729	5.66	<0.001
Gender Match	-0.7890	0.4062	-1.94	0.052
Group	-0.2039	0.3313	-0.62	0.538
Verbal Intelligence	0.0355	0.0225	1.58	0.115
Dominance	-0.4494	0.4008	-1.12	0.262
Gender Match*Group	0.0735	0.3891	0.19	0.850
Gender Match*Verbal Intelligence	-0.0478	0.0285	-1.68	0.093
Group*Verbal Intelligence	-0.0787	0.0385	-2.04	0.041
Gender Match*Group*Verbal Intelligence	0.1041	0.0464	2.24	0.025

- (b) Model with the bilingual children and verbal intelligence and dominance as predictor
`modelAPT_bil_Dom <-glmer(accuracy ~ condition*Dominance*APT.cent + (1 | subject) + (1 | itemno), data=d_TD_L2, family=binomial, control=glmerControl(optimizer="bobyqa"))`

Table A6. Summary of the model for accuracy and the effect of verbal intelligence and language dominance in the production task for the bilingual children.

Predictors	Estimates	std. Error	z Value	p
Intercept	1.5137	0.3286	4.61	<0.001
Gender Match	-0.6537	0.4781	-1.37	0.172
Dominance	-0.3641	0.5621	-0.65	0.517
Verbal Intelligence	-0.0216	0.0339	-0.64	0.524
Gender Match*Dominance	0.4575	0.7194	-0.64	0.525
Gender Match*Verbal Intelligence	0.0396	0.0437	0.91	0.364
Dominance*Verbal Intelligence	-0.0574	0.0679	-0.85	0.398
Gender Match*Dominance*Verbal Intelligence	-0.0899	0.0873	-1.03	0.303

- (c) Model with the whole dataset and sentence repetition and dominance as predictor
`modelSRT_Dom <-glmer(accuracy ~ condition*subtype*SRT.cent + Dominance+ (1+condition | subject) + (1 | itemno), data=d_TD, family=binomial, control=glmerControl(optimizer="bobyqa"))`
- (d) Model with the bilingual children and sentence repetition and dominance as predictor
`modelSRT_B_Dom <-glmer(accuracy ~ condition*Dominance*SRT.cent + (1+condition | subject) + (1 | itemno), data=d_TD_L2, family=binomial, control=glmerControl(optimizer="bobyqa"))`

Table A7. Summary of the model for accuracy and the effect of sentence repetition and language dominance in the production task for the whole dataset.

Predictors	Estimates	std. Error	z Value	p
Intercept	1.8668	0.3562	5.24	<0.001
Gender Match	−0.6705	0.4144	−1.62	0.106
Group	−0.0621	0.322	−0.19	0.847
Sentence Repetition	0.1118	0.0744	1.5	0.133
Dominance	−0.1496	0.3316	−0.45	0.652
Gender Match*Group	0.0642	0.4019	0.16	0.873
Gender Match*Sentence Repetition	−0.163	0.0987	−1.65	0.099
Group*Sentence Repetition	−0.0716	0.0828	−0.86	0.387
Gender Match*Group*Sentence Repetition	0.2107	0.1095	1.92	0.054

Table A8. Summary of the model for accuracy and the effect of sentence repetition and language dominance in the production task for the bilingual children.

Predictors	Estimate	std. Error	z Value	p
Intercept	1.6142	0.3275	4.93	<0.001
Gender Match	−0.8771	0.5342	−1.64	0.101
Dominance	−0.2693	0.5529	−0.49	0.626
Sentence Repetition	0.0381	0.0791	0.48	0.63
Gender Match*Dominance	0.0769	0.8343	0.09	0.927
Gender Match* Sentence Repetition	0.2369	0.12	1.97	0.048
Dominance*Sentence Repetition	0.0697	0.1582	0.44	0.66
Gender Match*Dominance*Sentence Repetition	−0.4353	0.2399	−1.81	0.07

Appendix D

Table A9. Mean raw reaction times and standard deviation for all four sentence types in the precritical, critical, postcritical and final segment for the L1 group, Match: gender match, Mismatch: gender mismatch, Gram: grammatical, Ungram: ungrammatical, RT: reaction times.

Segment	Gender Match	Grammaticality	Mean RTs	Standard Deviation RTs
Precritical	Mismatch	Gram	1029.4	196.3
		Ungram	1197.0	361.1
	Match	Gram	1088.6	212.9
		Ungram	1386.3	406.9
Critical	Mismatch	Gram	1185.8	338.2
		Ungram	1199.5	367.8
	Match	Gram	1211.5	288.3
		Ungram	1386.0	448.7
Postcritical	Mismatch	Gram	1223.2	321.1
		Ungram	1536.9	626.3
	Match	Gram	1265.7	331.9
		Ungram	1952.1	1138.8
Final	Mismatch	Gram	2411.5	1011.6
		Ungram	2622.6	1078.6
	Match	Gram	2449.8	912.4
		Ungram	2384.1	796.7

Table A10. Mean raw reaction times and standard deviation for all four sentence types in the precritical, critical, postcritical and final segment for the 2L1 group and in parentheses for the subset of the 2L1 group with MG as dominant language, Match: gender match, Mismatch: gender mismatch, Gram: grammatical, Ungram: ungrammatical, RT: reaction times.

Segment	Gender Match	Grammaticality	Mean RTs	Standard Deviation RTs
Precritical	Mismatch	Gram	1237.7 (1269.7)	317.9 (300.4)
		Ungram	1214.1 (1259.0)	242.9 (268.1)
	Match	Gram	1242.4 (1267.1)	192.4 (136.8)
		Ungram	1388.0 (1383.4)	289.2 (337.7)
Critical	Mismatch	Gram	1230.6 (1269.5)	162.3 (145.5)
		Ungram	1388.2 (1507.6)	452.3 (488.7)
	Match	Gram	1342.6 (1315.6)	210.5 (239.1)
		Ungram	1404.5 (1487.1)	442.3 (503.2)
Postcritical	Mismatch	Gram	1376.6 (1425.8)	239.6 (247.4)
		Ungram	1411.2 (1458.0)	339.3 (379.8)
	Match	Gram	1383.9 (1362.3)	166.8 (148.9)
		Ungram	1801.1 (1870.8)	757.5 (836.0)
Final	Mismatch	Gram	2592.2 (2810.4)	973.6 (1071.5)
		Ungram	3032.9 (3232.8)	1318.9 (1519.1)
	Match	Gram	3138.1 (3097.1)	1092.0 (1102.8)
		Ungram	2551.2 (2749.7)	1196.8 (1375.2)

Appendix E

Statistical analysis of the comprehension task

1. Models for the reaction times and the effect of verbal intelligence in the comprehension task

The following fixed effects were specified in the initial model, motivated by the research questions and the predictions: (1) group, (2) gender match, (3) grammaticality, (4) segment, (5) their four-way interaction, (6) the interaction group by verbal intelligence (motivated by the interaction found in production) and (7) the interaction grammaticality by verbal intelligence, motivated by the hypothesis that the higher the proficiency the more the sensitivity to grammatical violations. The contrasts of all categorical variables were set up with sliding contrast coding, which means that the intercept was the mean of the two groups across conditions and segments. Moreover, the main effect of group refers to the difference between the two groups, the main effect of gender match refers to the difference between sentences with gender match and sentences with gender mismatch, and the main effect of grammaticality refers to the difference between grammatical and ungrammatical sentences. Concerning the four-level factor segment, the first main effect refers to the

difference between the precritical and the critical segment, the second main effect refers to the difference between the critical and the postcritical and the third main effect refers to the difference between postcritical and final segment. The score in verbal intelligence was centred to the mean of both groups, which means that the main effect refers to the mean value of the groups. The following random components were specified: (1) an adjustment of each child's individual average (random intercept for subjects), (2) an adjustment for each child's individual grammaticality effect, gender match effect and region effect and their interaction, (3) an adjustment of the average of each item (random intercept for items) and (4) an adjustment for each item's individual grammaticality effect, since all items were used for both grammatical and ungrammatical sentences (random slopes for items).

- (a) Initial model motivated by the research hypothesis
- ```
modelAPTGroupGramm_cent <- lmer(rt.r ~ gramm*gender_match*subt*region +
APT.cent*subt + APT.cent*gramm +
(1+gramm*gender_match*region | subject) + (1+gramm | Picture),
data=stat1,
REML=F, lmerControl(optimizer = "bobyqa"))
```

**Model selection procedure:** This initial model did not converge, which suggests that it was too complex to be supported by the data. We successively simplified the structure of the random effects. We ran models in which we dropped first the random effect of grammaticality for items and then the random effect of gender match for subjects. Both models did not converge. Subsequently, we removed the random effect of region for subjects and the model converged. We checked for singularity (check was negative) and tried to identify the number of dimensions supported by the data by performing a principal component analysis on the model. The principal component analysis showed that both dimensions (intercept and slope) cumulatively account for 100% of the variance. Moreover, the correlation between intercept and slope was very low (0.04). Thus, we used this model for the analysis of the data.

- (b) Model supported by the data
- ```
modelAPTGroupGramm_cent <- lmer(rt.r ~ gramm*subt*gender_match*region +
APT.cent*subt + APT.cent*gramm +
(1+gramm | subject) + (1 | Picture),
data=stat1,
REML=F, lmerControl(optimizer = "bobyqa"))
```

Table A11. Summary of the model for the reaction times and the effect of verbal intelligence in the comprehension task.

Predictors	Estimates	std. Error	t Value	p
Intercept	−0.93	0.02	−40.23	<0.001
Grammaticality	0.04	0.03	1.45	0.148
Group	−0.10	0.04	−2.33	0.020
Gender Match	−0.00	0.03	−0.01	0.988
Segment Precritical–Critical	0.03	0.03	0.95	0.344
Segment Critical–Postcritical	−0.00	0.03	−0.01	0.991
Segment Postcritical–Final	0.27	0.03	9.66	<0.001
Verbal Intelligence	−0.01	0.00	−2.06	0.039
Grammaticality*Group	−0.03	0.05	−0.63	0.527
Grammaticality*Gender Match	−0.01	0.05	−0.14	0.891
Group*Gender Match	−0.01	0.04	−0.20	0.840
Grammaticality*Segment Precritical–Critical	−0.00	0.06	−0.03	0.975
Grammaticality*Segment Critical–Postcritical	0.03	0.06	0.60	0.549
Grammaticality*Segment Postcritical–Final	−0.06	0.06	−1.11	0.269
Group*Segment Precritical–Critical	0.01	0.06	0.22	0.824
Group*Segment Critical–Postcritical	−0.02	0.06	−0.29	0.770
Group*Segment Postcritical–Final	−0.03	0.06	−0.49	0.625
Gender Match*Segment Precritical–Critical	0.02	0.06	0.30	0.765
Gender Match*Segment Critical–Postcritical	−0.06	0.06	−1.11	0.268
Gender Match*Segment Postcritical–Final	−0.05	0.06	−0.80	0.421
Group*Verbal Intelligence	0.00	0.01	0.93	0.355
Grammaticality*Verbal Intelligence	0.00	0.00	0.65	0.518
Grammaticality*Group*Gender Match	−0.01	0.08	−0.11	0.912
Grammaticality*Group*Segment Precritical–Critical	0.04	0.11	0.37	0.712
Grammaticality*Group*Segment Critical–Postcritical	−0.07	0.11	−0.64	0.522
Grammaticality*Group*Segment Postcritical–Final	0.00	0.11	0.02	0.981
Grammaticality*Gender Match*Segment Precritical–Critical	−0.13	0.11	−1.13	0.260
Grammaticality*Gender Match*Segment Critical–Postcritical	0.10	0.11	0.91	0.361
Grammaticality*Gender Match*Segment Postcritical–Final	−0.25	0.11	−2.25	0.025
Group*Gender Match*Segment Precritical–Critical	−0.08	0.11	−0.74	0.459
Group*Gender Match*Segment Critical–Postcritical	0.08	0.11	0.71	0.478
Group*Gender Match*Segment Postcritical–Final	−0.17	0.11	−1.51	0.130
Grammaticality*Group*Gender Match*Segment Precritical–Critical	0.34	0.22	1.50	0.133
Grammaticality*Group*Gender Match*Segment Critical–Postcritical	−0.28	0.22	−1.25	0.211
Grammaticality*Group*Gender Match*Segment Postcritical–Final	0.19	0.23	0.85	0.396

2. Model for the reaction times and the effect of sentence repetition in the comprehension task

The contrast coding was identical with the model for verbal intelligence. The initial model we ran included the following fixed effects: (1) group, (2) gender match, (3) grammaticality, (4) segment, (5) their four-way interaction, (6) the interaction group by sentence repetition (parallel to the model of the effect of verbal intelligence) and (7) the interaction grammaticality by sentence repetition (motivated by the hypothesis that the higher the syntactic production abilities, the more the sensitivity to grammatical violations). The following random components were specified: (1) an adjustment of each child’s individual average (random intercept for subjects), (2) an adjustment for each child’s individual grammaticality effect, gender match effect, region effect and their interaction, (3) an adjustment of the average of each item (random intercept for items) and (4) an adjustment for each item’s individual grammaticality effect, since all items were used for both grammatical and ungrammatical sentences (random slopes for items).

```
(a) Initial model motivated by the hypothesis
modelSRTGroupGramm_cent <- lmer(rt.r ~ gramm*gender_match*subt*region +
SRT.cent*subt + SRT.cent*gramm +
(1+gramm*region*gender_match | subject) +
(1+gramm | Picture),
data=stat1,
REML=F, lmerControl(optimizer = "bobyqa"))
```

Model selection procedure: The initial model did not converge, which suggests that it was too complex to be supported by the data. We successively simplified the structure of the random effects for subjects, as before. First, we dropped the random slope of gender match for subjects. The model was singular, which suggests that the random effects structure is still too complex, and then, we removed the random slope for regions. We checked for singularity (check was negative) and used a principal component analysis on the model to identify the number of dimensions supported by the data. The analysis showed that both dimensions (intercept and slope) cumulatively account for 100% of the variance. Moreover, the correlation between intercept and slope was very low (0.30). Thus, we used this model for the analysis of the data.

```
(b) Model supported by the data
modelSRTGroupGramm_cent <- lmer(rt.r ~ gramm*gender_match*subt*region +
SRT.cent*subt + SRT.cent*gramm +
(1+gramm | subject) +
(1 | Picture),
data=stat1,
REML=F, lmerControl(optimizer = "bobyqa"))
```

Table A12. Summary of the model for the reaction times and the effect of sentence repetition in the comprehension task.

Predictors	Estimates	std. Error	t Value	p
Intercept	−0.93	0.02	−38.35	<0.001
Grammaticality	0.04	0.03	1.46	0.143
Group	−0.09	0.05	−1.95	0.051
Gender Match	−0.00	0.03	−0.02	0.983
Segment Precritical–Critical	0.03	0.03	0.95	0.344
Segment Critical–Postcritical	−0.00	0.03	−0.01	0.995
Segment Postcritical–Final	0.27	0.03	9.65	<0.001
Sentence Repetition	−0.00	0.01	−0.13	0.897
Grammaticality*Group	−0.03	0.05	−0.58	0.559

Table A12. *Cont.*

Predictors	Estimates	std. Error	t Value	p
Grammaticality*Gender Match	−0.01	0.05	−0.13	0.896
Group*Gender Match	−0.01	0.04	−0.22	0.827
Grammaticality*Segment	−0.00	0.06	−0.03	0.976
Precritical–Critical				
Grammaticality*Segment	0.03	0.06	0.60	0.546
Critical–Postcritical				
Grammaticality*Segment	−0.06	0.06	−1.10	0.270
Postcritical–Final				
Group*Segment Precritical–Critical	0.01	0.06	0.22	0.824
Group*Segment Critical–Postcritical	−0.02	0.06	−0.30	0.766
Group*Segment Postcritical–Final	−0.03	0.06	−0.48	0.631
Gender Match*Segment	0.02	0.06	0.30	0.765
Precritical–Critical				
Gender Match*Segment	−0.06	0.06	−1.10	0.269
Critical–Postcritical				
Gender Match*Segment	−0.05	0.06	−0.81	0.418
Postcritical–Final				
Group*Sentence Repetition	0.01	0.01	0.65	0.516
Grammaticality*Sentence Repetition	−0.01	0.01	−1.23	0.220
Grammaticality*Group*Gender Match	−0.01	0.08	−0.12	0.901
Grammaticality*Group*Segment	0.04	0.11	0.37	0.712
Precritical–Critical				
Grammaticality*Group*Segment	−0.07	0.11	−0.65	0.518
Critical–Postcritical				
Grammaticality*Group*Segment	0.00	0.11	0.02	0.982
Postcritical–Final				
Grammaticality*Gender Match*Segment	−0.13	0.11	−1.13	0.260
Precritical–Critical				
Grammaticality*Gender Match*Segment	0.10	0.11	0.92	0.359
Critical–Postcritical				
Grammaticality*Gender Match*Segment	−0.25	0.11	−2.25	0.025
Postcritical–Final				
Group*Gender Match*Segment	−0.08	0.11	−0.74	0.459
Precritical–Critical				
Group*Gender Match*Segment	0.08	0.11	0.71	0.479
Critical–Postcritical				
Group*Gender Match*Segment	−0.17	0.11	−1.51	0.132
Postcritical–Final				
Grammaticality*Group*Gender	0.34	0.22	1.50	0.133
Match*Segment Precritical–Critical				
Grammaticality*Group*Gender	−0.28	0.22	−1.25	0.210
Match*Segment Critical–Postcritical				
Grammaticality*Group*Gender	0.19	0.23	0.85	0.398
Match*Segment Postcritical–Final				

3. Models for the effect of language dominance on reaction times

- (a) Model with the whole dataset and verbal intelligence and dominance as predictors
`modelAPTGroupGramm_Dominance <- lmer(rt.r ~ gramm*subt*gender_match*region + APT.cent*subt + APT.cent*gramm + Dominance + (1+gramm | subject) + (1 | Picture), data=stat1, REML=F, lmerControl(optimizer = "bobyqa"))`

Table A13. Summary of the model with the whole dataset and verbal intelligence and dominance as predictors.

Predictors	Estimates	std. Error	t Value	p
Intercept	−0.97	0.07	−13.89	<0.001
Grammaticality	0.04	0.03	1.45	0.148
Group	−0.11	0.05	−2.38	0.017
Gender Match	0	0.03	−0.02	0.987
Segment Precritical–Critical	0.03	0.03	0.95	0.344
Segment Critical–Postcritical	0	0.03	−0.01	0.991
Segment Postcritical–Final	0.27	0.03	9.65	<0.001
Verbal Intelligence	0	0	−1.72	0.085
Dominance	0.05	0.08	0.61	0.544
Grammaticality*Group	−0.03	0.05	−0.63	0.526
Grammaticality*Gender Match	−0.01	0.05	−0.14	0.889
Group*Gender Match	−0.01	0.04	−0.2	0.839
Grammaticality*Segment Precritical–Critical	0	0.06	−0.03	0.975
Grammaticality*Segment Critical–Postcritical	0.03	0.06	0.6	0.549
Grammaticality*Segment Postcritical–Final	−0.06	0.06	−1.1	0.269
Group*Segment Precritical–Critical	0.01	0.06	0.22	0.824
Group*Segment Critical–Postcritical	−0.02	0.06	−0.29	0.77
Group*Segment Postcritical–Final	−0.03	0.06	−0.49	0.625
Gender Match*Segment Precritical–Critical	0.02	0.06	0.3	0.765
Gender Match*Segment Critical–Postcritical	−0.06	0.06	−1.11	0.268
Gender Match*Segment Postcritical–Final	−0.05	0.06	−0.81	0.421
Group*Verbal Intelligence	0	0.01	0.62	0.536
Grammaticality*Verbal Intelligence	0	0	0.65	0.518
Grammaticality*Group*Gender Match	−0.01	0.08	−0.11	0.912
Grammaticality*Group*Segment Precritical–Critical	0.04	0.11	0.37	0.712
Grammaticality*Group*Segment Critical–Postcritical	−0.07	0.11	−0.64	0.522
Grammaticality*Group*Segment Postcritical–Final	0	0.11	0.02	0.982
Grammaticality*Gender Match*Segment Precritical–Critical	−0.13	0.11	−1.13	0.26
Grammaticality*Gender Match*Segment Critical–Postcritical	0.1	0.11	0.91	0.361
Grammaticality*Gender Match*Segment Postcritical–Final	−0.25	0.11	−2.24	0.025
Group*Gender Match*Segment Precritical–Critical	−0.08	0.11	−0.74	0.459
Group*Gender Match*Segment Critical–Postcritical	0.08	0.11	0.71	0.478
Group*Gender Match*Segment Postcritical–Final	−0.17	0.11	−1.51	0.13
Grammaticality*Group*Gender Match*Segment Precritical–Critical	0.34	0.22	1.5	0.134
Grammaticality*Group*Gender Match*Segment Critical–Postcritical	−0.28	0.22	−1.25	0.211
Grammaticality*Group*Gender Match*Segment Postcritical–Final	0.19	0.23	0.85	0.397

```
(b) Model with the bilingual children and verbal intelligence and dominance as predictors
modelAPTGroupGramm_B_Dom<- lmer(rt.r ~ gramm*Dominance*gender_match*region
+ APT.cent*Dominance + APT.cent*gramm +
(1+gramm | subject) +
(1 | Picture),
data=stat1_L2,
REML=F, lmerControl(optimizer = "bobyqa"))
```

Table A14. Summary of the model with the bilingual children and verbal intelligence and dominance as predictors.

Predictors	Estimates	std. Error	t Value	p
Intercept	−0.92	0.07	−12.99	<0.001
Grammaticality	0	0.08	0	1
Dominance	0.05	0.08	0.58	0.563
Gender Match	0.01	0.07	0.09	0.928
Segment Precritical–Critical	0.11	0.09	1.23	0.221
Segment Critical–Postcritical	−0.01	0.09	−0.09	0.931
Segment Postcritical–Final	0.2	0.09	2.21	0.027
Verbal Intelligence	−0.01	0.01	−1.07	0.284
Grammaticality*Dominance	0.07	0.09	0.79	0.427
Grammaticality*Gender Match	−0.08	0.14	−0.58	0.564
Dominance*Gender Match	0	0.08	0.02	0.983
Grammaticality*Segment Precritical–Critical	−0.36	0.18	−1.96	0.05
Grammaticality*Segment Critical–Postcritical	0.2	0.18	1.08	0.28
Grammaticality*Segment Postcritical–Final	−0.14	0.18	−0.76	0.446
Dominance*Segment Precritical–Critical	−0.13	0.11	−1.2	0.23
Dominance*Segment Critical–Postcritical	0.02	0.11	0.2	0.842
Dominance*Segment Postcritical–Final	0.11	0.11	1.05	0.295
Gender Match*Segment Precritical–Critical	0.13	0.18	0.7	0.481
Gender Match*Segment Critical–Postcritical	0.02	0.18	0.11	0.911
Gender Match*Segment Postcritical–Final	−0.02	0.18	−0.12	0.903
Dominance*Verbal Intelligence	0	0.01	−0.07	0.943
Grammaticality*Verbal Intelligence	0.01	0.01	1.9	0.058
Grammaticality*Dominance*Gender Match	0.11	0.16	0.69	0.49
Grammaticality*Dominance*Segment Precritical–Critical	0.48	0.22	2.18	0.029
Grammaticality*Dominance*Segment Critical–Postcritical	−0.18	0.22	−0.83	0.409
Grammaticality*Dominance*Segment Postcritical–Final	0.11	0.22	0.49	0.626
Grammaticality*Gender Match*Segment Precritical–Critical	−0.75	0.37	−2.03	0.042
Grammaticality*Gender Match*Segment Critical–Postcritical	0.18	0.37	0.49	0.626
Grammaticality*Gender Match*Segment Postcritical–Final	−0.76	0.37	−2.07	0.039
Dominance*Gender Match*Segment Precritical–Critical	−0.1	0.22	−0.47	0.635

Table A14. *Cont.*

Predictors	Estimates	std. Error	t Value	p
Dominance*Gender Match*Segment Critical–Postcritical	−0.17	0.22	−0.78	0.433
Dominance*Gender Match*Segment Postcritical–Final	0.09	0.22	0.4	0.689
Grammaticality*Dominance*Gender Match*Segment Precritical–Critical	0.64	0.44	1.47	0.143
Grammaticality*Dominance*Gender Match*Segment Critical–Postcritical	0.09	0.44	0.2	0.84
Grammaticality*Dominance*Gender Match*Segment Postcritical–Final	0.58	0.44	1.32	0.188

(c) Model with the whole dataset and sentence repetition and dominance as predictors
`modelSRTGroupGramm_Dominance <- lmer(rt.r ~ gramm*subt*gender_match*region + SRT.cent*subt + SRT.cent*gramm + Dominance + (1+gramm | subject) + (1 | Picture), data=stat1, REML=F, lmerControl(optimizer = "bobyqa"))`

Table A15. Summary of the model with the bilingual children and sentence repetition and dominance as predictors.

Predictors	Estimates	std. Error	t Value	p
Intercept	−1	0.07	−15.16	<0.001
Grammaticality	0.04	0.03	1.46	0.143
Group	−0.11	0.05	−2.29	0.022
Gender Match	0	0.03	−0.02	0.981
Segment Precritical–Critical	0.03	0.03	0.95	0.344
Segment Critical–Postcritical	0	0.03	−0.01	0.995
Segment Postcritical–Final	0.27	0.03	9.65	<0.001
Sentence Repetition	0	0.01	−0.02	0.988
Dominance	0.08	0.07	1.13	0.258
Grammaticality*Group	−0.03	0.05	−0.59	0.558
Grammaticality*Gender Match	−0.01	0.05	−0.14	0.892
Group*Gender Match	−0.01	0.04	−0.21	0.83
Grammaticality*Segment Precritical–Critical	0	0.06	−0.03	0.976
Grammaticality*Segment Critical–Postcritical	0.03	0.06	0.6	0.546
Grammaticality*Segment Postcritical–Final	−0.06	0.06	−1.1	0.27
Group*Segment Precritical–Critical	0.01	0.06	0.22	0.824
Group*Segment Critical–Postcritical	−0.02	0.06	−0.3	0.766
Group*Segment Postcritical–Final	−0.03	0.06	−0.48	0.631
Gender Match*Segment Precritical–Critical	0.02	0.06	0.3	0.765
Gender Match*Segment Critical–Postcritical	−0.06	0.06	−1.11	0.269
Gender Match*Segment Postcritical–Final	−0.05	0.06	−0.81	0.418
Group*Sentence Repetition	0.01	0.01	0.53	0.594
Grammaticality*Sentence Repetition	−0.01	0.01	−1.23	0.22

Table A15. *Cont.*

Predictors	Estimates	std. Error	t Value	p
Grammaticality*Group*Gender Match	−0.01	0.08	−0.12	0.905
Grammaticality*Group*Segment Precritical–Critical	0.04	0.11	0.37	0.712
Grammaticality*Group*Segment Critical–Postcritical	−0.07	0.11	−0.65	0.518
Grammaticality*Group*Segment Postcritical–Final	0	0.11	0.02	0.982
Grammaticality*Gender Match*Segment Precritical–Critical	−0.13	0.11	−1.13	0.26
Grammaticality*Gender Match*Segment Critical–Postcritical	0.1	0.11	0.92	0.359
Grammaticality*Gender Match*Segment Postcritical–Final	−0.25	0.11	−2.25	0.025
Group*Gender Match*Segment Precritical–Critical	−0.08	0.11	−0.74	0.458
Group*Gender Match*Segment Critical–Postcritical	0.08	0.11	0.71	0.479
Group*Gender Match*Segment Postcritical–Final	−0.17	0.11	−1.51	0.132
Grammaticality*Group*Gender Match*Segment Precritical–Critical	0.34	0.22	1.5	0.134
Grammaticality*Group*Gender Match*Segment Critical–Postcritical	−0.28	0.22	−1.25	0.21
Grammaticality*Group*Gender Match*Segment Postcritical–Final	0.19	0.23	0.85	0.398

(d) Model with the bilingual children and sentence repetition and dominance as predictors
`modelSRTGroupGramm_B_Dom <- lmer(rt.r ~ gramm*Dominance*gender_match*
region + SRT.cent*Dominance + SRT.cent*gramm +
(1+gramm | subject) +
(1 | Picture),
data=stat1_L2,
REML=F, lmerControl(optimizer = "bobyqa"))`

Table A16. Summary of the model with the bilingual children and sentence repetition and dominance as predictors.

Predictors	Estimates	std. Error	t Value	p
Intercept	−0.91	0.07	−13.02	<0.001
Grammaticality	0.07	0.08	0.97	0.331
Dominance	0.04	0.08	0.51	0.611
Gender Match	0	0.07	0.05	0.961
Segment Precritical–Critical	0.11	0.09	1.23	0.221
Segment Critical–Postcritical	−0.01	0.09	−0.09	0.931
Segment Postcritical–Final	0.2	0.09	2.21	0.028
Sentence Repetition	−0.04	0.02	−1.53	0.127
Grammaticality*Dominance	−0.03	0.09	−0.29	0.773
Grammaticality*Gender Match	−0.08	0.14	−0.61	0.543
Dominance*Gender Match	0.01	0.08	0.06	0.948
Grammaticality*Segment Precritical–Critical	−0.36	0.18	−1.96	0.05
Grammaticality*Segment Critical–Postcritical	0.2	0.18	1.08	0.28

Table A16. Cont.

Predictors	Estimates	std. Error	t Value	p
Grammaticality*Segment Postcritical–Final	−0.14	0.18	−0.76	0.449
Dominance*Segment Precritical–Critical	−0.13	0.11	−1.2	0.23
Dominance*Segment Critical–Postcritical	0.02	0.11	0.21	0.836
Dominance*Segment Postcritical–Final	0.11	0.11	1.04	0.298
Gender Match*Segment Precritical–Critical	0.13	0.18	0.7	0.481
Gender Match*Segment Critical–Postcritical	0.02	0.18	0.11	0.911
Gender Match*Segment Postcritical–Final	−0.02	0.18	−0.13	0.899
Dominance*Sentence Repetition	0.04	0.02	1.43	0.152
Grammaticality*Sentence Repetition	−0.01	0.01	−1.12	0.264
Grammaticality*Dominance*Gender Match	0.11	0.16	0.72	0.47
Grammaticality*Dominance*Segment Precritical–Critical	0.48	0.22	2.18	0.029
Grammaticality*Dominance*Segment Critical–Postcritical	−0.18	0.22	−0.82	0.413
Grammaticality*Dominance*Segment Postcritical–Final	0.11	0.22	0.48	0.629
Grammaticality*Gender Match*Segment Precritical–Critical	−0.75	0.37	−2.03	0.042
Grammaticality*Gender Match*Segment Critical–Postcritical	0.18	0.37	0.49	0.626
Grammaticality*Gender Match*Segment Postcritical–Final	−0.76	0.37	−2.06	0.039
Dominance*Gender Match*Segment Precritical–Critical	−0.1	0.22	−0.48	0.635
Dominance*Gender Match*Segment Critical–Postcritical	−0.17	0.22	−0.78	0.436
Dominance*Gender Match*Segment Postcritical–Final	0.09	0.22	0.4	0.691
Grammaticality*Dominance*Gender Match*Segment Precritical–Critical	0.64	0.44	1.46	0.143
Grammaticality*Dominance*Gender Match*Segment Critical–Postcritical	0.09	0.44	0.21	0.837
Grammaticality*Dominance*Gender Match*Segment Postcritical–Final	0.58	0.44	1.32	0.189

Notes

¹ These assumptions do not apply to clitic left dislocation structures.

² Tsakali and Wexler (2004) assume that case is checked in Agr-O, whereas according to Tsakali and Anagnostopoulou (2008), gender and number are also checked in this phrase.

References

- Agresti, Alan. 2019. *An Introduction to Categorical Data Analysis*, 3rd ed. Hoboken: John Wiley & Sons.
- Alexiadou, Artemis, and Elena Anagnostopoulou. 2000. Greek Syntax: A Principles and Parameters Perspective. *Journal of Greek Linguistics* 1: 171–222. [CrossRef]
- Alexiadou, Artemis, Liliane Haegeman, and Melita Stavrou. 2008. *Noun Phrase in the Generative Perspective*. Studies in Generative Grammar. Berlin: Walter de Gruyter, vol. 71.
- Andreou, Maria, Eva Knopp, Christiane Bongartz, and Ianthi Maria Tsimpli. 2015. Character Reference in Greek-German Bilingual Children's Narratives. In *EUROSLA Yearbook*. Amsterdam: John Benjamins Publishing Company, vol. 15, pp. 1–40. [CrossRef]
- Avrutin, Sergey. 2006. Weak Syntax. In *Broca's Region*. Edited by Yosef Grodzinsky and Katrin Amunts. Oxford and New York: Oxford University Press, pp. 49–62.

- Baayen, Harald, Douglas Bates, Reinhold Kliegl, and Shravan Vasishth. 2015. *RePsychLing: Data Sets from Psychology and Linguistics Experiments. R Package Version 0.0.4*, Computer Software.
- Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using Lme4. *Journal of Statistical Software* 67: 1–48. [CrossRef]
- Bates, Douglas, Reinhold Kliegl, Shravan Vasishth, and Harald Baayen. 2018. Parsimonious Mixed Models. *arXiv* arXiv:1506.04967v2.
- Boersma, Paul, and David Weenink. 2017. *Praat: Doing Phonetics by Computer (Version 6.0.36)*, Computer Software.
- Cardinaletti, Anna, and Michal Starke. 1994. *The Typology of Structural Deficiency. On the Three Grammatical Classes*. Working Papers in Linguistics. Venice: University of Venice, vol. 4, pp. 41–109.
- Chondrogianni, Vicky. 2007. Acquiring Clitics and Determiners in Child L2 Modern Greek*. *Selected Papers on Theoretical and Applied Linguistics* 17: 356–66. [CrossRef]
- Chondrogianni, Vicky. 2008. Comparing Child and Adult L2 Acquisition of the Greek DP: Effects of Age and Construction. In *Current Trends in Child L2 Acquisition: Generative Approaches*. Edited by Belma Haznedar and Elena Gavrusseva. Amsterdam: John Benjamins, pp. 97–142.
- Chondrogianni, Vicky, Theodoros Marinis, Susan Edwards, and Elma Blom. 2015. Production and On-Line Comprehension of Definite Articles and Clitic Pronouns by Greek Sequential Bilingual Children and Monolingual Children with Specific Language Impairment. *Applied Psycholinguistics* 36: 1155–91. [CrossRef]
- Clackson, Kaili, Claudia Felser, and Harald Clahsen. 2011. Children’s Processing of Reflexives and Pronouns in English: Evidence from Eye-Movements during Listening. *Journal of Memory and Language* 65: 128–44. [CrossRef]
- De Houwer, Annick. 2009. *Bilingual First Language Acquisition*. MM Textbooks. Bristol and Buffalo: Multilingual Matters.
- Dimitropoulou, Maria, Jon Adoni Dunabeitia, Panagiotis Blitsas, and Manuel Carreiras. 2009. A standardized set of 260 pictures for Modern Greek: Norms for name agreement, age of acquisition, and visual complexity. *Behavior Research Methods* 41: 584–89. [CrossRef]
- Egger, Evelyn, Aafke Hulk, and Ianthi Maria Tsimpli. 2018. Crosslinguistic Influence in the Discovery of Gender: The Case of Greek–Dutch Bilingual Children. *Bilingualism: Language and Cognition* 21: 694–709. [CrossRef]
- Gavarró, Anna, Vicenç Torrens, and Ken Wexler. 2010. Object Clitic Omission: Two Language Types. *Language Acquisition* 17: 192–219. [CrossRef]
- Kaltsa, Maria, Ianthi Maria Tsimpli, and Froso Argyri. 2019. The Development of Gender Assignment and Agreement in English-Greek and German-Greek Bilingual Children. *Linguistic Approaches to Bilingualism* 9: 253–88. [CrossRef]
- Kaltsa, Maria, Alexandra Prentza, Despina Papadopoulou, and Ianthi Maria Tsimpli. 2020. Language External and Language Internal Factors in the Acquisition of Gender: The Case of Albanian-Greek and English-Greek Bilingual Children. *International Journal of Bilingual Education and Bilingualism* 23: 981–1002. [CrossRef]
- Klem, Marianne, Monica Melby-Lervåg, Bente Hagtvet, Solveig-Alma Halaas Lyster, Jan-Eric Gustafsson, and Charles Hulme. 2015. Sentence Repetition Is a Measure of Children’s Language Skills Rather than Working Memory Limitations. *Developmental Science* 18: 146–54. [CrossRef]
- Lenth, Russel. 2020. Emmeans: Estimated Marginal Means, Aka Least-Squares Means. Available online: <https://CRAN.R-project.org/package=emmeans> (accessed on 8 September 2022).
- Li, Ping. 2013. Successive Language Acquisition. In *The Psycholinguistics of Bilingualism*. Edited by François Grosjean and Ping Li. Hoboken: Wiley-Blackwell/John Wiley & Sons, pp. 145–67.
- Lüdecke, Daniel. 2019. Strengjacker: Load Packages Associated with Strengjacker!. R Package Version 0.5.0. Available online: <https://github.com/strengjacker/strengjacker> (accessed on 8 September 2022).
- Lüdecke, Daniel, Dominique Makowski, Philip Waggoner, and Indrajeet Patil. 2020. Performance: Assessment of Regression Models Performance. R Package Version 0.4.7. Available online: <https://CRAN.R-project.org/package=performance> (accessed on 8 September 2022).
- Manika, Sophia, Spyridoula Varlokosta, and Ken Wexler. 2010. The Lack of Omission of Clitics in Greek Children with SLI: An Experimental Study. In *BUCLD 35: Proceedings of the 35th Annual Boston University Conference on Language Development*. Edited by Nick Danis, Kate Mesh and Sung Hyunsuk. Somerville: Cascadilla Press, pp. 427–39.
- Marinis, Theodore. 2000. The Acquisition of Clitic Objects in Modern Greek: Single Clitics, Clitic Doubling, Clitic Left Dislocation. *ZAS Papers in Linguistics* 15: 259–81. [CrossRef]
- Marinis, Theodore. 2003. *The Acquisition of the DP in Modern Greek*. Language Acquisition and Language Disorders 31. Amsterdam: Benjamins.
- Marinis, Theodore. 2010. Using On-Line Processing Methods in Language Acquisition Research. In *Experimental Methods in Language Acquisition Research*. Edited by Elma Blom and Sharon Unsworth. Amsterdam and Philadelphia: John Benjamins Publishing Company, pp. 139–62.
- Mastropavlou, Maria. 2006. *The Role of Phonological Salience and Feature Interpretability in the Grammar of Typically Developing and Language Impaired Children*. Thessaloniki: Aristotle University of Thessaloniki. Available online: <https://thesis.ekt.gr/thesisBookReader/id/14071#page/1/mode/2up> (accessed on 8 September 2022).
- Mathôt, Sebastiaan, Daniel Schreij, and Jan Theeuwes. 2012. OpenSesame: An Open-Source, Graphical Experiment Builder for the Social Sciences. *Behavior Research Methods* 44: 314–24. [CrossRef]

- Mavrogiorgos, Marios. 2010a. *Clitics in Greek: A Minimalist Account of Proclisis and Enclisis*. *Linguistik Aktuell/Linguistics Today*, v. 160. Amsterdam: John Benjamins.
- Mavrogiorgos, Marios. 2010b. Internal Structure of Clitics and Cliticization. *Journal of Greek Linguistics* 10: 3–44. [CrossRef]
- Patil, Umesh, Shravan Vasishth, and Richard L. Lewis. 2016. Retrieval Interference in Syntactic Processing: The Case of Reflexive Binding in English. *Frontiers in Psychology* 7: 329. [CrossRef]
- Prentza, Alexandra, Maria Kaltsa, Ianthi Maria Tsimpli, and Despina Papadopoulou. 2019. The Acquisition of Greek Gender by Bilingual Children: The Effects of Lexical Knowledge, Oral Input, Literacy and Bi/Monolingual Schooling. *International Journal of Bilingualism* 23: 901–20. [CrossRef]
- R Core Team. 2020. *R: A Language and Environment for Statistical Computing (Version 3.5.3)*. Vienna: R Foundation for Statistical Computing. Available online: <https://www.R-project.org/> (accessed on 8 September 2022).
- Radford, Andrew. 1997. *Syntactic Theory and the Structure of English: A Minimalist Approach*, 1st ed. Cambridge: Cambridge University Press. [CrossRef]
- Ralli, Angela. 2002. The Role of Morphology in Gender Determination: Evidence from Modern Greek. *Linguistics* 40: 519–51. [CrossRef]
- Raven, Jean, John C. Raven, and John C. Court. 1998. *Manual for Raven's Progressive Matrices and Vocabulary Scales*. San Antonio: Harcourt Assessment.
- Renfrew, Catherine. 1997. *Action Picture Test*. Bicester: Winslow Press.
- Revithiadou, Anthi, and Vassilios Spyropoulos. 2020. Cliticisation in Greek: A Contrastive Examination and Cross-Linguistic Remarks. In *Contrastive Studies in Morphology and Syntax*. Edited by Michalis Georgiades, Giannoula Giannoulou, Maria Koliopoulou and Angeliki Tsokoglou. London: Bloomsbury Academic, pp. 225–44.
- Rivero, Maria Luisa. 1994. Clause Structure and V-Movement in the Languages of the Balkans. *Natural Language & Linguistic Theory* 12: 63–120. [CrossRef]
- Rossi, Eleonora, Judith F. Kroll, and Paola E. Dussias. 2014. Clitic Pronouns Reveal the Time Course of Processing Gender and Number in a Second Language. *Neuropsychologia* 62: 11–25. [CrossRef]
- Schaeffer, Jeannette C. 2000. *The Acquisition of Direct Object Scrambling and Clitic Placement: Syntax and Pragmatics*. *Language Acquisition & Language Disorders*, v. 22. Amsterdam and Philadelphia: J. Benjamins Pub. Co.
- Smith, Nafsika, Susan Edwards, Vesna Stojanovic, and Spyridoula Varlokosta. 2008. Object Clitic Pronouns, Definite Articles and Genitive Possessive Clitics in Greek Preschool Children with Specific Language Impairment (SLI): Implications for Domain-General and Domain-Specific Accounts of SLI. In *Supplement Proceedings of 32nd Boston University Conference of Child Language Development*. Edited by Harvey Chan, Enkeleida Kapia and Heather Jacob. Boston: Cascadilla Press.
- Sorace, Antonella. 2004. Native Language Attrition and Developmental Instability at the Syntax-Discourse Interface: Data, Interpretations and Methods. *Bilingualism: Language and Cognition* 7: 143–45. [CrossRef]
- Spathas, Giorgos, and Yasutada Sudo. 2020. Gender on Animal Nouns in Greek. *Catalan Journal of Linguistics* 19: 25–48. [CrossRef]
- Sportiche, Dominique. 1996. Clitic Constructions. In *Phrase Structure and the Lexicon*. Edited by Johan Rooryck and Laurie Zaring. *Studies in Natural Language and Linguistic Theory*. Dordrecht: Springer, vol. 33, pp. 213–76. [CrossRef]
- Stavrakaki, Stavroula. 2001. *Specific Language Impairment in Greek: Aspects of Syntactic Production and Comprehension*. Thessaloniki: Aristotle University of Thessaloniki.
- Stavrakaki, Stavroula, and Areti Okalidou. 2016. Gr-LARSP: Towards a Greek Version of LARSP. In *Profiling Grammar: More Languages of LARSP*. Edited by Paul Fletcher, Martin J. Ball and David Crystall. Blue Ridge Summit: Multilingual Matters. [CrossRef]
- Stavrakaki, Stavroula, and Ianthi Maria Tsimpli. 2000. Diagnostic Verbal IQ Test for Greek Preschool and School Age Children: Standardization, Statistical Analysis, Psychometric Properties. In *Proceedings of the 8th Conference on Speech Therapy*. Athens: Ellinika Grammata, pp. 95–106. (In Greek)
- Stavrakaki, Stavroula, and Heather van der Lely. 2010. Production and Comprehension of Pronouns by Greek Children with Specific Language Impairment. *British Journal of Developmental Psychology* 28: 189–216. [CrossRef] [PubMed]
- Stavrakaki, Stavroula, Marie-Annick Chrysomallis, and Evangelia Petraki. 2011. Subject-Verb Agreement, Object Clitics and Wh-Questions in Bilingual French-Greek SLI: The Case Study of a French-Greek-Speaking Child with SLI. *Clinical Linguistics & Phonetics* 25: 339–67. [CrossRef]
- Stewart, Andrew J., Judith Holler, and Evan Kidd. 2007. Shallow Processing of Ambiguous Pronouns: Evidence for Delay. *Quarterly Journal of Experimental Psychology* 60: 1680–96. [CrossRef] [PubMed]
- Talli, Ioanna, and Stavroula Stavrakaki. 2020. Short-Term Memory, Working Memory and Linguistic Abilities in Bilingual Children with Developmental Language Disorder. *First Language* 40: 437–60. [CrossRef]
- Tedeschi, Roberta. 2008. Referring Expressions in Early Italian: A Study on the Use of Lexical Objects, Pronouns and Null Objects in Italian Pre-School Children. In *LOT Occasional Series*. Utrecht: LOT, Netherlands Graduate School of Linguistics, vol. 8, pp. 201–16. Available online: <https://dspace.library.uu.nl/handle/1874/296782> (accessed on 8 September 2022).
- Terzi, Arhonto. 1996. The Linear Correspondence Axiom and the Adjunction Site of Clitics. In *Configurations*. Edited by Anna-Maria Di Sciullo. Boston: Cascadilla Press, pp. 185–89.
- Torregrossa, Jacopo, Maria Andreou, Christiane Bongartz, and Ianthi Maria Tsimpli. 2021. Bilingual acquisition of reference: The role of language experience, executive functions and crosslinguistic effects. *Bilingualism: Language and Cognition* 24: 694–706. [CrossRef]
- Tsakali, Vina. 2014. Acquisition of Clitics: The State of the Art. In *Developments in the Acquisition of Clitics*. Edited by Theoni Neokleous and Kleantes K. Grohmann. Cambridge: Cambridge Scholars Publishing, pp. 161–87.

- Tsakali, Vina, and Elena Anagnostopoulou. 2008. Rethinking the Clitic Doubling Parameter: The Inverse Correlation between Clitic Doubling and Participle Agreement. In *Clitic Doubling in the Balkan Languages*. Edited by Dalina Kallulli and Liliane Tasmowski. Amsterdam: John Benjamins, pp. 321–57.
- Tsakali, Vina, and Ken Wexler. 2004. Why Children Omit Clitics in Some Languages but Not in Others: New Evidence from Greek. In *Proceedings of Generative Approaches to Language Acquisition 2003*. Edited by Jacqueline van Kampen and Sergio Baauw. Utrecht: LOT, vol. II, pp. 493–504.
- Tsimpli, Ianthi Maria. 2001. LF-Interpretability and Language Development: A Study of Verbal and Nominal Features in Greek Normally Developing and SLI Children. *Brain and Language* 77: 432–48. [[CrossRef](#)]
- Tsimpli, Ianthi Maria. 2003. Clitics and Determiners in L2 Greek. In *Proceedings of the 6th Generative Approaches to Second Language Acquisition Conference (GASLA 2002)*. Edited by Juana Liceras. Somerville: Cascadilla Proceedings Project, pp. 331–39.
- Tsimpli, Ianthi Maria. 2014. Early, Late or Very Late?: Timing Acquisition and Bilingualism. *Linguistic Approaches to Bilingualism* 4: 283–313. [[CrossRef](#)]
- Tsimpli, Ianthi Maria, and Aafke Hulk. 2013. Grammatical Gender and the Notion of Default: Insights from Language Acquisition. *Lingua* 137: 128–44. [[CrossRef](#)]
- Tsimpli, Ianthi Maria, and Maria Mastropavlou. 2008. Feature Interpretability in L2 Acquisition and SLI: Greek Clitics and Determiners. In *The Role of Formal Features in Second Language Acquisition*. Edited by Juana Liceras, Helmut Zobl and Helen Goodluck. New York: Lawrence Erlbaum, pp. 142–83.
- Tsimpli, Ianthi Maria, and Stavroula Stavrakaki. 1999. The Effects of a Morphosyntactic Deficit in the Determiner System: The Case of a Greek SLI Child. *Lingua* 108: 31–85. [[CrossRef](#)]
- Unsworth, Sharon, Froso Argyri, Leonie Cornips, Aafke Hulk, Antonella Sorace, and Ianthi Tsimpli. 2014. The Role of Age of Onset and Input in Early Child Bilingualism in Greek and Dutch. *Applied Psycholinguistics* 35: 765–805. [[CrossRef](#)]
- Varlokosta, Spyridoula. 2002. (A)Symmetries in the Acquisition of Principle B in Typically Developing and Specifically Language Impaired Children. In *The Process of Language Acquisition*. Edited by Ingeborg Lasser. Berlin: Peter Lang Verlag, pp. 81–98.
- Varlokosta, Spyridoula, and Michaela Nerantzini. 2013. Grammatical gender in specific language impairment: Evidence from determiner-noun contexts in Greek. *Psychologia* 20: 338–57. [[CrossRef](#)]
- Varlokosta, Spyridoula, Katerina Konstantzou, and Michaela Nerantzini. 2014. On the Production of Direct Object Clitics in Greek Typical Development and Specific Language Impairment: The Effect of Task Selection. In *Developments in the Acquisition of Clitics*. Edited by Theoni Neokleous and Kleanthes K. Grohmann. Cambridge: Cambridge Scholars Publishing, pp. 188–211.
- Varlokosta, Spyridoula, Adriana Belletti, João Costa, Naama Friedmann, Anna Gavarró, Kleanthes K. Grohmann, Maria Teresa Guasti, Laurice Tuller, Maria Lobo, Darinka Anđelković, and et al. 2016. A Cross-Linguistic Study of the Acquisition of Clitic and Pronoun Production. *Language Acquisition* 23: 1–26. [[CrossRef](#)]
- Vogindroukas, Ioannis, Athanassios Protopapas, and Stavroula Stavrakaki. 2010. *The Greek Version of the Action Picture Test (Renfrew 1997)*. Chania: Glafki.
- Wexler, Ken. 1998. Very Early Parameter Setting and the Unique Checking Constraint: A New Explanation of the Optional Infinitive Stage. *Lingua* 106: 23–79. [[CrossRef](#)]
- Wolf, Florian, Edward Gibson, and Timothy Desmet. 2004. Discourse Coherence and Pronoun Resolution. *Language and Cognitive Processes* 19: 665–75. [[CrossRef](#)]
- Yip, Virginia. 2013. Simultaneous Language Acquisition. In *The Psycholinguistics of Bilingualism*. Edited by François Grosjean and Ping Li. Hoboken: Wiley-Blackwell/John Wiley & Sons, pp. 119–44.