Adding discourse to sentence repetition tasks: Under which conditions does bilingual children's performance improve?

Jacopo Torregrossa [a,*], Andrea Listanti [b], Christiane Bongartz [c], Theodoros Marinis [d]

[a] *Institute of Romance Languages and Literatures, Goethe University of Frankfurt, IG 6.157, Goethe University, Frankfurt am Main 60629, Germany*
[b] *Collaborative Research Centre "Prominence in Language", University of Cologne, Germany*
[c] *English Seminar I, University of Cologne, Germany*
[d] *Department of General Linguistics and Multilingualism, University of Konstanz, Germany*

ABSTRACT

Sentence repetition tasks (SRTs) have been extensively used as measures of bilinguals' language abilities. Most studies relied on SRTs in which the target sentences were not connected to each other. However, participants' performance may differ if these sentences are embedded in discourse, since discourse provides participants with additional cues for sentence comprehension and interpretation. For the present study, we designed a discourse-based SRT, whereby the target sentences were connected to each other in a story. We examined the effect of discourse on bilinguals' performance in the SRT and investigated whether this effect varied based on the language of administration, bilinguals' dominance score and type of target structure. We tested 32 Italian-German bilingual children (7–12 years) living in Germany with two SRTs in each language, one with discourse and one without discourse. Participants showed a better performance in the SRTs with discourse, especially in the heritage language (Italian). The effect of discourse was visible across the board with all target structures. On the whole, SRTs with discourse seem to reduce the processing costs associated with lexical retrieval and shifts in scenarios, thus tapping more directly into children's processing abilities, compared to more traditional SRTs. The results are discussed in terms of ecological validity of different assessment instruments.

1. Introduction

Scholars working in psycholinguistics and language acquisition have advocated for the ecological validity of the research instruments used for data collection (de Groot & Hagoort, 2017 for a general overview). The expression "ecological validity" can refer to different aspects of an experiment, ranging from the design of the stimuli to its administration (Holleman et al., 2020 for a review). In the present study, we examined how far the use of experiments informed by criteria of ecological validity affected participants' results in these experiments. In particular, we dealt with the use of sentence repetition tasks (SRTs, henceforth) as instruments to assess bilingual children's abilities in one or the other language. SRTs involve listening to target sentences and reproducing them (see Section 1.1 below). For the present study, we designed a SRT in which the target sentences were connected to each other in a narrative. In this way, children got involved in a goal-oriented activity: by listening and reproducing the target sentences, they were able to advance in the plot of the narrative.

---

* Corresponding author.
E-mail address: Torregrossa@lingua.uni-frankfurt.de (J. Torregrossa).

In traditional SRTs, the sentences are presented in isolation from each other and are not contextualized. This may make the task artificial, since children may have a hard time figuring out why they should repeat sentences that are unrelated to each other. As a result, children may have little motivation to participate in this type of task. Storyfying the task, then, should make SRTs more child-friendly and increase children's engagement and motivation. In other words, using a narrative-based version of the SRT allows us to measure children's language abilities through a task which is similar to their real-world practices: Adult-child shared story-telling is a practice to which children are familiar (from their past experience, at least) in the age range considered in this study. Therefore, children may be able to experience continuity between language practices in and outside of the lab.

The aim of the present study was to investigate the extent to which children perform differently in a SRT that is embedded within a narrative compared to a traditional SRT. Towards that aim we administered two versions of a SRT to Italian-German bilingual children ranging in age between 7 and 12 years: a "traditional" version – in which the sentences were independent from each other – and a narrative-based one. We tested the children in both languages with the two SRT types and examined their performance in terms of the syntactic complexity of the target structures, language dominance, and the respective social status of German and Italian as societal and heritage language, respectively.

1.1. Sentence repetition task as a measure of bilinguals' language abilities

Sentence Repetition Tasks have been shown to be reliable instruments for assessing speakers' language abilities at many different levels (Devescovi & Caselli, 2007; Klem et al., 2015) and tap into different aspects of sentence comprehension and production. The present study focuses on children's grammar abilities. In a SRT, participants listen to a sentence and reproduce it as accurately as possible. In this sense, SRT is a type of Elicited Imitation Task (EIT). EITs often include both grammatical and ungrammatical sentences. Test-takers usually correct ungrammatical sentences if they acquired the structure corresponding to the target sentence (see, e. g., Spada et al., 2015). On the contrary, SRTs usually include only grammatical sentences.[1] Each sentence in a SRT taps into the speaker's mastery of a specific grammatical structure. To successfully reproduce the stimulus sentence, participants must rely on their online processing for comprehension, in order to reconstruct what they heard. From the production point of view, this involves lexical retrieval as well as grammatical and phonological encoding (Klem et al., 2015; Marinis & Armon-Lotem, 2015; Schönström & Hauser, 2022).

SRTs have been used extensively in first and second-language acquisition research. In general, it is assumed that learners' ability to reproduce a target structure is an indicator of their successful acquisition of this structure (e.g., Yan et al., 2016). Furthermore, some studies in second language acquisition have shown that SRTs are a measure of implicit language knowledge: since they are mainly reconstructive in nature, they tend to induce focus on meaning that detracts from focus on form (Tomita et al., 2009; Trofimovich et al., 2009). In this sense, they differ from tasks inducing focus on form, such as error correction tasks, in which participants are supposed to abstract away from sentence meaning. For example, Spada et al. (2015) tested 73 adult learners of English using an error correction task, an oral narration test, an EIT featuring grammatical and ungrammatical sentences and two timed grammaticality judgement tasks, one in the oral and one in the written mode. Based on an exploratory factor analysis, the authors showed that learners' scores in the EIT loaded on the same factor as the two timed grammaticality judgement tasks. This factor was crucially different from the factor onto which the error correction task loaded. They interpreted this result as showing that EITs tap into learners' grammatical processing and implicit language knowledge.

In bilingualism research, the COST Action IS0804 underlined the importance of testing bilingual children's abilities in both their languages when it comes to diagnoses of Developmental Language Disorders (DLD; Marinis & Armon-Lotem, 2015). To this purpose, the scholars working in this project developed SRTs that were comparable across languages, considering linguistic structures that had been shown to be problematic for children with DLD across languages. These structures involved different degrees of complexity related to embedding and syntactic movement (Marinis & Armon-Lotem, 2015). The instrument that we developed for the present study was not designed for the assessment of children with DLD. Rather, we only considered typically developing bilingual children in both of their languages, in order to examine their grammar abilities and understand how far they performed better in a narrative-based SRT than in a non-narrative-based one, since the former should be more ecologically valid than the latter. However, the study may have implications for the assessment of bilingual children with DLD. Children may benefit from the embedding of the SRT in a narrative across the board, independently of their being typically developing or with DLD. We deal with these implications in the discussion section.

Building on the comparability guidelines proposed in the COST project, studies conducted with typically developing bilingual children showed that SRTs could be considered as reliable instruments for assessing children's dominance in one or the other language. For example, Andreou et al. (2021) designed two comparable versions of a SRT in Greek and Italian, respectively. They administered these SRTs to 38 Greek-Italian bilingual children ranging in age between 8 and 12 years together with a vocabulary task and a questionnaire tapping into children's exposure to Greek or Italian in different contexts over time. The study showed that the difference scores between the SRTs in Greek and Italian correlated with the difference scores in vocabulary and in several language-exposure variables. This suggests that children's performance in SRTs is sensitive to their dominance in one or the other language. Therefore, the present study considered dominance as a relevant factor.

Across the literature, several factors have been shown to affect bilingual children's performance in SRTs. Among child-internal

---

[1] However, there are some exceptions to this generalization across the literature. For example, Polišenská (2011) designed a SRT featuring ungrammatical sentences.

factors, some relate to bilingualism *per se* and others are relevant for monolinguals and bilinguals alike. For example, age of acquisition of a language seems to have a strong impact on bilinguals' scores in the SRT in this language, whereby earlier-onset learners tend to perform better than later-onset ones (Armon-Lotem et al., 2011).

Cross-linguistic effects may also affect bilingual children's accuracy to reproduce certain structures. For example, Meir et al. (2016) tested L1-Russian-L2-Hebrew bilingual children in preschool age and showed cross-linguistic effects from Hebrew to Russian in the domain of knowledge of case marking. Likewise, Torregrossa et al. (2023a) showed that Italian-German bilingual children's accuracy of reproduction of non-finite complementizers was affected by the fact that German allows for only one complementizer, whereas Italian has more than one. In other words, SRTs were sensitive to the typology of the language pairs at issue. Children were shown to adjust their stimulus reproduction in the direction of their less complex system(s).

Age is another factor that seems to play a relevant role in repetition accuracy in SRTs among both monolingual and bilingual children (see, e.g., Polišenská et al., 2015). SRTs usually include sentences of increasing levels of complexity (Marinis & Armon-Lotem, 2015). Therefore, a certain degree of cognitive maturity is needed to achieve full mastery of the most complex structures (Torregrossa et al., 2023b for similar considerations). The precise role of memory in children's performance in SRTs is controversial. On the one hand, some studies suggested that SRTs tap into the episodic buffer, a component of memory involved in the integration of information from working memory and long-term memory (Alloway et al., 2004). On the other hand, other studies showed that SRTs measure children's language abilities independently of their memory capacity (Klem et al., 2015). In general, when designing a SRT, it should be ensured that speakers do not rely exclusively on rote repetition for the reproduction of the target sentence (Tomita et al., 2009). For example, to prevent speakers from "parroting", target sentences should be relatively long (Marinis & Armon-Lotem, 2015). The way of coding children's responses can also minimize the impact of memory capacity on children's performance in the SRT if performance is measured through the accurate reproduction of the target structures rather than the exact repetition of the sentences. For example, a response can be coded as accurate if children reproduce the target structure accurately even if they do not use exactly the same words they heard.

In the current study, we coded response accuracy based on the accurate reproduction of the target structure, in line with the main aims and research questions of the study (Section 2). We did not use the SRT designed for this study as a clinical instrument or an instrument for the assessment of children's cognitive as well as grammar abilities. Therefore, we chose to adopt a coding that would be the least sensitive to variation in participants' working memory capacity (Hamann & Abed Ibrahim, 2017; see Section 3.3). However, it should be mentioned that coding response accuracy based on *verbatim* repetitions may be more straightforward for coders who are not trained in linguistics, and hence, less subject to interrater reliability issues than the coding adopted in this study (see Vinther, 2002 for methodological considerations related to different coding procedures).

Furthermore, children's vocabulary knowledge also plays a role in children's performance in a SRT. For example, Simon-Cereijido and Méndez (2018) administered two SRTs to English-Spanish bilingual children, one in English and one in Spanish. They showed that the children's performance was predicted by their expressive vocabulary score in the respective language.

Bilingual children's performance in SRTs is also affected by child-external variables, such as variables related to children's language and literacy exposure. For example, the study by Pratt et al. (2021) involving English-Spanish bilingual typically developing children and children with DLD showed that among the typically developing children, the amount of exposure to English was a strong predictor of their performance in the SRT in English (see also Fleckstein et al., 2018 for similar results with typically developing French bilinguals). Andreou et al. (2021) offered a more nuanced view of the relation between language exposure variables and accuracy of reproduction of linguistic structures in a SRT. The authors considered the accuracy of reproduction of structures differing from each other in their degree of complexity (e.g., involving embedding, movement, etc.). Then, they analysed how sensitive these structures were to different language-exposure variables, for instance, current language exposure, language exposure between 0 and 3 years and between 3 and 6. The results showed that accuracy of reproduction of certain structures was sensitive to language-exposure variables related to the stage in which these structures usually emerge in monolingual language acquisition. Another relevant variable that affects performance in a SRT is literacy exposure. For example, Torregrossa et al. (2022) showed that Italian heritage children living in Greece were very accurate in their reproduction of target structures in Italian, with the lowest result corresponding to the accurate repetition of still almost 65 % of target structures. Crucially, all of these children attended an Italian immersion school in Greece. Likewise, De Cat (2020) administered a SRT in English to bilingual children of different language combinations in the UK ranging in age between 5 and 7 years. These children attended English monolingual schools. Crucially, De Cat showed that exposure to English at school was the unique predictor of children's accuracy of repetition in English.

As a final remark, we would like to come back to the issue of the ecological validity of the existing SRTs. As mentioned in Section 1, in all existing SRTs to our knowledge, the target sentences are independent from each other, which may make the task artificial and distant from real-world practices. In order to overcome this shortcoming, Marinis and Armon-Lotem (2015) proposed to embed the SRT into a board game. However, the connection between the reproduction of the sentence and the way the game unfolds may not be intuitive for children. As an alternative, in this study, we decided to embed the SRT into a storytelling activity with which children are usually familiar (see Section 1).

1.2. The role of discourse in bilingual language processing

Several studies have suggested that bilingual speakers may exhibit difficulties in integrating information at the sentence and discourse level. Grüter and Rohde (2021) showed that L1 speakers of English were able to anticipate the upcoming mention of a referent based on verb semantics and grammatical aspect. The authors used sentences like 'Patrick gave/was giving Emma a bottle of wine', featuring a transfer-of-possession verb which denoted either a complete or an incomplete event (*gave* vs. *was giving*). Based on a

visual-world eye-tracking paradigm study, they showed that L1 speakers were more likely to anticipate reference to the goal argument (Emma) when the verb was in simple past (*gave*) than when it was in past progressive (*was giving*). By contrast, adult L2 speakers of English did not seem to be sensitive to the aspectual properties of the verb in predicting upcoming reference. These results are in line with the Interface Hypothesis (Sorace, 2011), which argues that bilinguals and L2 speakers may experience difficulties when integrating morphosyntactic (the aspectual properties of a verb, for instance) with discourse information, which – in the case at issue – is involved in the accurate anticipation of an upcoming referent. Difficulty in integrating morphosyntactic with discourse information may be due to processing limitations: processing in a L2 is usually slower than processing in a L1 (Schlenter, 2022 for a review), especially if proficiency in the L2 is limited (Karaca et al., 2021). As a result, L2 speakers may exhibit "patterns of non-convergence and residual optionality" (Sorace, 2011: 1) in tasks that involve a considerable amount of processing resources, such as the ones requiring the integration of morphosyntactic and discourse information. Based on such considerations, a narrative-based SRT – as the one considered in this study – may complexify the task of comprehending and reproducing the target structures, since participants have not only to attend to incoming input (as in traditional SRTs), but also to keep track of previous information and generate expectations about upcoming information, which is associated with additional processing costs.

Very few studies have investigated bilingual children's ability to integrate morphosyntactic and discourse information during online processing. With bilingual children, we refer here to children that have been exposed to another language between birth and 6 years (upon entering in school). Some production studies indicated that bilingual children may exhibit a different mastery of syntax-discourse interface phenomena than their monolingual peers (Serratrice & De Cat, 2020; Torregrossa & Bongartz, 2018 and Torregrossa et al., 2021 on the production of referring expressions; Listanti & Torregrossa, 2023 on postverbal subjects). However, studies comparing production with online comprehension showed a discrepancy between these two modalities, with differences between L2 children and monolinguals in production but similar performance in online comprehension (Chondrogianni & Marinis, 2012, 2016; Pontikas et al., 2022). Moreover, studies on bilingual children's ability to generate expectations about upcoming information at the sentence level revealed that bilingual children behaved on a par with or even better than monolinguals (Brouwer et al., 2017; Meir et al., 2020).

Importantly, integrating grammar with discourse information has been found to be a challenge also for children growing up with just one language. For example, Papadopoulou et al. (2015) showed that at the age of 10, Greek monolingual children are not yet adult-like in their interpretation and processing of null subjects in Greek. In order to interpret pronouns, children have to refer to the information contained in previous discourse (e.g., whether the event expressed by a previous sentence is complete or not, as in the example shown above), the discourse relation linking the previous sentence and the current one and information contained in the current sentence (e.g., verb agreement morphology). In general, linguistic phenomena that require the integration of morphosyntax with other linguistic domains tend to be acquired late in monolingual as well as bilingual children (Tsimpli, 2014).

Although overall the integration of morphosyntax with discourse has been shown to be challenging in monolinguals and bilinguals, this does not seem to hold across the board. The integration of morphosyntax with discourse can have a facilitating effect for the production and comprehension of certain syntactic structures that are sensitive to discourse constraints. For example, several studies have indicated that adult monolingual speakers process marked word order (such as topicalizations, clefts and scrambling) more easily if they are preceded by a relevant context than if they are presented in isolation (e.g., Kaiser & Trueswell, 2004; López-Beltrán et al., 2022; Yano & Koizumi, 2018). For the purposes of the present study, this predicts that children may benefit from discourse embedding in association with structures whose meaning is sensitive to discourse context, such as structures exhibiting marked word orders. For example, object-topicalizations (e.g., 'The book he is reading') and passive sentences (e.g., 'The book is read by him') are usually licensed by previous discourse, which, for instance, gives prominence to objects/themes. Likewise, processing of object relative clauses is facilitated when they are embedded into an appropriate discourse context (Brandt et al., 2009; Mak et al., 2008). Based on these findings, we expected to find an effect of discourse on children's performance in the SRT in association with discourse-sensitive structures.

However, discourse may benefit bilinguals' performance in a SRT independently of the type of target structure at stake. This may be related to the following reasons: First, SRTs with discourse are more ecologically valid (Section 1). Second, they offer more cues for sentence comprehension and reproduction compared to SRTs without discourse. In particular, in a SRT with discourse, certain words may be repeated and other words may be activated because they are connected semantically to each other throughout the story (e.g., they belong to the same semantic field). This should facilitate lexical access. Furthermore, the interpretation and prediction of thematic roles does not rely only on morphological cues and word order information, but also on considerations of coherence with respect to previous discourse.

Finally, it should be mentioned that the outcome of our investigation might be more complex than the two hypotheses sketched above, which predict either a facilitative or hampering effect of discourse on bilingual children's performance in a SRT. In Section 1.1, we mentioned that performance in a SRT benefits from literacy exposure in the target language. The main language of instruction at school usually corresponds to the main language in society, which is also the case for the children considered in this study (Section 3.1). Therefore, we expect these children to perform better in a SRT administered in the societal language than the heritage language. Within this general pattern, one may observe variation depending on whether children are more or less exposed to the societal or the home language across different contexts beyond school, which also affects their abilities in one or the other language. Whether discourse has a facilitative or a hampering effect, this effect should be more visible among the children with less developed grammar abilities in the target language. In other words, discourse may either exacerbate the processing limitations experienced by bilingual children in the non-dominant language (in society and in their daily language experience) or benefit performance in it. By contrast, the dominant language should be more stable, and hence, less vulnerable to the effects of discourse.

2. The present study

We investigated Italian-German bilingual children who were exposed to Italian from birth and to German between birth and 72 months (upon entering school). At the time of testing, the children were attending a German-Italian bilingual school in which German was the main medium of instruction, with instruction in German amounting to around 20 h per week and instruction in Italian to around 12 h per week. We aimed to investigate how far children's performance in SRTs – as administered in both German and Italian – differed depending on whether the SRTs included a discourse dimension or not. In the two SRTs including discourse, the target sentences were connected to each other to build a narrative, whereas in the two SRTs without discourse, the target sentences were independent from each other.

We addressed the following research questions:

1. Does discourse enhance children's performance in the SRT?

We expected a SRT with discourse to enhance children's performance. SRTs with discourse should be ecologically more valid than SRTs without discourse.

2. Does discourse have a different effect on children's performance in the SRT, depending on whether the task is administered in the societal (German) or the heritage language (Italian)?

We expected discourse to enhance children's performance only in the heritage language. Bilingual children's language processing tends to be stable in their societal/school language, whereby they should be able to perform in the SRT in the same way independently of the mode of administration (with or without discourse). By contrast, their processing abilities in the heritage language should be more vulnerable: in this case, discourse should support their sentence comprehension and production. For comprehension, the felicitousness of a target structure with respect to previous discourse should facilitate its processing (Section 1.3). Likewise, a preceding appropriate discourse should render the production of the target structures more natural (Section 1). Children's production may also be facilitated by the repetition of certain words throughout the task (see, e.g., alien, beaver and spaceship in the SRTs with discourse in Supplementary Materials 1) and the possibility for children to build expectations on upcoming words, referents and events.

3. Does children's dominance in one or the other language modulate the effect of discourse on their performance in the SRT in each language?

Based on the above considerations related to the facilitative role of discourse on children's performance in the SRT, we expected children who were less dominant in the target language to perform better in the SRT with discourse than the one without discourse.

4. Is a positive effect of discourse (if any) visible for all target structures or only in association with specific ones?

Based on the considerations in Section 1.3, we hypothesized the effect of discourse to be more visible with the structures that are more sensitive to discourse constraints, such as marked word orders or object relative clauses.

3. Method

3.1. Participants

We tested 32 Italian-German bilingual children living in Germany who had Italian as their heritage language and German as their societal language. We had to exclude 3 children from the analysis because they missed some experimental sessions. Therefore, the study included 29 children ranging in age from 7 years and 8 months to 12 years and 6 months (M_{age} = 9 years and 10 months, SD = 20 months). We recruited the children in two schools in Germany where German was the main medium of instruction. Italian was offered both as a language subject and as vehicular language for Italian history and geography. In some classes, some modules of natural science were also taught in Italian. We obtained written consent from the parents. Ethical approval for the study was obtained from the Ethics Committee of the German Society for Linguistics (Deutsche Gesellschaft für Sprachwissenschaft – DGfS). Furthermore, the children were told that they were free to take part in the study and could withdraw from it whenever they wanted. The sample included 13 simultaneous bilinguals, i.e., exposed to Italian and German from birth, 3 early sequential bilinguals who were exposed to Italian from birth and German at the age of 3, and 7 late successive bilinguals who were exposed to Italian from birth and German when they entered primary school at the age of 6. Six parents did not provide the relevant information related to age of acquisition of Italian or German. The parents and the teachers reported that none of the children had a previously identified speech, language or hearing impairment.

3.2. Research instruments

3.2.1. Background questionnaires
The parents were administered a background questionnaire tapping into children's exposure to Italian and German across different

contexts (with family members or friends, during literacy-related activities outside school or other types of leisure activities, etc.) over time (currently or in the past). The questionnaire was based on Torregrossa et al. (2021) and Torregrossa et al. (2022). It was structured into four parts. The first considered children's *home language history*, i.e., the amount of exposure they received at the age of 3, between 3 and 6 and at the age of 6. The second tapped children's *early literacy practices*, i.e., whether and in which language(s) their parents read book to them in their preliterate years. The third examined children's *current use* of one or the other language with family members and friends and during after-school activities. The fourth considered children's *current literacy practices*, i.e., in which language(s) children conducted literacy practices outside of the school, such as writing e-mails or reading books. In Supplementary Materials 2, we report some of the questions used in each part of the questionnaire and show how the partial score related to each part was calculated, based on Torregrossa et al. (2021) and Torregrossa et al. (2022).

3.2.2. Vocabulary task

We tested children's vocabulary knowledge in Italian and German, respectively. For Italian, we used the expressive vocabulary task by Renfrew (1995). The choice of this instrument was related to the fact that no normed instruments were available for testing vocabulary knowledge in Italian in the age range at issue at the moment of testing. Furthermore, this vocabulary test was used as a reliable measure of vocabulary in some previous studies of ours (e.g., Torregrossa et al., 2022). The task consisted of 50 pictures of objects; the target words were all nouns. For German, we used the expressive vocabulary task by Petermann et al. (2010), which is a normed instrument. It consists of 40 pictures meant to elicit 30 nouns and 10 verbs. In both tasks, the children were asked to name the pictures that they were shown. If they could not do it immediately, they were provided with a semantic cue: we used the same cues for all participants. In this way, we made sure that the children were able to recognize the object represented in the picture. If they were still not able to name the object, we provided a phonemic cue, consisting in the first syllable of the target word.

3.2.3. Sentence repetition tasks

We designed four comparable versions of a Sentence Repetition Task (SRT), two for German and two for Italian. The two Italian versions and the two German versions differed from each other depending on whether the target sentences were embedded or not in a narrative. In particular, we distinguished between a SRT *with discourse*, in which the target sentences were linked to each other in a narrative, and a SRT *without discourse*, in which the sentences were independent from each other. The narrative in the discourse condition was about an alien who landed on Earth in a broken spaceship and met a beaver at the crash site. The beaver helped him to fix his spaceship, so that he could go back to space (see Supplementary Materials 1). The scenarios corresponding to the target sentences in the SRTs without discourse were adapted from the ones used in Costa and Guasti (2021). In this way, we could rely on a task already validated in a previous study involving several children (121 bilinguals and 71 monolinguals in Costa & Guasti, 2021) as a baseline. Most sentences in Costa and Guasti (2021) were modified to make them comparable to the corresponding sentences in the SRT with discourse, especially in terms of sentence length (see the statistics reported below). However, there were still some minor details in which the two tasks (with and without discourse) differed from each other. For example, some pronouns in the SRT with discourse were replaced by full nouns in the SRT without discourse (see e.g., (3) in Table S1.1 as opposed to (3) in Table S1.2 in Supplementary Materials 1), in order to prevent children from being confused by the occurrence of a pronoun that was not preceded by an antecedent in a previous sentence. Likewise, among the three relative clauses used across the two types of SRT (with vs. without discourse), the SRT without discourse featured two reversible object relative clauses (e.g., (5) and (10) in Table S1.2) – based on Costa and Guasti (2021), whereas the SRT with discourse only one (i.e., (5) in Table S1.1), which was done to ensure coherence with previous discourse.

The choice of the target structures was based on the criteria introduced in Marinis and Armon-Lotem (2015). In particular, the structures differed from each other with respect to the presence (or absence) of embedding or movement. As a result, all four SRTs included:

- structures involving *no embedding and no movement* (*N*: 13), such as SVO sentences; some of these structures featured a modal or a negation. Some involved only one sentence and some two, as derived by means of coordination. An example of coordination is given in (1) and (2) for German and Italian, respectively (corresponding to (2) in Table S1.1 and (2) in Table S1.3, respectively):

(1)	Auf einmal geht	das Raumschiff kaputt und fällt	in einen Wald.
	at once go.$_{3SG.PRES.}$	the spaceship broken and fall.$_{3SG.PRES.}$	into a forest
(2)	Ad un certo punto	la navicella si rompe e cade	in una foresta.
	at one certain point	the spaceship CL.$_{REFL.}$ break.$_{3SG.PRES.}$ and fall.$_{3SG.PRES.}$	into a forest
	'At one point the spaceship breaks and falls into a forest'		

- structures involving *embedding but no movement* (*N*: 10), such as finite and non-finite complement clauses and adverbial clauses. An example containing both a finite and non-finite complement clause is given in (3) and (4) for German and Italian, respectively (corresponding to (8) in Table S.1.1 and (8) in Table S.1.3, respectively):

(3)	Der Alien erzählt ihm,	dass das Raumschiff	aufhörte	zu funktionieren.
	the alien tell.3SG.PRES. he.DAT.3SG.	that the spaceship	cease.3SG.PAST	to work
(4)	L'alieno gli dice	che la navicella	ha smesso	di funzionare.
	The alien CL.DAT.3SG. tell.3SG.PRES.	that the spaceship	AUX.3SG.PRES. ceased	to work
	'The alien tells him that the spaceship ceased to work'			

- structures involving *movement but no embedding* (*N*: 8), such as topicalizations, object *wh*-questions, passive sentences. An example of topicalization is given in (5) and (6) for German and Italian, respectively (corresponding to (14) in Table S.1.1 and (14) in Table S.1.3, respectively):

(5)	Den Motor	hat	der Biber	nicht geschafft	zu reparieren.
	the motor.ACC.SG.	AUX.3SG.PRES.	the beaver.NOM.SG.	NEG. managed	to repair
(6)	Ma il motore,	il castoro	non riesce	ad accenderlo.	
	but the motor	the beaver	NEG. manage.3SG.PRES.	to activate-CL.ACC.MASC.SG.	
	'The motor, the beaver did not manage to repair (it)'				

- structures involving *embedding and movement* (*N*: 10), such as subject and object relative clauses and *wh*-complement clauses. An example of an object relative clause is given in (7) and (8) for German and Italian, respectively (corresponding to (5) in Table S.1.1 and (5) in Table S.1.3, respectively):

(7)	Dann kommt	ein Biber, den	ein Jäger	im Wald	verfolgt.
	then come.3SG.PRES.	a beaver REL.ACC.MASC.SG.	a hunter	in the forest	follow.3SG.PRES.
(8)	Arriva	un castoro che	un cacciatore	sta inseguendo	nella foresta.
	come.3SG.PRES.	a beaver that	a hunter	is follow.GERUND	in the forest
	'Then comes a beaver that a hunter is following in the forest'				

Overall, each SRT consisted of 28 sentences targeting 41 different structures, since one sentence could target more than one structure. The classification of structures presented above ensured comparability across languages: the structures at issue in the analysis are derived in the same way via movement and/or embedding in Italian and German (Marinis & Armon-Lotem, 2015 for similar considerations). However, it should be noted that some structures showed crosslinguistic differences between German and Italian. For example, the German topicalization in (5) involves movement of an accusative-marked constituent. Its counterpart in Italian involves movement of the object constituent – which is not case-marked – and resumption by means of the clitic pronoun *lo* (it) in sentence-internal position (Rizzi, 1997). Likewise, object relative clauses are marked by an accusative-marked relative pronoun in German (as shown in (7)), whereas their counterparts in Italian feature the complementizer *che* (that) which is not case-marked – as shown in (8). To enhance comparability across tasks, we tried to match the sentences for number of syllables across tasks (without vs. with discourse) and languages (German and Italian). The two Italian versions did not differ from each other in number of syllables: with discourse ($M = 18.07$, SD: 2.43) vs. without discourse (M: 18.75, SD: 1.80); paired *t*-test: $t(27) = -1.62$, $p = .12$. The same held for the two German versions: with discourse ($M = 17.71$, SD: 3.18) vs. without discourse (M: 16.64, SD: 2.70); paired *t*-test: $t(27) = 1.71$, $p = .10$. Only the German version without discourse differed from both Italian versions, the one with discourse ($t(27) = 2.34$, $p = .03$) and the one without discourse ($t(27) = 4.31$, $p < .001$). We refer to Supplementary Materials 1 for a complete list of the target structures included in each SRT. The four SRTs are available through a public OSF profile (https://osf.io/dsf3w/?view_only=47e7281832b44459a636f67a433fbad6).

For the administration of the task, we followed the guidelines in Marinis and Armon-Lotem (2015). We administered the tasks as a series of Power Point slides. The sentences as well as the task instructions had been pre-recorded by a German and an Italian female native speaker, respectively. In the SRT with discourse, the children were told that they were going to hear a story about an alien and a beaver. They had to repeat each sentence of the story as accurately as possible in order to know how the story went. After repeating each sentence, the children were shown the corresponding picture on the next slide (see Fig. S1.1 in Supplementary Materials 1). The SRT without discourse was administered in a board-game format. By repeating each sentence as accurately as possible, the children could go forward one space in the game (following the design by Costa & Guasti, 2021). In both tasks, the children received positive feedback (such as "well done") and were shown the picture on the next slide or went forward one space, independently of whether they repeated the target sentence correctly or not. They could listen to each sentence only once. If they were not able to repeat anything, they were told not to worry and were shown the next picture or could advance to the next sentence anyway. Right after the instructions and before the actual task, there was a practice session in which the children could ask questions and the experimenter provide

feedback.

3.2.4. Administration

The children were tested in two different sessions with at least one-week interval. In each session, they were administered a SRT with discourse in one language and a SRT without discourse in the other language. This was done because we did not want to burden the children with two SRTs in the same language or with the same SRT in two different languages in one session. The order of administration of the four tasks was counterbalanced across sessions. Likewise, within each session, the order of administration between the two languages was counterbalanced. In one session, they were tested for vocabulary in Italian and in the other session, for vocabulary in German, counterbalancing the order across participants. The tests were administered by a native speaker of Italian and second language learner of German in the session in which the vocabulary test in Italian was administered and a native speaker of German and second language learner of Italian in the session with the vocabulary test in German. The administration of the different SRTs proceeded by using the corresponding Power Point presentations.

3.3. Data analysis

3.3.1. Analysis of the vocabulary task and the background questionnaires

We calculated a vocabulary score for Italian and German, giving 1 point to each item that the children were able to name directly or with the help of a semantic cue and 0.5 points to each item named after a phonemic cue. We gave 0 points to the items that the children could not name or named incorrectly. Both the Italian and German scores were expressed in proportion relative to the maximum score of the corresponding task for the sake of comparability, since the two tests were based on a different number of items. Then, we subtracted the score in German from the score in Italian, whereby a negative score indicated dominance in German and a positive score indicated dominance in Italian. In this way, each child received a dominance score in vocabulary. Furthermore, we calculated a dominance score for each component of the background questionnaire (home language history, early literacy practices, current language use and current literacy practices), following the procedure described in Supplementary Materials 2.

We conducted an exploratory factor analysis to reduce the number of variables related to children's dominance in language proficiency and experience. The analysis was conducted on five variables: dominance in vocabulary, home language history, early literacy practices, current language use and current literacy practices. The factor analysis involved three steps. First, we checked the assumptions for conducting a factor analysis using the Kaiser-Meyer-Olkin test (taking 0.50 as the threshold for suitability of the data for factor analysis, based on Kaiser & Rice, 1974) and the Barlett's Test of Sphericity. Then, we established the optimal number of factors to retain in the analysis by visualizing a scree plot. For each factor, we considered the loading of each variable associated with it and calculated an index. This index was the weighted sum of the values related to the variables at issue, where the weights (w) were the loadings of the respective variables. In order to calculate the dominance score for each child, we used the following formula (see Listanti & Torregrossa, 2023 for methodology):

Dominance score = ($w_{\text{vocabulary}}$ * differential_score_vocabulary) + ($w_{\text{home\_language\_history}}$ * differential_score_home_language_history) + ($w_{\text{early\_literacy}}$ * differential_score_early_literacy) + ($w_{\text{current\_literacy}}$ * differential_score_current_literacy) + ($w_{\text{current\_language\_use}}$ * differential_score_current_language_use)

3.3.2. Analysis of the sentence repetition task

Children's answers in Italian and German were transcribed by an Italian and a German native speaker respectively (i.e., the second author of this study and a student assistant). Each answer was coded based on whether or not children were able to reproduce the target structure. We assigned 1 point to accurate reproductions of the target structures without taking into account whether the children changed some words or whether or not the whole target sentence was grammatical. By contrast, we gave 0 points if the children were inaccurate in the reproduction of the target structure or substituted it, by producing, for instance, an active sentence instead of a passive one (Marinis & Armon-Lotem, 2015, for this methodology). As explained in Section 1.1, this way of coding taps directly into children's grammar abilities and is less sensitive to children's vocabulary knowledge and memory capacity than other coding methodologies focusing on verbatim repetitions or word omissions, substitutions or additions.

3.3.3. Statistical analysis

We fitted a series of generalized linear mixed-effects models, adding predictors progressively according to the research questions formulated in Section 2. In each model, reproduction accuracy (0 vs. 1) was considered as the dependent variable. In the first model, we used type of task (with discourse vs. without discourse) as fixed effect, choosing the SRTs without discourse as the reference level. In the second model, we considered the interaction between type of SRT (with vs. without discourse) and language (Italian vs. German) as predictor, with the SRT without discourse and Italian being the reference levels of the corresponding variables. In the third model, we used the interaction between type of SRT (with vs. without discourse), language (Italian vs. German) and dominance score as predictor. In this model, we also included children's age as fixed effect, since their age range was relatively wide. The values related to age were mean centered. Finally, in the fourth model, we considered the interaction between type of SRT (with vs. without discourse) and type of structure (no embedding-no movement vs. embedding vs. movement vs. movement and embedding) as predictor, with the SRT without discourse and structures exhibiting no embedding and no movement as the reference levels of the corresponding variables. We

fit all models with random intercepts for participants and items and a by-participant random slope for type of SRT. We used the function emmeans in the emmeans R package (Lenth, 2023) to identify relevant pairwise contrasts whenever needed.

4. Results

4.1. Vocabulary task and background questionnaires

The factor analysis was conducted on five variables. We refer to Table 1 for the descriptive statistics of these variables. In Table S3.1 in Supplementary Materials 3, we report the correlation matrix between the five dominance scores reported in Table 1.

Concerning their language experience, the children were dominant in Italian in their home language history and early literacy practices (positive scores in Table 1), and dominant in German in current literacy practices (negative scores). They were slightly dominant in German in their current language use. These results are not surprising since children usually experience a shift in dominance from the heritage to the societal language throughout their lifespan (after entering school, in particular; Caloi & Torregrossa, 2021). The vocabulary scores indicated that overall, the children tended to be German dominant.

The Kaiser-Meyer Olkin for the correlation matrix in Table S3.1 is 0.87. The Barlett's Test of Sphericity was significant ($\chi^2(2) = 15.98, p = .003$), indicating that the correlation matrix was not an identity matrix. Both analyses confirmed that the data were suitable for a factor analysis. The scree plot in Fig. S3.2 showed that the optimal number of components to be considered in the factor analysis equaled 1. The loading for *vocabulary score* was 0.97, for *home language history* 0.93, *early literacy practices* 0.81, *current language use* 0.91 and *current literacy practices* 0.83. The mean value for the dominance score was 0.20, which indicated a relatively balanced profile.[2] Finally, to examine the internal consistency of the variables at issue as measures of language dominance, we conducted a Cronbach's alpha reliability analysis on the six variables of the above analysis (including the final dominance score) using the *alpha ()* function of the 'psych' package in R (Revelle, 2024). The alpha test indicated a high internal consistency (α = 0.85).

4.2. Sentence repetition task with and without discourse

The first research question (RQ1) of the study concerned whether children exhibited overall better grammar abilities in the SRT with discourse than in the SRT without discourse. The *glmer*-analysis revealed a significant effect of type of SRT: The children tended to be more accurate in the SRT with discourse than in the SRT without discourse ($\beta = 0.34, SE = 0.10, z = 3.54, p < .001$).[3]

Research Question 2 considered how far the positive effects of discourse differed depending on whether the SRT was conducted in the heritage or in the societal language. The bar graph in Fig. 1 reports the mean proportions of participants' accuracy (and ± 1.5 standard error) in the two versions of the SRT (with and without discourse) across the two languages (Italian and German). The mean proportion of accuracy was: 0.73 ($SD = 0.44$) for the Italian task with discourse, 0.65 ($SD = 0.48$) for the Italian task without discourse, 0.81 ($SD = 0.39$) for the German task with discourse and 0.81 ($SD = 0.39$) for the German task without discourse.

The results of the generalized linear mixed-effects model reported in Table 2[4] revealed a significant lower-order effect of type of SRT: in Italian, children tended to be more accurate in the SRT with discourse than in the SRT without discourse (positive estimate). The lower-order effect of language indicated that children were more accurate in the German SRT without discourse than in the Italian SRT without discourse (positive estimate). Finally, the significant interaction between type of SRT and language indicated that the positive effect of discourse tended to be less visible (negative estimate) in German than in Italian (see also Fig. 1). This interpretation is also confirmed by the emmeans-analysis, which showed a significant difference between the SRT with and without discourse in Italian ($\beta = -0.50, SE = 0.11, z = -4.43, p < .001$) but not in German ($\beta = -0.12, SE = 0.13, z = -0.96, p = .34$).

Research Question 3 concerned whether the interaction between type of SRT (with and without discourse) and language (Italian vs. German) was modulated by the children's dominance score.[5] In other words, we aimed at finding out how far the effect of discourse on performance in the SRT in a language varied based on children's dominance in one or the other language. The results of the corresponding generalized linear-mixed effects model are reported in Table 3. We found a significant lower-order effect of type of SRT (with discourse) and language (German) as well as a significant interaction between type of SRT and language. These results are consistent with the results of the previous model. There was also a significant lower-order effect of dominance: children who were more dominant in Italian tended to be more accurate in the Italian version of the SRT without discourse (the reference level of the corresponding variable; positive estimate). The model revealed also a significant interaction between dominance score and language (German), which suggests that in the German SRT without discourse, the children who were more dominant in Italian exhibited lower accuracy rates (negative estimate). Finally, we found no evidence that in the two languages, the effect of dominance on accuracy changed across the

---

[2] Overall, it seems that the simultaneous bilinguals (n = 13) tended to be more dominant in German than the early and late sequential bilinguals (n = 10) – see Section 3.1 for this distinction. The simultaneous bilinguals had a mean dominance score of −1.05 (SD = 1.51), whereas the sequential bilinguals a mean dominance score of 1.83 (SD = 1.08).

[3] The resulting R model was: m1 <- glmer (accuracy ~ 1 + type of SRT + (1+type of SRT|CH) + (1|item), data = SRT, family = binomial, control = glmerControl(optimizer = "bobyqa")). The statistics for the intercept was: $\beta = 1.26, SE = .23, z = 5.39, p < .001$.

[4] The corresponding R model was: m2 <- glmer (accuracy ~ 1 + type of SRT * language + (1+type of SRT|CH) + (1|item), data = SRT, family = binomial, control = glmerControl(optimizer = "bobyqa")).

[5] The corresponding R model was: m3 <- glmer (accuracy ~ 1 + type of SRT * language * dominance score + age + (1+type of SRT|CH) + (1|item), data = SRT, family = binomial, control = glmerControl(optimizer = "bobyqa"))

Table 1

Descriptive statistics for the dominance-related variables included in the exploratory factor analysis and the final dominance score. The maximum and minimum values of the dominance-related variables could range between 1.00 (exposure to Italian only) and −1.00 (exposure to German only) – see Supplementary Materials 2. The maximum and minimum values for the overall dominance score could range between 4.45 (i.e., indicating exposure only to Italian) and −4.45 (indicating exposure only to German).

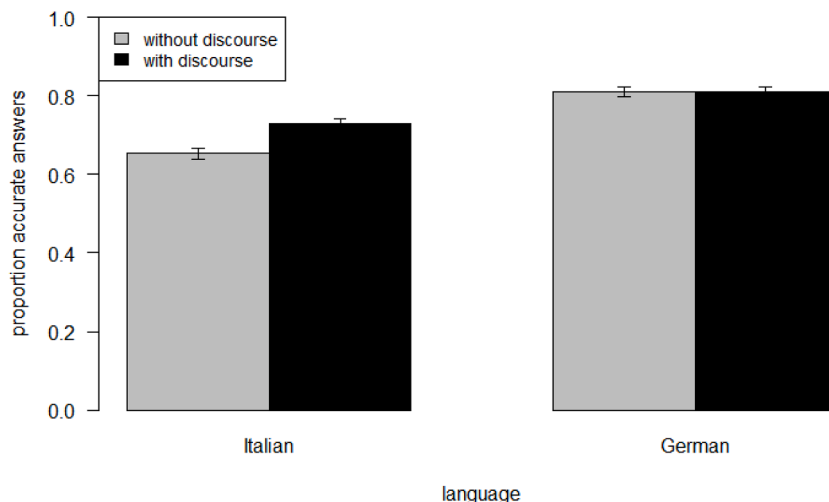|	Mean	SD	Minimum	Maximum
Dominance in vocabulary score	−0.19	.29	−0.77	.35
Dominance in home language history	.28	.53	−0.89	1.00
Dominance in early literacy practices	.45	.67	−1.00	1.00
Dominance in current language use	−0.03	.62	−1.00	1.00
Dominance in current literacy practices	−0.28	.40	−1.00	.52
Dominance score	.20	1.97	−4.12	3.36

Fig. 1. Mean proportions of accuracy (bar plots) and ± 1.5 standard errors across SRT (with or without discourse) and language (Italian and German).

Table 2

Parameters of the generalized linear mixed-effects analysis concerning the accuracy of target structure reproduction across SRT-type (with or without discourse) and language (Italian and German).

Fixed effects	B	SE	95% CI	z	p
Intercept	.84	.24	[0.37, 1.33]	3.52	<0.001
Type of SRT (discourse)	.50	.11	[0.28, 0.73]	4.43	<0.001
Language (German)	.96	.11	[0.74, 1.17]	8.78	<0.001
Type of SRT * Language	−0.37	.16	[−0.68, −0.07]	−2.41	.02

Table 3

Parameters of the generalized linear mixed-effects analysis concerning the accuracy of target structure reproduction as a function of SRT-type (with or without discourse), language (Italian and German), dominance score and age.

Fixed effects	β	SE	95% CI	Z	P
Intercept	1.12	0.28	[0.57,1.66]	4.02	<0.001
Type of SRT (discourse)	0.57	0.15	[0.27, 0.88]	3.71	<0.001
Language (German)	1.09	0.16	[0.79, 1.40]	6.99	<0.001
Dominance score	0.62	0.12	[0.37, 0.86]	4.97	<0.001
Age	0.27	0.23	[−1.18, 0.71]	1.18	.24
Type of SRT (discourse) * Language (German)	−0.67	0.21	[−1.09, −0.25]	−3.13	.002
Type of SRT (discourse) * Dominance score	−0.0003	0.08	[−0.16, 0.16]	−0.004	1.00
Language (German) * Dominance score	−1.33	0.09	[−1.50, −1.16]	−15.16	<0.001
Type of SRT (discourse) * Language (German) * Dominance score	0.10	0.11	[−0.12, 0.32]	0.87	.38

two types of SRT (with or without discourse). We also found no effect of age.

Fig. 2 plots the predicted probability of reproducing a target structure accurately as a function of dominance. Higher dominance scores (on the right) indicate dominance in Italian, whereas lower dominance scores (on the left) dominance in German. In the figure, the two conditions (type of SRT with vs. without discourse) were merged, since we did not find any interaction between type of SRT, language and dominance score. It should be noticed that Italian-dominant children (on the right of the figure) are more accurate in German than German-dominant children (on the left of the figure) are in Italian.

Research Question 4 was related to whether the effect of discourse on children's response accuracy was modulated by the type of target structure, being mostly visible in association with structures which are more sensitive to discourse.[6] The results of the generalized linear mixed-effects model reported in Table 4 revealed a marginal lower-order effect of type of SRT, whereby with structures involving no movement and no embedding, the children were slightly more accurate in the SRT with discourse than in the SRT without discourse. There was a lower-order effect of type of structure in association with structures involving movement and structures involving both embedding and movement in the SRT without discourse (the reference level of the corresponding variable). However, we did not find any interaction between type of SRT (with or without discourse) and type of structure. This indicates that children exhibited lower response accuracy with structures involving movement or structures involving movement and embedding independently of whether these structures occurred in a SRT with or without discourse.

5. Discussion

This study investigated how far discourse enhanced the performance of bilingual children when they completed a SRT in their heritage and societal language. It also addressed the extent to which language dominance modulated the effect of discourse in each language and whether the effect of discourse was visible with all structures tested or only with structures that are more sensitive to discourse constraints.

The first result of the study was that the bilingual children who took part in the study showed a better performance in the SRT with discourse than the SRT without discourse. The positive effect of discourse was visible only in the heritage language Italian, in line with our predictions related to the second research question (Section 2). This was the language to which the children were less exposed at school (see Section 1.1 on the effects of literacy on children's performance in a SRT). Italian was also the language in which the children, as a group, tended to be less dominant in vocabulary (see, e.g., the results related to the vocabulary scores reported in Table 1). As a result, processing in this language may be less efficient than processing in the dominant language at school and in the society (German). The facilitative role of discourse in the less dominant language may be related to the fact that discourse provides children with more cues for sentence comprehension and interpretation compared to a format in which the target sentences are not connected to each other. The repetition of certain words as well as the possibility to build expectations about upcoming referents and events may make the task less effortful. Therefore, we found no evidence that the children had difficulty to integrate grammar and discourse information while performing the task. Rather, the SRT with discourse seemed to facilitate the comprehension and reproduction of the target sentences, although the children had to maintain and update information as the story unfolded, which, in principle, added complexity to the task (Section 1.2).

Furthermore, we would like to suggest that the facilitative role of discourse for children's performance in their non-dominant language was related to the nature of the task: when performing the SRT with discourse, children were engaged in a goal-oriented, playful activity. By repeating each target sentence, they could advance in the story plot. Furthermore, the SRT with discourse was embedded in an activity which was familiar to the children, who are usually engaged in story telling at home and school. On top of this, in the SRT with discourse, the reproduction of each sentence was followed by the visualization of the corresponding picture. As a result, children's comprehension and production relied on several semiotic resources at the same time (visual, oral, etc.), which added to the ecological validity of the SRT with discourse. Crucially, we did not find a positive effect of discourse in the dominant language because children's ability in it were more stable and, hence, less likely to be vulnerable to external factors, such as the presence of discourse and the ecological validity of the task. Discourse may benefit children's performance in a SRT up to a certain threshold of grammar abilities. The children of the present study were likely below this threshold in the heritage language (Italian), but above it in the societal language (German).

In Section 1.1, we reported that many SRTs have been designed to assess DLD among monolingual and bilingual children. The present study was not designed to be used with bilinguals with DLD. However, its use could be extended to children with atypical language development, in order to investigate how far they are able to integrate discourse and grammar information in the target language. Children with DLD or other atypically developing children may benefit from discourse to the same extent as bilingual children in their non-dominant language, as shown in the present study. In this sense, the introduction of discourse may be seen as an accommodation of existing SRTs, which facilitates performance without affecting the construct (see Kormos & Taylor, 2020 for a discussion of the concept of accommodation in language testing). The construct remains identical because both versions of the SRT (with and without discourse) tapped into children's grammar abilities, as operationalized in terms of their ability to reproduce target structures. We return to the issue of construct validity below, when discussing the effect of type of structure on children's performance.

Another result that emerged from our analysis is the absence of any interaction between type of SRT (with or without discourse), language (heritage vs. societal) and children's dominance score. In other words, children's dominance in the heritage or societal

---

[6] The corresponding R model was: m4 <- glmer (accuracy ~ 1 + type of SRT * type_structure + (1+type of SRT|CH) + (1|item), data = SRT, family = binomial, control = glmerControl(optimizer = "bobyqa"))
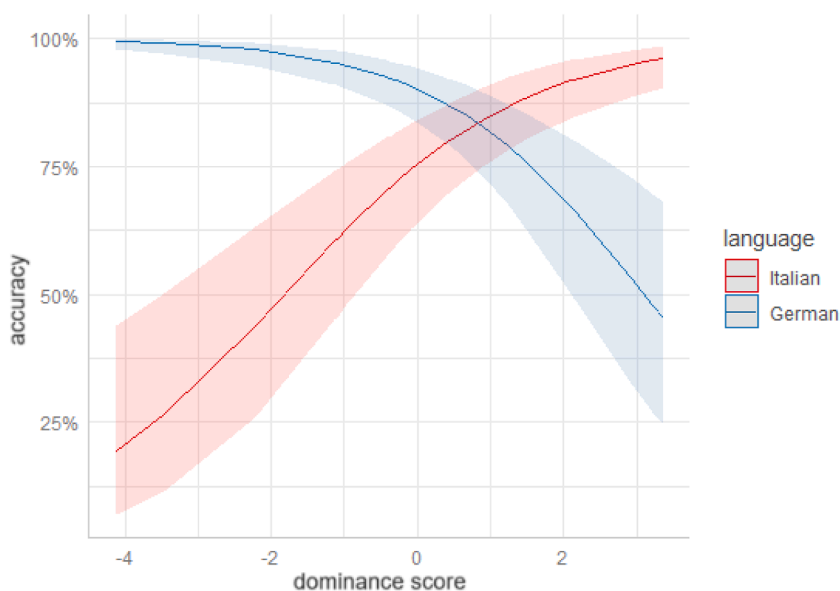
Fig. 2. Predicted probability of accurate reproduction of a target structure as a function of dominance. We merged the data from the two types of SRT (with and without discourse). Higher dominance scores (on the right) indicate dominance in Italian, whereas lower dominance scores (on the left) dominance in German. The shaded lines correspond to a 95 % confidence interval. The predicted probabilities were derived by using the ggpredict() function in the 'ggeffects' package (Lüdecke, 2018).

Table 4

Parameters of the generalized linear mixed-effects analysis concerning the accuracy of target structure reproduction across SRT-type (with or without discourse) and type of target structure.

Fixed effects	B	SE	95% CI	z	p
Intercept	1.64	0.28	[1.09, 2.20]	5.78	<0.001
Type of SRT (discourse)	0.29	0.16	[−0.01, 0.60]	1.88	.06
Type of structure (embedding)	−0.14	0.31	[−0.75, 0.46]	−0.47	.64
Type of structure (movement)	−0.66	0.33	[−1.30, - 0.02]	−2.02	.04
Type of structure (embedding & movement)	−0.89	0.31	[−1.49, −0.29]	−2.91	.004
Type of SRT (discourse) * Type of structure (embedding)	.07	.21	[−0.35, 0.49]	.33	.74
Type of SRT (discourse) * Type of structure (movement)	.03	.21	[−0.39, 0.45]	.16	.87
Type of SRT (discourse) * Type of structure (embedding & movement)	.09	.20	[−0.30, 0.49]	.46	.64

language did not affect the beneficial effect of discourse on children's performance. As a measure of dominance, we relied on a complex score which encompassed a language proficiency component (vocabulary score) and several measures of language exposure, which were combined together through a factor analysis (Section 3.3). We expected children who were less dominant in the target language to benefit more from discourse than children who were more dominant in it, based on the above considerations related to the effects of discourse on children's performance in the heritage language. Contrary to our expectations, we only observed that children's performance in the SRT was affected by language dominance independently of the nature of the task (with or without discourse; Fig. 2). This result can be interpreted as showing that language dominance affected children's grammar abilities across the board. This conclusion is in line with previous studies showing that bilingual children's performance in SRTs is sensitive to their degree of dominance in the target language (Andreou et al., 2021; Pratt et al.; 2021 and references in Section 1.1). In this sense, the role of discourse is not so much related to the improvement of children's grammar abilities, but to a facilitation of processing, as has been discussed above.

In terms of the impact of type of target structure on children's performance in the SRT, we noticed that structures involving movement and structures involving both movement and embedding were the most difficult for the children in the present study. This finding is not surprising given that these structures exhibit a higher degree of syntactic complexity than structures involving only embedding or neither embedding nor movement (e.g., Jakubowicz, 2011). In Section 1.3, we mentioned that different syntactic structures exhibit a different sensitivity to discourse constraints. For example, marked word orders and object relative clauses need to be licensed by an appropriate discourse context. Therefore, we expected the abovementioned positive effect of discourse on children's performance in the SRTs to be modulated by the type of structure. Contrary to this expectation, we found no interaction between type of SRT (with and without discourse) and type of target structure. In other words, discourse seemed to be beneficial across the board for all structures. Albeit not in line with our expectations, this result is relevant for the assessment of the construct validity of the SRT with discourse (see our considerations above). The observation that the impact of the complexity of the target structures did not change

across the two SRTs (with discourse and without discourse) indicates that the construct remained unaltered independently of the presence vs. absence of discourse.

Overall, the results of the study suggest that the SRT with discourse was in general easier for the children. This may be related to several factors beyond the abovementioned ecological validity of these tasks. For example, the SRT with discourse seemed to reduce the processing costs related to shifting between scenarios, whereas in the SRT without discourse each sentence referred to a different situation. Furthermore, the SRT with discourse was likely to reduce the costs of lexical retrieval given that certain words were repeated and other activated by previous discourse (e.g., the semantic relation between spaceship, alien and planet). Finally, in the SRT with discourse, it was easier for children to generate expectations about upcoming events as usually happens during narrative comprehension.

In conclusion, the present study suggested that the SRT with discourse is a promising instrument for the assessment of children's language abilities and may be able to provide a better picture of children's grammar abilities compared to traditional SRTs (without discourse) because they are less vulnerable to other types of processing difficulties (as mentioned above, with reference, for instance, to lexical access). In particular, the addition of a discourse component to SRTs should enhance performance among individuals who exhibit processing limitations for one reason or another. By contrast, the effect of discourse may not be visible with individuals exhibiting difficulties to integrate grammar and discourse information during online processing.

Financial support

CRediT authorship contribution statement

Jacopo Torregrossa: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Andrea Listanti:** Writing – review & editing, Methodology, Investigation, Data curation. **Christiane Bongartz:** Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Visualization, Writing – review & editing. **Theodoros Marinis:** Writing – review & editing, Visualization, Validation, Supervision, Methodology, Investigation, Conceptualization.

Declaration of competing interest

I hereby declare that I have no competing financial and/or non-financial interest that could influence the work reported in this paper. This statement is on behalf of all co-authors of the paper.

Acknowledgments

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.rmal.2024.100107.

References

Alloway, T. P., Gathercole, S. E., Willis, C., & Adams, A. M. (2004). A structural analysis of working memory and related cognitive skills in young children. *Journal of Experimental Child Psychology, 87*(2), 85–106.

Andreou, M., Torregrossa, J., & Bongartz, C. (2021). Sentence repetition task as a measure of language dominance. D. Dionnne & L.A. Vidal Covas (Eds.). In *Proceedings of the 45th annual boston university conference on language development* (pp. 14–25). Cascadilla Press.

Armon-Lotem, S., Walters, J., & Gagarina, N. (2011). The impact of internal and external factors on linguistic performance in the home language and in L2 among Russian-Hebrew and Russian-German preschool children. *Linguistic Approaches to Bilingualism, 1*(3), 291–317.

Brandt, S., Kidd, E., Lieven, E., & Tomasello, M. (2009). The discourse bases of relativization: An investigation of young German and English-speaking children's comprehension of relative clauses. *Cognitive Linguistics, 20*(3), 539–570.

Brouwer, S., Özkan, D., & Küntay, A. C. (2017). Semantic prediction in monolingual and bilingual children. E. Blom, J. Schaeffer, L. Cornips (Eds.). *Cross-linguistic influence in bilingualism* (pp. 49–74). John Benjamins.

Caloi, I., & Torregrossa, J. (2021). Home and school language practices and their effects on heritage language acquisition: A view from heritage Italians in Germany. *Languages, 6*, 50. https://doi.org/10.3390/languages6010050

Chondrogianni, V., & Marinis, T. (2012). Production and processing asymmetries in the acquisition of tense morphology by sequential bilingual children. *Bilingualism: Language and Cognition, 15*(1), 5–21.

Chondrogianni, V., & Marinis, T. (2016). L2 children do not fluctuate. B. Haznedar & F. Nihan Ketrez (Eds.). *The acquisition of Turkish in childhood* (pp. 361–388). John Benjamins.

Costa, F., & Guasti, M. T. (2021). Is bilingual education sustainable? *Sustainability, 13*, 13766. https://doi.org/10.3390/su132413766

De Cat, C. (2020). Predicting language proficiency in bilingual children. *Studies in Second Language Acquisition, 42*(2), 279–325.

De Groot, A. M., & Hagoort, P. (2017). *Research methods in psycholinguistics and the neurobiology of language: A practical guide*. John Wiley & Sons.

Devescovi, A., & Caselli, M. C. (2007). Sentence repetition as a measure of early grammatical development in Italian. *International Journal of Language & Communication Disorders, 42*(2), 187–208.

Fleckstein, A., Prévost, P., Tuller, L., Sizaret, E., & Zebib, R. (2018). How to identify SLI in bilingual children: A study on sentence repetition in French. *Language Acquisition, 25*(1), 85–101.

Grüter, T., & Rohde, H. (2021). Limits on expectation-based processing: Use of grammatical aspect for co-reference in L2. *Applied Psycholinguistics, 42*(1), 51–75.

Hamann, C., & Abed Ibrahim, L. (2017). Methods for identifying specific language impairment in bilingual populations in Germany. *Frontiers in Communication, 2*, 16.

Holleman, G. A., Hooge, I. T., Kemner, C., & Hessels, R. S. (2020). The 'real-world approach'and its problems: A critique of the term ecological validity. *Frontiers in Psychology, 11*, 721.

Jakubowicz, C. (2011). Measuring derivational complexity: New evidence from typically developing and SLI learners of L1 French. *Lingua, 121*(3), 339–351. International Review of general linguistics. Revue internationale de linguistique generale.

Kaiser, E., & Trueswell, J. C. (2004). The role of discourse context in the processing of a flexible word-order language. *Cognition, 94*(2), 113–147.

Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark IV. *Educational and Psychological Measurement, 34*(1), 111–117.

Karaca, F., Brouwer, S., Unsworth, S., & Huettig, F. (2021). Predictions in bilingual children: The missing piece of the puzzle. In E. Kaan, & T. Grüter (Eds.), *Prediction in second language processing and learning* (pp. 116–137). John Benjamins.

Klem, M., Melby-Lervåg, M., Hagtvet, B., Lyster, S. A. H., Gustafsson, J. E., & Hulme, C. (2015). Sentence repetition is a measure of children's language skills rather than working memory limitations. *Developmental Science, 18*(1), 146–154.

Kormos, J., & Taylor, L. B. (2020). Testing the L2 of learners with specific learning differences. Winke, P. & Brunfaut, T. (eds.). *TheRoutledgeHandbook of second language acquisition and language testing*. Routledge.

Lenth, R. V (2023). *EMMEANS: Estimated marginal means, aka least-squares means*. R package version 1.8.9.

Listanti, A., & Torregrossa, J. (2023). The production of preverbal and postverbal subjects by Italian children: Timing of acquisition matters. *First Language, 43*(4), 431–460.

López-Beltrán, P., Johns, M. A., Dussias, P. E., Lozano, C., & Palma, A. (2022). The effects of information structure in the processing of word order variation in the second language. *Second Language Research, 38*(3), 639–670.

Lüdecke. (2018). ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software, 3*(26), 772.

Mak, W. M., Vonk, W., & Schriefers, H. (2008). Discourse structure and relative clause processing. *Memory & Cognition, 36*(1), 170–181.

Marinis, T., & Armon-Lotem, S. (2015). Sentence repetition. S. Armon-Lotem, J. de Jong & N. Meir (Eds.). *Assessing multilingual children: Disentangling bilingualism from language impairment (95–124)*. De Gruyter.

Meir, N., Parshina, O., & Sekerina, I. A. (2020). The interaction of morphological cues in bilingual sentence processing: An eye-tracking study. In *Proceedings of the 44th Annual Boston University Conference on Language Development* (pp. 376–389). Cascadilla Press.

Meir, N., Walters, J., & Armon-Lotem, S. (2016). Disentangling SLI and bilingualism using sentence repetition tasks: The impact of L1 and L2 properties. *International Journal of Bilingualism, 20*(4), 421–452.

Papadopoulou, D., Peristeri, E., Plemenou, E., Marinis, T., & Tsimpli, I. (2015). Pronoun ambiguity resolution in Greek: Evidence from monolingual adults and children. *Lingua, 155*, 98–120. International Review of General linguistics. Revue internationale de linguistique generale.

Petermann, F., Fröhlich, L. P., & Metz, D. (2010). *SET 5–10. sprachstandserhebung für kinder im alter von 5–10 jahren*. Hochgrefe.

Polišenská, K. (2011). *The influence of linguistic structure on memory span: Repetition tasks as a measure of language ability. Phd dissertation*. City University London.

Polišenská, K., Chiat, S., & Roy, P. (2015). Sentence repetition: What does the task measure? *International Journal of Language & Communication Disorders, 50*(1), 106–118.

Pontikas, G., Cunnings, I., & Marinis, T. (2022). Online processing of which-questions in bilingual children: Evidence from eye-tracking. *Journal of Child Language,* 1–37.

Pratt, A. S., Peña, E. D., & Bedore, L. M. (2021). Sentence repetition with bilinguals with and without DLD: Differential effects of memory, vocabulary, and exposure. *Bilingualism: Language and Cognition, 24*(2), 305–318.

Renfrew, C. E. (1995). *Word finding vocabulary test*. Speechmark Publishing.

Revelle, W. (2024). *psych: Procedures for psychological, psychometric, and personality research*. Evanston, Illinois: Northwestern University. R package version 2.4.3.

Rizzi, L. (1997). The fine structure of the left periphery. L. Haegeman (Ed.). *Elements of grammar* (pp. 281–337). Kluwer Academic Publishers.

Schlenter, J. (2022). Prediction in bilingual sentence processing: How prediction differs in a later learned language from a first language. *Bilingualism: Language and Cognition,* 1–15.

Schönström, K., & Hauser, P. C. (2022). The sentence repetition task as a measure of sign language proficiency. *Applied Psycholinguistics, 43*(1), 157–175.

Serratrice, L., & De Cat, C. (2020). Individual differences in the production of referential expressions: The effect of language proficiency, language exposure and executive function in bilingual and monolingual children. *Bilingualism: Language and Cognition, 23*(2), 371–386.

Simon-Cereijido, G., & Méndez, L. I. (2018). Using language-specific and bilingual measures to explore lexical–grammatical links in young Latino dual-language learners. *Language, Speech, and Hearing Services in Schools, 49*(3), 537–550.

Sorace, A. (2011). Pinning down the concept of "interface" in bilingualism. *Linguistic Approaches to Bilingualism, 1*(1), 1–33.

Spada, N., Shiu, J. L. J., & Tomita, Y. (2015). Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies. *Language Learning, 65*(3), 723–751.

Tomita, Y., Suzuki, W., & Jessop, L. (2009). *Elicited imitation: Toward valid procedures to measure implicit second language grammatical knowledge* (pp. 345–350). TESOL Quarterly.

Torregrossa, J., Andreou, M., Bongartz, C., & Tsimpli, I. M. (2021). Bilingual acquisition of reference: The role of language experience, executive functions and cross-linguistic effects. *Bilingualism: Language and Cognition, 24*(4), 694–706.

Torregrossa, J., & Bongartz, C. (2018). Teasing apart the effects of dominance, transfer, and processing in reference production by German–Italian bilingual adolescents. *Languages, 3*(3), 36.

Torregrossa, J., Caloi, I., & Listanti, A. (2023a). The acquisition of syntactic structures in Italian: Assessing the role of language exposure at critical periods. F. Romano (Ed.). *Studies in Italian as a heritage language* (pp. 155–194). De Gruyter Mouton.

Torregrossa, J., Eisenbeiß, S., & Bongartz, C. (2022). *Boosting bilingual metalinguistic awareness under dual language activation: Some implications for bilingual education.* Language Learning. https://doi.org/10.1111/lang.12552. published online.

Torregrossa, J., Flores, C., & Rinke, E. (2023b). What modulates the acquisition of difficult structures in a language? A study on Portuguese in contact with French, German and Italian. *Bilingualism: Language and Cognition, 26*(1), 179–192.

Trofimovich, P., Lightbown, P. M., Halter, R. H., & Song, H. (2009). Comprehension-based practice: The development of L2 pronunciation in a listening and reading program. *Studies in Second Language Acquisition, 31*(4), 609–639.

Tsimpli, I. M. (2014). Early, late or very late?: Timing acquisition and bilingualism. *Linguistic Approaches to Bilingualism, 4*(3), 283–313.

Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics, 12*(1), 54–73. https://doi.org/10.1111/1473-4192.00024

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing, 33*(4), 497–528.

Yano, M., & Koizumi, M. (2018). Processing of non-canonical word orders in (in)felicitous contexts: Evidence from event-related brain potentials. Language. *Cognition and Neuroscience, 33*(10), 1340–1354.