

Research Article

Cite this article: Lam, H.B.-L., Paradis, J., Soto-Corominas, A., Al-Janaideh, R., Chen, X. and Gottardo, A. (2025). The longitudinal development of L2 complex syntax in Arabic-English refugee children: sources of individual differences and comparison of measures of syntax. *Bilingualism: Language and Cognition* 1–14. <https://doi.org/10.1017/S1366728925100862>

Received: 15 November 2024

Revised: 12 August 2025


Accepted: 23 October 2025

Keywords:

child-L2; refugees; narratives; complex syntax; individual differences

Corresponding author:

Hannah Bou-Lai Lam;
Email: hblam@ualberta.ca


 This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.



CAMBRIDGE
UNIVERSITY PRESS

The longitudinal development of L2 complex syntax in Arabic-English refugee children: sources of individual differences and comparison of measures of syntax

Hannah Bou-Lai Lam¹ , Johanne Paradis¹, Adriana Soto-Corominas², Redab Al-Janaideh³, Xi Chen³ and Alexandra Gottardo⁴

¹Department of Linguistics, University of Alberta, Canada; ²Department of English and German Studies, Universitat Autònoma de Barcelona, Spain; ³University of Toronto, Canada and ⁴Wilfrid Laurier University, Canada

Abstract

We examined the growth of English-L2 clausal density (CD) in narrative language samples from 129 school-age Syrian refugee children during their first 5 years of residency in Canada. First, we found that CD showed unique developmental trajectories from MLUw, and relatively rapid acquisition, consistent with studies with non-refugee participants. Second, faster growth in CD was associated with superior cognitive abilities and higher maternal education. An older-age advantage was found at Time 1, but a younger-age advantage emerged across Time 2–3. Factors more specific to the refugee experience (time in refugee camps and wellbeing difficulties) also predicted variance in CD and MLUw development but to a lesser extent. Finally, modeling performance on sentence repetition tasks revealed stronger contributions of lexical diversity and MLUw than CD. We conclude that complex syntax is relatively resilient in the L2 acquisition of refugee children and that CD in naturalistic production and SRT capture different abilities.

Highlights

- Growth in general and complex syntax observed L2-English of Arabic-L1 refugee youth.
- General and complex syntax show different developmental trajectories.
- Age, maternal education, cognition and refugee-specific factors affect syntax.
- General syntax and lexical diversity predict performance in sentence repetition.
- Complex syntax in L2-acquisition is resilient for vulnerable language learners.

1. Introduction

First-generation migrant children who enter the school system in a host country at various ages face many challenges, such as navigating acculturation transitions, coping with new and sometimes advanced academic content and acquiring a new language, all at the same time. Regarding first-generation refugee children, adverse pre-migration experiences like deprivation, violence and interrupted education, as well as ongoing difficulties in mental health and well-being post-migration, exacerbate such challenges, and therefore, could impact their educational outcomes and social inclusion (Browder, 2018; Fazel & Stein, 2002; Graham et al., 2016; Paradis, Soto-Corominas, et al., 2022; Sirin & Rogers-Sirin, 2015). To date, there is limited research examining whether pre- and post-migration adversity has an impact on refugee children's acquisition of their second language (L2), separately from their academic progress or general learning capacities.

Bilingual children's L2 abilities are foundational for their academic success (National Academy of Sciences, Engineering, and Medicine, 2017; Picot & Sweetman, 2012; Strand & Demie, 2005). Moreover, because production and comprehension of multiclausal or complex sentences are associated with reading development in monolinguals (MacKay et al., 2021; Scott, 2009), acquiring complex syntax in the language of schooling is important for first-generation bilinguals. Nevertheless, research on bilingual children's acquisition of L2 complex syntax in naturalistic production is scarce, with few studies that are longitudinal, and none we are aware of with children from refugee backgrounds. Existing research suggests that bilingual children use complex sentences frequently and accurately relatively early in their L2 development, in contrast with morphology (Paradis et al., 2017; Scheidnes & Tuller, 2019). Therefore, on one hand, it is possible that complex syntax is a relative strength in the L2 of all bilingual children, including those from refugee backgrounds, which would be beneficial for reading and academic learning in their L2. On the other hand, premigration adversity and current wellbeing could modulate individual refugee children's development of L2 complex syntax. Thus, the primary objective of

this study was to examine the use of complex sentences in English language samples from school-age, Syrian refugee children during their first 5 years of residency in Canada. We examined change over time and the sources of individual differences in children's development to investigate the extent to which productive L2 complex syntax is sensitive to extra-linguistic factors, and in particular, factors beyond age and L2 input which are expected to be associated with L2 development in all bilingual children.

The use of sentence repetition tasks (SRT) to measure bilingual children's syntactic abilities has become increasingly common for both research and clinical purposes (Armon-Lotem & Meir, 2016; Hamann et al., 2020; Kaltsa et al., 2020; Meir, 2017; Paradis et al., 2021; Soto-Corominas et al., 2022; Zebib et al., 2020). SRTs are a more practical alternative to estimate syntactic abilities than language sampling, since they are quicker to administer and less time-consuming to score than a language sample. However, there is a lack of research on the relationship between SRT performance and naturalistic production of syntax in bilingual populations; little is known about how SRT are indexing similar L2 abilities to those exhibited on a more ecologically valid task like storytelling. Toward this end, a secondary objective of this study was to compare measures of syntax from language samples with SRT scores from the same participants over time to ascertain their degree of overlap.

1.1. Measures of syntax in naturalistic production data

Mean length of utterance in words (MLUw) and clausal density (CD) are both used to measure syntactic development, but they do not measure the same (morpho)syntactic abilities. MLUw is calculated by dividing the total number of words by the number of utterances, while CD is calculated by dividing the total number of clauses by the number of utterances. MLUw growth over time indexes clause expansion that includes the addition of adverbs, prepositional phrases, expanded noun phrases with adjectives, as well as coordinated and embedded clauses but it does not specifically measure the use of multiple clauses (Cahill et al., 2020; Castilla-Earls et al., 2021; Frizelle et al., 2018). In contrast, CD growth over time indexes clause expansion only with respect to multi-clausal sentences, such as embedded and coordinated clauses. Therefore, MLUw is more a measure of general syntax, while CD is a specific measure of complex syntax. Furthermore, MLUw can be calculated automatically in language samples using programs like SALT or CLAN (Miller & Iglesias, 2012; MacWhinney, 2000); whereas, calculating CD requires more fine-grained coding, or minimally, could be estimated through searches for coordinating and subordinating connectives (conjunctions, relative pronouns, etc.) in transcripts (Squires et al., 2014). Automated systems for analyzing syntactic structures in transcripts are still in development (Agmon et al., 2024). Therefore, the automated calculation of MLUw makes it a more accessible and practical measure of syntax than CD.

MLUw is a long-standing and widely used measure of early general syntactic growth in preschool children (Brown, 1973; Miller & Chapman, 1981); however, MLUw still grows past the early years (Frizelle et al., 2018; Nippold et al., 2008). Frizelle et al. (2018) found that CD and MLUw are not identical measures of syntactic development in school-age monolinguals. Some studies with bilinguals also suggest that MLUw and CD are not identical measures of L2 syntax (Castilla-Earls et al., 2021; Paradis & Kirova, 2014; Paradis, Sorenson Duncan, et al., 2022; except see Cahill et al., 2020). For example, MLUw and CD differentially predict error types and rates within L2 utterances (Castilla-Earls et al., 2021), L2 children converge with monolingual norms more rapidly for CD than MLUw

(Paradis & Kirova, 2014), and CD discriminates between bilinguals with typical and atypical development better than MLUw (Paradis, Sorenson Duncan, et al., 2022). Therefore, in this study, we compared developmental trajectories of CD and MLUw and how these are modulated by individual difference factors.

1.2. Measuring syntax using SRT

The LITMUS English SRT, adapted for Canadian English and used in the present study, includes several computationally complex structures, such as sentences with subordination and relative clauses, *wh*-object questions and passive sentences, and thus, is designed to implicate children's morphosyntactic knowledge in their performance (Marinis & Armon-Lotem, 2015). For example, more complex targets are expected to provoke more errors in repetition compared to simpler structures because they involve more resources in processing and production (Marinis & Armon-Lotem, 2015). Children's short-term and working memory skills contribute to their SRT performance (Hamann et al., 2020; Polišenská et al., 2015; Pratt et al., 2021); nevertheless, SRTs are considered to be primarily measures of *linguistic* ability in typically developing bilinguals and monolinguals, encompassing phonological, lexical, morphosyntactic and speech production abilities (Hamann et al., 2020; Polišenská et al., 2015; Pratt et al., 2021; Zebib et al., 2020). To assess relative contributions to SRT performance, previous studies have used performance on independent tests of lexical and morphosyntactic abilities as predictors of variance in SRT performance (Hamann et al., 2020; Polišenská et al., 2015; Zebib et al., 2020). Studies to date have not compared SRT performance with morphosyntactic measures from *naturalistic* production, like MLUw and CD. Therefore, it remains unknown how SRT performance and MLUw and CD are measuring similar constructs. If they do measure similar constructs, this would constitute evidence for SRT being reasonable estimates of spontaneous, expressive morphosyntactic abilities. Toward this end, the second aim of this study was to examine the associations between SRT performance on one hand, MLUw, CD and lexical diversity from language samples on the other hand, in the same bilingual participants over time.

2. L2 complex sentences in bilingual language development

Studies point to production of complex sentences emerging early in child English-L2 development. Paradis and Kirova (2014) found that 4- to 5-year-old bilingual children had a mean CD score on an English narrative task within the normal range for monolinguals, while their mean MLUw was below this range. Similarly, a study using the same narrative task with English-dominant bilingual 4-year-olds with developmental language disorder found no differences between their complex syntax scores and those of monolinguals with developmental language disorder (Cleave et al., 2010). Paradis et al. (2017) examined complex sentence use in language samples from 187 5- to 7-year-old bilinguals with 17 months of English-L2 exposure. Despite being in the early stages of L2 acquisition, 18% of the children's sentences were multiclausal; this proportion exceeds what has been reported for 4-year-old monolinguals and is similar to same-age monolinguals (Frizelle et al., 2018). Furthermore, all bilingual children in Paradis et al. (2017) used different types of complex sentences, including relative clause sentences, and only 4% of all their complex sentences contained errors with word order or connectives. Scheidnes and Tuller (2019)'s study of complex syntax in spontaneous speech included

school-age English-L1, French-L2 children in France, 6–12 years old, along with French-speaking monolinguals the same age, and younger French monolinguals. They found that the French-L2 children had similar CD scores to their monolingual age-peers and higher scores than those of younger monolinguals. Degree of convergence with monolinguals in the use of complex syntax increased with L2 exposure. Importantly, Scheidnes and Tuller (2019) found that, in contrast with CD, the bilingual children were dissimilar and lagging behind their monolingual peers in skills with grammatical morphology. Like Scheidnes and Tuller (2019), Cahill et al. (2020) found no differences in MLUw or CD from a narrative task between French-English bilinguals and English monolinguals, 7–8 and 11–12 years old. Overall, existing studies suggest that complex sentences emerge relatively early in L2 development, but there are limitations in this body of research due to small sample sizes (e.g., Cahill et al., 2020; Cleave et al., 2010; Paradis & Kirova, 2014; Scheidnes & Tuller, 2019), cross-sectional design (only Scheidnes & Tuller, 2019 is longitudinal) and lack of focus on first-generation bilinguals.

Studies of child L2 syntax in general, comprehension or production, have found that older age of L2 acquisition and stronger cognitive abilities are associated with stronger L2 syntactic abilities, when amount of L2 exposure is controlled (Paradis et al., 2017; Paradis et al., 2021; Rothman et al., 2016; Soto-Corominas et al., 2022; Unsworth et al., 2019). Cognitive abilities that have been associated with stronger L2 syntax in production are short-term verbal memory, analytic reasoning and working memory (Hamann et al., 2020; Paradis et al., 2017, 2021; Soto-Corominas et al., 2022). In previous studies with the same participants in the present study, analytic reasoning was associated with morphosyntax (Paradis et al., 2021; Soto-Corominas et al., 2022). Analytic reasoning is a component of language aptitude associated with pattern detection, and thus particularly relevant for learning morphosyntax (Paradis, 2011). Regarding input factors, more cumulative and current exposure to the L2 predict stronger L2 syntactic abilities (Hamann et al., 2020; Paradis et al., 2017, 2021; Rojas et al., 2016; Sorenson Duncan & Paradis, 2020; Soto-Corominas et al., 2022; Unsworth et al., 2019). Finally, higher family socio-economic status (SES) is associated with stronger L2 syntactic abilities in bilingual children (De Cat, 2021; Hamann et al., 2020; Paradis et al., 2021; Rojas et al., 2016; Soto-Corominas et al., 2022). To date, the influence of factors more specific to the refugee experience on the development of children's L2 complex syntax has been understudied (cf. Paradis, Soto-Corominas, et al., 2022).

2.1. The present study

The primary objective of this study was to examine general (MLUw) and complex syntax (CD) development longitudinally in language samples from a cohort of first-generation Arabic-L1, English-L2 refugee children during their first 5 years of residency in Canada. Given the importance of L2 complex syntactic abilities to schooling in the host country, we sought to understand what individual difference factors modulated this development, especially factors more specific to refugee populations. The secondary, more methodological objective was to determine the overlap between SRT performance and naturalistic syntactic production in the same bilingual children. Our specific research questions were:

- (1) What are the distributional patterns of MLUw and CD at each time point? What kinds of complex sentence types are being used? Do MLUw and CD display similar or dissimilar developmental trajectories over time? Based on research with

school-age monolinguals and bilinguals, we expected growth in MLUw and CD show some dissimilarities in terms of growth. These English-L2 children were expected to show similar distribution of complex sentence types as other English-L2 children (Paradis et al., 2017; Paradis, Sorenson Duncan, et al., 2022).

- (2) Which factors (age, maternal education and cognitive abilities) modulate individual variation in MLUw and CD over time? Do time in refugee camps and wellbeing factors reflecting refugee experiences predict additional variance in the models? We chose time in refugee camps as a pre-migration adversity factor, as this could index experiences like deprivation and interrupted schooling (Sirin & Rogers-Sirin, 2015). We chose current wellbeing because this indexes possible socio-emotional difficulties in response to both pre- and post-migration experiences and is coincident with our language measures. It is important to separate well-being and mental health from adverse experience because not all children show a trauma response to adverse experience (Bartlett & Sacks, 2019).
- (3) Does performance on the LITMUS SRT correlate with naturalistic production measures such as MLUw and CD? What are the contributions of MLUw, CD and lexical diversity in naturalistic production to variation in SRT performance? Lexical diversity was included in this analysis because lexical abilities have been shown to contribute to SRT performance. As individual differences have been analyzed in previous studies at Time 1 and Time 2 with the same participants, this was not investigated in the present study for SRT (Paradis, Soto-Corominas, et al., 2022; Soto-Corominas et al., 2022).

3. Method

3.1. Participants

Three time points were examined in this longitudinal study within a larger project investigating the bilingual development of Syrian refugee children and youth who arrived in Canada in 2016–2017 as part of a targeted resettlement program of UNHCR refugees organized by the Canadian government in response to the civil war in Syria. Participants and their families resided in Edmonton, Waterloo, or Toronto and ranged from 6;0–13;6 at the first data collection interval. At this interval, participants had different lengths of exposure to English (Table 1) due to different waves of arrival in Canada but were considered a cohort for the first Time period for the study. Therefore, at each Time period, the cohort accumulated more exposure to their English-L2. All participants spoke Syrian Arabic and upon arrival in Canada were placed in mainstream classrooms with additional English second language instruction at English medium schools. None of the participants spoke any additional non-English language. Recruitment took place through school boards, welcome centers for newcomers, mosques and word of mouth.

Since this study was based on a narrative task, data that were included for analysis in this study were based on participants with sufficient English ability to produce narratives. As a result, the number of participants included for analysis at Time 1 (T1; December 2017 to September 2018) was 108 participants (56 females), at Time 2 (T2; February 2019 to August 2019) was 118 participants (61 females) and at Time 3 (T3; October 2020 to December 2021) was 110 participants (61 females). The smaller number of participants at T1 compared to T2 was due to

Table 1. Participant characteristics

	T1 (N = 108) ^a			T2 (N = 118)			T3 (N = 110)		
	M	SD	Range	M	SD	Range	M	SD	Range
Age at testing (years)	9.48	1.97	6–14	10.19	1.99	7–14	11.82	1.95	8–16
Age of arrival (months) ^b	89.38	25.08	45–140	88.92	24.61	45–140	88.39	23.98	45–140
Length of L2 exposure (months)	21.63	7.50	3–36	32.68	7.07	12–47	50.63	7.08	29–63
Time in refugee camp (months) ^b	8.30	15.33	0–48	7.64	15.20	0–48	7.85	15.14	0–48
Maternal education (years) ^b	9.78	3.81	0–17	9.50	3.89	0–17	9.68	4.04	0–17
KBIT (cognitive) Score ^b	96.15	19.88	50–136	97.16	18.94	50–136	97.14	19.83	50–136

Abbreviations: M = mean, SD = standard deviation.

^aThe lower number of participants at T1 is due to proficiency issues that barred children's participation in producing narrative data in addition to some technical issues.

^bAlthough these variables remain stable over the three time points, the means are specified for each due to small differences in participant sample size over the course of the study. KBIT standard scores from T2 are presented across all three time points to account for different participant data being entered into models.

excluding 16 participants (average length of English exposure of 18.6 months) who were unable to produce any utterances for the narrative task. Due to the onset of the COVID-19 pandemic where testing was shifted from in-person to remote, there was a slightly longer interval between T2 and T3 than between T1 and T2. However, none of our research questions assumed equal intervals between T1, T2 and T3. Although our sample includes both children and adolescents, for the sake of simplicity, we refer to them as “children.” Further details of participant characteristics are reported in Table 1 in Results.

3.2. Procedures

3.2.1. Data collection

Data collection took place in children's homes or at school at T1 and T2. For T3, data collection took place remotely for both parents and children.

3.2.2. Materials

Alberta language environment questionnaire-4 (ALEQ-4; Paradis, 2011). We administered the ALEQ-4 in Arabic to parents in an interview format with a native speaker of Arabic to gather demographic and language exposure information on the child and the parents: child age measured in years, length of exposure to English measured in months, time in refugee camps measured in months and maternal education measured in years.

Kaufman brief intelligence test 2 (KBIT; Kaufman & Kaufman, 2004). We used a subtest from the KBIT to measure participants' nonverbal analytic reasoning. Participants were presented with five to six images to form an association with another picture or matrix of pictures. Standardized scores were calculated against age-norms, with a mean of 97 (SD = 19) and Cronbach's alpha of .91. The KBIT was only administered once at T2, since it was expected to be a stable estimate of cognitive abilities that would not change with time as would language environment variables, and after 3 years of L2 exposure, participants were proficient enough in the L2 to perform on such a minimally verbal task without language being a barrier. Standard KBIT scores were entered into our models for analyses.

Sentence repetition task (SRT; Soto-Corominas et al., 2022). At each time point, we administered a version of the LITMUS SRT for English (Marinis & Armon-Lotem, 2015), adapted for Canadian English (Soto-Corominas et al., 2022). The 31 target stimuli of the SRT comprised 15 monoclausals ranging in complexity (declaratives, short and long passives and wh-object questions)

and 13 biclausal sentences (coordinated clauses, subordinate clauses, and relative clauses). The list of stimuli items is given in Supplementary Table S1.

Children were administered the task on a laptop with headphones, where they listened to prerecorded sentences. They were instructed to repeat the sentence they heard. SRT can be scored in different ways depending on the goals of the research (Marinis & Armon-Lotem, 2015). For this study, we chose to score items for structural accuracy because our interests concern syntactic development. For structural accuracy coding, participants are assigned a score of 1 or 0 depending on whether they had the same syntactic structure in the stimuli and their repetition. Other errors in production were ignored, such as lexical substitutions not affecting the realization of the targeted constructions. So, if the stimuli included an adjunct/subordinate clause, e.g., *if the weather is warm, we can go to the park*, children would be assigned 1 if they produced an exact repetition or a repetition like, *if the weather is warm, we will go to the park*, and 0 if they produced a sentence that omitted the conditional with other coordinators like, *Ø the weather is warm and we can go to the park*. For more details on this scoring technique, see Soto-Corominas et al. (2022).

Strengths and difficulties questionnaire (SDQ; Goodman, 1997). The SDQ provides an index of wellbeing for children aged 3–16 using 25 behavioral screening questions that assess four negative behaviors and one positive behavior. Each of the 25 items are categorized into 5 subscales: (1) emotional symptoms, (2) peer relationship problems, (3) hyperactivity/inattention, (4) conduct problems and (5) prosocial behaviors. For instance, the emotional symptoms subscale comprised items like “Often complains of headaches, stomach-aches or sickness” and “Nervous in new situations, easily loses confidence,” where parents or children were asked to respond with “Not True,” “Somewhat True,” and “Certainly True” to produce scores between 0 and 2, respectively. With five items per subscale, the total possible score was 10, with higher scores for the behaviors indicating a higher incidence of the behaviors.

At each of the three time points, we administered the parent version to all participants. For children aged 7–16, we additionally administered the youth self-report version of the questionnaire, which consists of the same questions as the parental version with slightly different wording to reflect age-appropriate literacy and understanding. Both questionnaires were included for analysis in this study as the youth self-reports have been found to have high internal consistency (Theunissen et al., 2019), but due to the wide age range of the participants, only a subset of our sample was able to

complete the self-report at T1 and T2. The Cronbach's alpha was .79 for the parental SDQ and .80 for the youth self-reported SDQ.

Test of narrative language (TNL; Gillam & Pearson, 2017). Narrative language samples were collected using a subtest of the TNL. Participants wore noise-cancelling headphones and the research assistant in charge of the testing session shared their computer screen to present the visual stimuli for narrative elicitation. Participants were first asked to listen to a recording of a model story about a single picture ("Treasure") and then generate a narrative of their own based on a different picture ("Aliens"). The recorded production of the child's own narrative comprised the basis of the language samples used in the present study.

Transcription of narratives. The children's story generation was recorded, transcribed, and checked using formatting from Systemic Analysis of Language Transcripts (SALT; Miller & Iglesias, 2012), before being reformatted according to CHAT conventions (MacWhinney, 2000).

Given the age of the participants, utterance boundaries were based solely on grammatical structure to ensure systematic segmentation, and prosodic information was not used (Scott & Stokes, 1995; Wong et al., 2004); each utterance that was a sentence consisted of a main clause and its dependent clauses, if present. Ten percent of the recordings ($n = 30$) were randomly selected and transcribed independently by a different research assistant to calculate inter-rater reliability. Inter-rater reliability between original and new transcriptions was calculated as the proportion of agreed words over the total number of agreements and disagreements (Sackett, 1978), yielding a mean reliability score of 95% (82%–100%; $SD = 4\%$).

Coding of narratives. Once participants' transcriptions were reformatted into CHAT files, MLU in words for each participant was calculated at each time point by running the command "MLUw +*CHI -t%mor *.cha" in the CLAN program. Utterances from the participants' transcribed narratives were extracted from the narratives and entered into an Excel sheet for syntactic analysis. Since this study was concerned with sentences, only utterances with at least one finite verb were included. Fragments, false starts and self-corrections, common phrases (e.g., "thank you"), and comments unrelated to the narrative (e.g., "I don't know what this is") were excluded. Therefore, only complete sentences were included for coding and calculations of clausal density.

Sentences were first coded as either simple or complex. Complex sentences were sentences containing two or more clauses as determined by the number of main verbs or copulas. Complex sentences were coded by type: coordinated clause, sentential complement clause, adjunct clause, and relative clause (see [Supplementary Table S2](#), for examples of each structure from participants). Following Frizelle et al. (2018), clauses linked by coordinating conjunctions were only coded as a single utterance when the subject of the second clause was omitted. For instance, *Mack got scared and told Sarah* would be considered a coordinated complex sentence, whereas *Mack got scared/and he told Sarah* would be considered two monoclausals. The exclusion of coordinated complex sentences with overt subjects in both clauses was because of the difficulty in determining true coordinated sentences of this kind from sequences of independent sentences joined with "and" where children describe a sequence of events. Sentential complements included embedded sentences that were the complement of the verb in the matrix clause. Reported speech, like *She said, "I'm scared"* was entered as two separate utterances and only considered a sentential complement within the same utterance if the complementizer "that" could be inserted, like *She said (that)*

her brother was scared. Adjunct clauses included subordinating conjunctions like *because*, *when*, *if*, *before*, or *after*, to specify modifying information relating to the main clause. Relative clauses modified noun phrases to provide additional information about the relativized noun. A small percentage of complex sentences were coded as Ambiguous (1.02%) and excluded when they included at least two verbs but were difficult to classify according to type. For instance, in a sentence like *Something coming landing on the floor*, the lack of any overt coordinator or relativizer between the two verbs requires assumptions about what complex clausal structure the child was intending to use. For coding reliability, the first author and research assistants separately coded 25% (86 files) of the transcripts for complex syntax, after which the first author checked for agreement between both sets of coding. The coding reliability was 96% (92%–100%, $SD = 2\%$). Instances of disagreement were resolved by the first author revisiting the original transcript and making her most informed decision.

Calculation of mean language of utterance in words (MLUw), clausal density (CD), and lexical diversity (NDW). MLUw in words was calculated via the CHAT program, by dividing the total number of words in each transcript by the total number of utterances. Utterances included set phrases or clauses without verbs, such as "thank you" or "the ship also." Only non-repeated, intelligible words were included in the total count of words. CD was calculated by dividing the total number of clauses in each transcript by the total number of sentences. For instance, in the single sentence *The girl saw the ship and ran home*, there are two clauses – (1) *The girl saw the ship* (2) *and ran home* – yielding a clausal density of 2. Lexical diversity was calculated using the "freq" command to determine the number of different words (NDW) used. NDW was included in the analyses addressing research question 3.

4. Results

4.1. Participant characteristics

Participant characteristics at each time point, such as age at testing, length of exposure to English, maternal education in years, and standardized scores on our test of cognitive abilities and nonverbal intelligence (KBIT) from T2, are summarized in [Table 1](#).

At the outset of the study, participants were 9 years old on average and had been in Canada just under 2 years. At T2, participants were exposed to English for little less than 3 years and by T3, participants had been exposed to English for an average of over 4 years. Participant age was weakly correlated to length of English exposure ($r(334) = .35, p < .001$). There was a large variation in time spent in refugee camps. This is because some children in our sample did not spend time in refugee camps prior to resettlement in Canada ($N = 76$), thus resulting in 0 for this variable. Among the children who spent time in a refugee camp ($N = 42$), the average stay was 20 months and ranged from one day to 48 months. The variable for time in camp was scaled prior to modeling as well all other variables. The levels of maternal education, and thus SES, were low on average. Most mothers had only primary school education levels, averaging less than 10 years; however, the sample included some mothers with post-secondary education.

4.1.1. SDQ parental report-on-youth and youth self-report

Due to the complexity of the SDQ scoring and interpretation, results of this measure are presented here, rather than in [Table 1](#). Wellbeing subscale scores from SDQ parental report and youth

self-report at each time point are visualized in Figure 1, in Panel A and Panel B, respectively. For negative behavior subscales (emotional, peer problems, hyperactivity, conduct problems), higher scores denote more difficulties with wellbeing, whereas for the positive behavior subscale (prosocial), higher scores indicate better wellbeing. Following Paradis, Sorenson Duncan, et al. (2022), the interpretation of the score categories was determined by using both US-population parental report (PRY) norms (Bourdon et al., 2005) and UK-population youth self-report (YSR) norms (Meltzer et al., 2000), where 80% of the population are classified as normal and the remaining 20% were classified as borderline-to-abnormal. These two separate norms were used because our participants are being studied in Canada, and the US-norms only included PRY while the UK-norms included YSR. The dashed lines in Figure 1 indicate the cut off scores for borderline-to-abnormal frequency of behaviors, above the line for negative subscales and below the line for the positive subscale. Details on norm-referenced subscale scores are provided in Supplementary Table S3. For PRY (Panel A), more than 20% of this study's participants scored in the borderline-to-abnormal range of the following subscales, averaged from T1 to T3: emotional = 20.92%, hyperactivity = 25.35% and prosocial = 24.72%;

this indicates that, among these refugee participants, difficulties with wellbeing were more common than the population average. A chi-squared test revealed a significant difference between borderline-to-abnormal scores for each subscale on the PRY for each time point compared to the expected 20% of the population ($\chi^2 = 46.86, p < .001$). Turning to the YSR (Panel B), the relatively lower number of participant scores at Time 1 ($N = 56, M_{\text{age}} = 10.84, SD_{\text{age}} = 1.30$) was due to age, since we only administered the YSR to children 9 years and older. For YSR, even larger percentages of children scored in borderline-to-abnormal ranges in all five of the subscales on average from T1 to T3: emotional = 49.37%, peer problems = 72.64%, hyperactivity = 39.72%, conduct problems = 27.78%, prosocial = 45.63%. A chi-squared test revealed a significant difference in the number of borderline-to-abnormal scores on the YSR for each subscale at each time point compared to the expected 20% of the population ($\chi^2 = 667.49, p < .001$). Finally, we ran Pearson correlations to understand the relationship between PRY and YSR for the individual participants (Supplementary Figure S1). Because the only significant correlations between corresponding subscales were weak ($r > .3$), YSR and PRY scores for each SDQ subscale were regarded as different

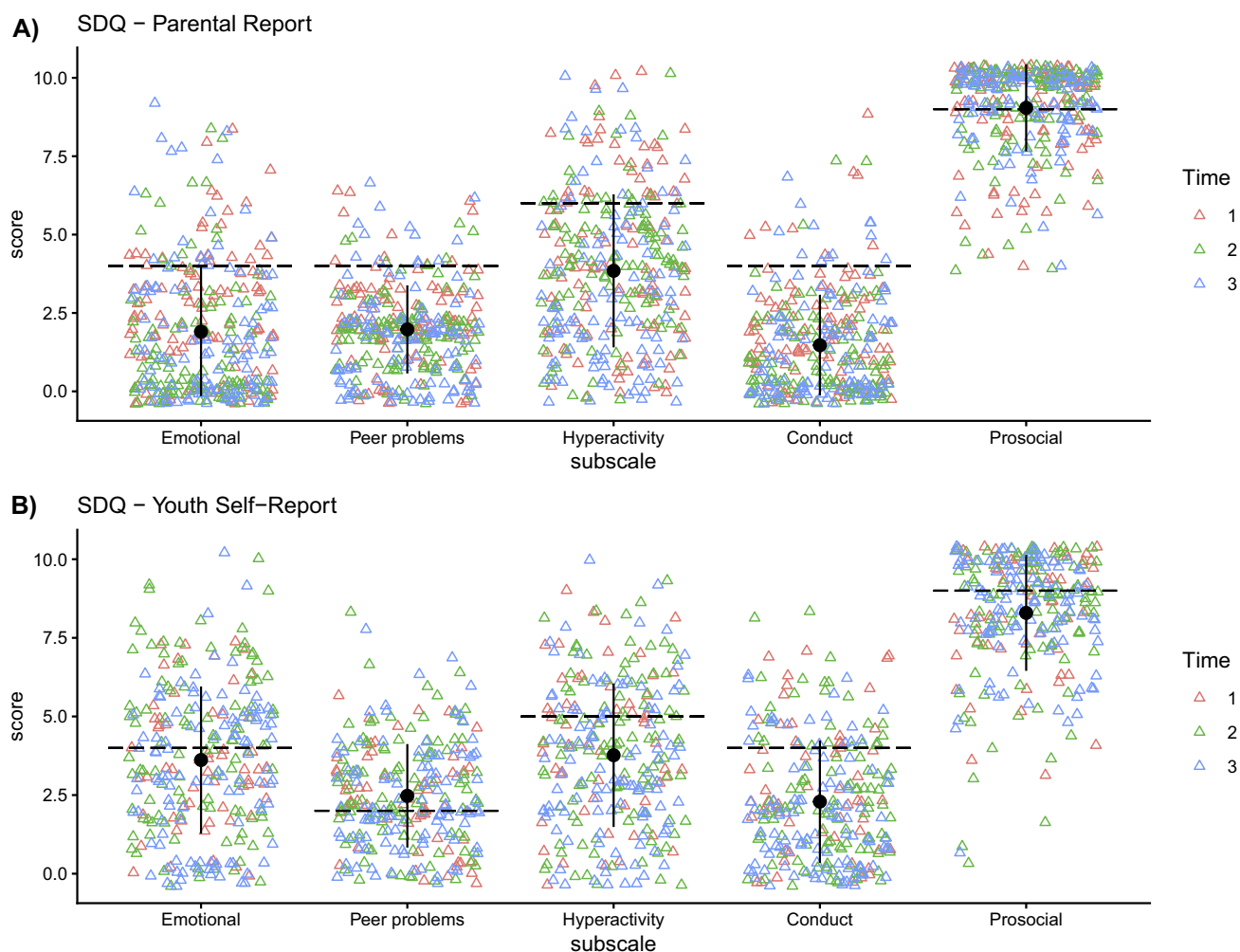


Figure 1. Participant scores by SDQ subscale for (A) parental report and (B) youth self-report.

Note. Each point denotes one participant and the different colors represent each of the three time points. The black dots in the middle of each column indicate the mean score for the subscale, and the error bar indicates one SD below and above the mean. The horizontal, dashed lines indicate borderline scores, which separate normal and abnormal scores: for the four negative SDQ subscales (emotional, peer problems, hyperactivity and conduct), borderline and abnormal scores are above the dashed lines, and for the one positive SDQ subscale (prosocial), borderline and abnormal scores are below the dashed line.

indicators of each child's wellbeing that do not covary. This result influenced how PSR and YSR subscales scores were entered in the regression models in *Sources of Individual Differences across Time*.

4.2. MLUw and CD over time

For our initial exploration of MLUw and CD, we examined their distributions among our participants at each time point, visualized in Figure 2. We grouped the participants into cohorts based on time points, in alignment with conventional treatment of longitudinal data (Collins, 2006). As a reminder, the participant groups at each time point had various lengths of exposure to the L2; nevertheless, they all had longer exposure with each time point, so time is a proxy for length of L2 exposure. We conducted Shapiro–Wilk tests at each time point using base R to investigate normality distribution. Only MLUw at T1 ($M = 6.56$, $SD = 1.96$, range = 1.5–14) was normally distributed ($W = .98$, $p = .141$), while it was not normally distributed at T2 ($M = 6.84$, $SD = 1.70$, range = 2.5–13.3, $W = .96$, $p = .002$) and at T3 ($M = 8.35$, $SD = 1.88$, range = 4.2–14.2, $W = .98$, $p = .040$). At no time point was CD normally distributed: T1 ($M = 1.34$, $SD = .25$, range = 1–2.11, $W = .95$, $p < .001$), T2 ($M = 1.46$, $SD = .28$, range = 1–2.4, $W = .93$, $p < .001$) and T3 ($M = 1.60$, $SD = .31$, range = 1–2.55, $W = .96$, $p = .002$). The plots for CD showed a slight bimodal distribution, with a secondary peak at around CD = 2. After applying various transformations, the T2 and T3 MLUw data and the CD data remained non-normally distributed; therefore, we opted to use non-transformed data in our regression models.

To understand the range of structures CD was accounting for, we examined the proportions of different types of sentences used by participants across time (Figure 3). Proportions were calculated by dividing the total tokens of each sentence type by the total number of utterances. At all three time points, over half of the sentences produced were simple monoclausal sentences. There was an increasing proportion of complex sentence types (sentential complement, adjunct/subordinate, relative and coordinate) between T1 through T3, with most complex sentences being sentential complement clauses at each time point.

Our first research question asked how MLUw and CD scores change along with greater exposure to English-L2 over time. To answer our first research question, we fit linear mixed-effects regression models using the lme4 package (Bates et al., 2014; R Core Team, 2024). MLUw and CD were included in their respective models as the dependent variable, with Time as the fixed effect and Participants as a random intercept. We nested Participant within Family for MLUw to let the model account for any variability that may arise due to 73% of participants being siblings, while nesting Participant within Family for CD resulted in a singular model. These models served as base models without other participant-level factors as predictors to observe the overall growth trends for MLUw and CD over time. Model outputs for the fixed effects are given in Table 2; additional details are provided in Appendix D, Model Output S1, SM.

In our model for MLUw, the fixed effect of Time had 3 levels with a reference level of T1. MLUw between T1 and T2 showed no significant changes. Between T1 and T3, MLUw increased significantly.

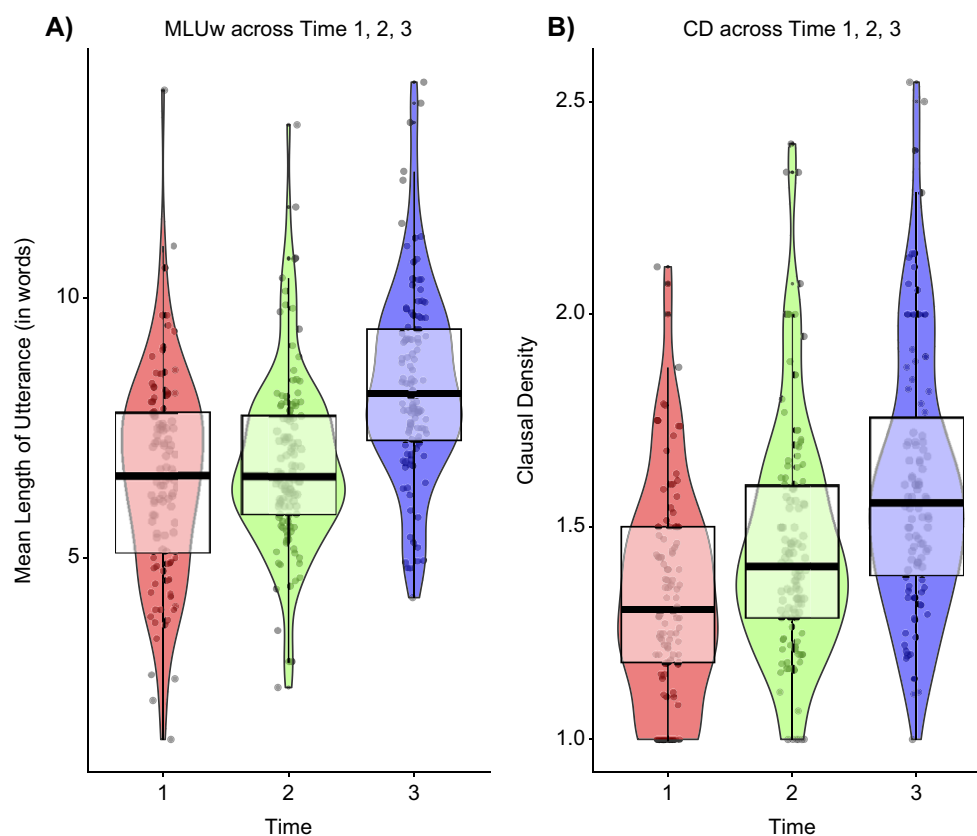


Figure 2. Plots of (A) MLUw and (B) CD across Time 1, Time 2 and Time 3.

Note. Each dot represents one participant. Medians are indicated by the dark black line, and interquartile ranges are indicated with the opaque box.

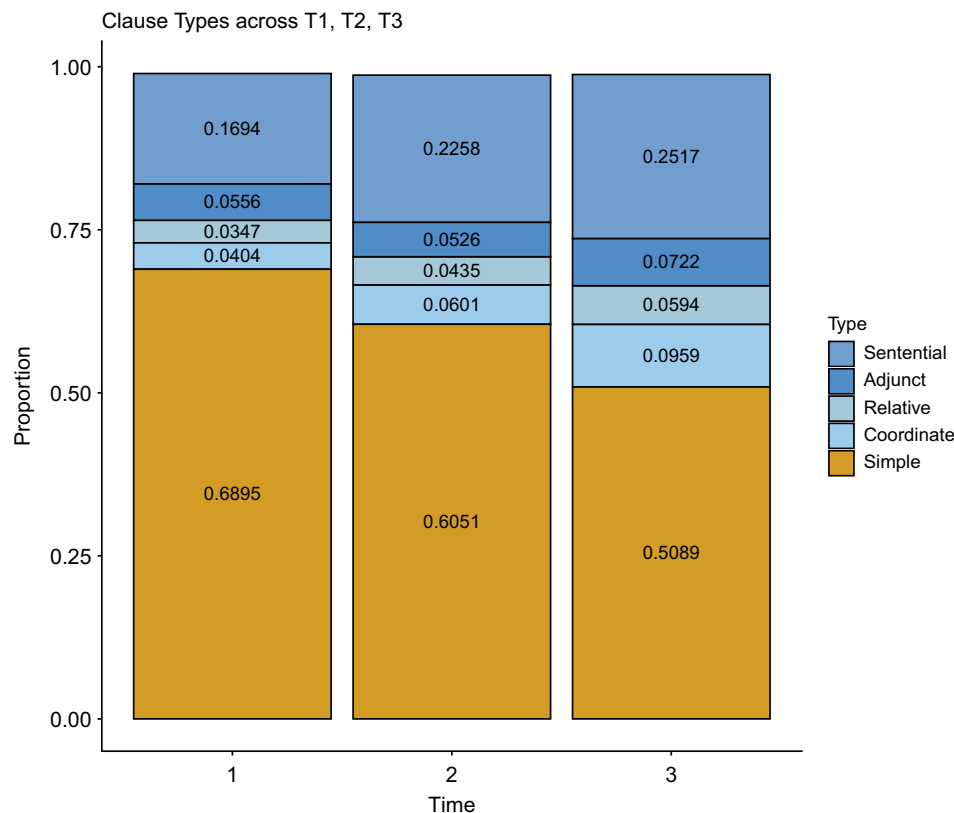


Figure 3. Proportions of clause types across Time 1 (T1), Time 2 (T2) and Time 3 (T3).

Note. Complex sentences comprise sentential complement, adjunct, relative and coordinate clauses, which are indicated in shades of blue. Simple sentences are monoclausal, indicated in orange. For examples of each sentence type, see [Supplementary Table S2](#).

Table 2. Linear mixed effects models pairwise coefficient table for MLUw and CD with Time as a fixed effect and Participants as a random intercept.

	Estimate	Std. Error	Pr(> z)
MLU			
T2–T1	.29	.21	.171
T3–T2	1.44	.21	< .001 ***
T3–T1	1.73	.21	< .001 ***
CD			
T2–T1	.11	.03	.004 **
T3–T2	.14	.03	< .001 ***
T3–T1	.26	.04	< .001 ***

Note: * $p < .05$; ** $p < .01$; *** $p < .001$.

In our model for CD, a fixed effect of Time and reference level of T1 showed that CD increased significantly between T1 and T2, T2 and T3, and T1 and T3. Using the R package emmeans (Lenth, 2023), our pairwise comparisons confirmed the significant increases. Finally, we ran Pearson correlations between MLUw and CD at each time point. MLUw and CD shared a moderate, positive correlation at T1 ($r(106) = .60, p < .001$) and T2 ($r(116) = .62, p < .001$), and a strong, positive correlation at T3 ($r(108) = .77, p < .001$).

In sum, these analyses revealed that both MLUw and CD increased with greater exposure to English-L2 from T1 to T3, although CD showed a steadier pattern of growth. Furthermore,

MLUw and CD were related, but not identical, measures of syntax at each time point.

4.3. Sources of individual differences across time

Our second question asked what individual difference factors modulated MLUw and CD scores over time. To address this question, we analyzed the additional influence of individual difference factors on MLUw and CD over time by fitting linear mixed-effects regression models using the lme4 package (Bates et al., 2014). Like the models we built to answer our first research question, MLUw and CD were dependent variables for their respective models, Time was a fixed effect, and random effects included Participants. Fixed effects included the following individual difference factors that have been found to be associated with syntactic development in previous research (see *Introduction*); these factors were scaled and centered for entry into each model: participant age, maternal education, KBIT standardized scores (cognitive ability = analytic reasoning) and time in refugee camp (adversity). For our sample, the SDQ fixed effects and time in camp are not strongly correlated ($r < .19$), so SDQ subscales (wellbeing) were added one at a time to the regression models with the other predictors, generating a large set of parallel rather than nested models. We chose this procedure for two reasons: (1) statistical power of the model given the number of participants in our study precluded adding all of them in the same model and (2) the low correlation between PRY and YSR warranted investigation of which version, if any, predicted MLUw and CD growth over time. Optimal models from this set were determined using a likelihood ratio test

(anova; Chambers & Hastie, 1992) to compare deviance between models. Dominance analyses were conducted on final models with the *domir* package (Luchman, 2024), using variance explained (R^2) comparisons between all possible combinations of predictors to determine the relative importance of fixed effects in the models. Dominance analyses permit model comparisons with a high number of fixed effects while minimizing incorrect interpretation of standardized beta coefficients (Mizumoto, 2023). Results are given in Table 3, and the interaction for our MLUw model is visualized in Figure 4. More details are given in Supplementary Appendix A.

The optimal model for MLUw revealed an interaction effect between Time and Participant age, with younger participants experiencing larger gains in MLUw between T1 to T2 and between T1 to T3 compared to older participants. Higher maternal education and KBIT scores were significantly associated with longer MLUw. Higher scores on the YSR prosocial were marginally predictive of longer MLUw. Dominance analysis of the model for MLUw showed that Time was the most important predictor in the model, followed by the interaction between Time and Participant age. The third-ranked predictor was KBIT scores, followed by Participant age, and maternal education. Time in refugee camp and YSR SDQ prosocial subscale scores were the two least dominant predictors.

The overall regression model for CD reveals no interaction effects between Time and the other factors. Older participant age, increased maternal education and higher KBIT scores significantly predicted higher CD scores over time. Increased time in refugee camps resulted in lower CD scores. Higher Hyperactivity subscale

scores from the PRY SDQ marginally predicted lower CD scores. Dominance analysis showed Time to be the most important fixed effect and Participant age to be the second. The third-ranked predictor was KBIT, followed by time in refugee camp and maternal education. The least dominant predictor was the Hyperactivity subscale score from the PRY SDQ.

In summary, individual difference factors emerged as predictors for MLUw and CD to different extents. Older participants had higher CD scores regardless of time, but for MLUw, younger participants showed steeper growth than older ones. Higher KBIT scores and increased maternal education were both significant predictors of longer MLUw and higher CD scores. Increased time in camp was a significant predictor of lower CD scores but not for MLUw. A higher prosocial subscale score from YSR version of the SDQ trended toward significance in predicting higher MLUw, while a higher hyperactivity score from the PRY version marginally predicted lower CD scores. Dominance analysis rankings showed that Time, Participant age, and KBIT scores were among the most dominant effects for both models, while wellbeing scores were among the least dominant predictors for both models. Time in camp was a less dominant predictor for MLUw than for CD.

4.4. Relationship between SRT, MLUw and CD

Our third research question asked about the relationship between measures of syntax from language samples and performance on the SRT. First, we first examined correlations between MLUw and CD and total SRT scores, irrespective of Time (e.g., each

Table 3. Linear mixed effects models coefficient table for MLUw and CD with as Time and individual difference factors as fixed effects and Participants as a random intercept.

	Estimate	Std. Error	t value	Pr(> z)	Dominance Ranking
MLU					
Intercept	6.95	.24	28.92	< .001 ***	
T2	.01	.26	.04	.968	1
T3	1.44	.28	5.18	< .001 ***	
Age	.78	.33	2.37	.019 *	4
Maternal education	.34	.13	2.56	.013 *	5
KBIT score	.30	.13	2.34	.021 *	3
Time in camp	-.20	.14	-1.47	.147	6
YSR Prosocial	.21	.11	2.00	.046 *	7
T2: Age	-.64	.35	-1.81	.072 .	2
T3: Age	-.88	.35	-2.48	.014 *	
CD					
Intercept	1.37	.03	49.18	< .001 ***	
T2	.09	.04	2.52	.012 *	1
T3	.19	.04	4.63	< .001 ***	
Age	.05	.02	2.94	.004 **	2
Maternal education	.05	.02	2.97	.004 **	5
KBIT score	.05	.02	3.13	.002 **	3
Time in camp	-.04	.02	-2.66	.009 **	4
PRY hyperactivity	-.03	.02	-1.82	.070 .	6

Note: * $p < .05$; ** $p < .01$; *** $p < .001$.

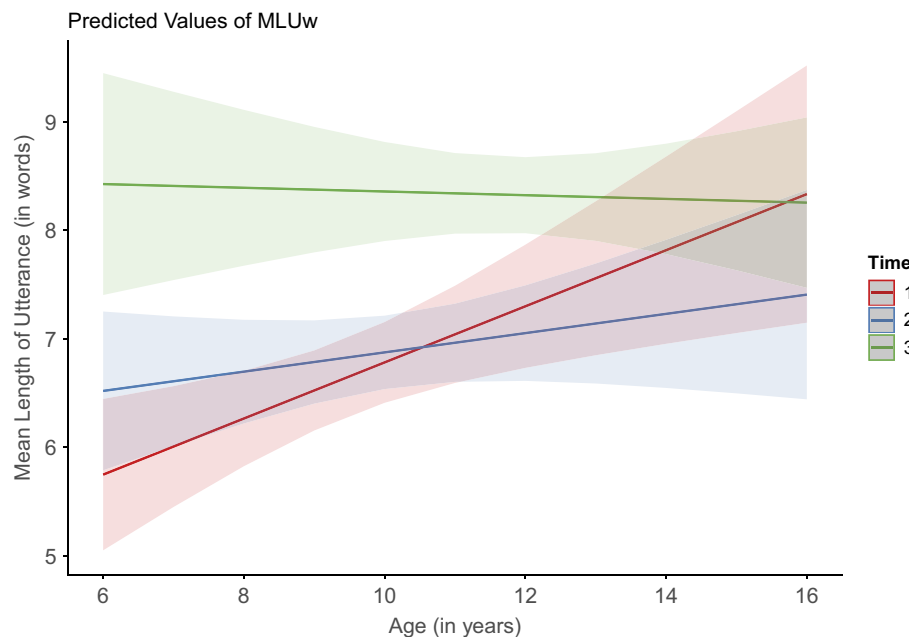


Figure 4. Interaction plot between time and participant age for MLUw.

Note. The interaction plot provides a visualization of the predicted linear trends for participants' MLU values based on their age in years. While older participants are not attested at T1 and younger participants are not attested at T3, the lines extrapolate from data of participants ages 8–14 who are present across all time points.

participant contributed three scores to the correlations). SRT scores were moderately correlated with MLUw ($r(324) = .51$, $p < .001$) and small-to-moderately correlated with CD ($r(324) = .42$, $p < .001$). Therefore, higher structural accuracy scores on SRT are associated with longer MLUw and more clausal density in naturalistic production.

To probe the relationship further, following Hamann *et al.* (2020) and Zebib *et al.* (2020), we next modeled SRT scores for each item as outcomes with MLUw, CD and NDW—as fixed effects. This strategy permits us to understand the *relative contribution* of each skill in naturalistic production to variance in SRT scores. In so doing, it reveals what SRT is indexing in children's language abilities. Therefore, we fitted binomial generalized linear mixed effects models to investigate how NDW (lexical abilities), MLUw (general syntax abilities) and CD (complex syntax abilities) predicted performance on (1) all SRT items and (2) biclausal items on the SRT (Supplementary Table S1). Biclausal SRT items included subordinate, coordinated and relative clauses. To best account for our measures of syntactic complexity from the narrative data, we also created models where coordinate clauses ($k = 3$) from the SRT were excluded because they involved two explicit subjects, one for each clause, that would not be considered complex clauses according to our language sample coding protocol. Exclusion or inclusion of coordinate clauses produced the same overall pattern of results. The outcome variable of SRT structural accuracy was an item-level either 0 or 1, with 1 being given to sentences that were repeated using the same syntactic structure as the stimulus item. We also added Time as an additional fixed effect since the data are longitudinal. SRT Item and Participant nested under Family were random intercepts. Model results are given in Table 4; details are given in Supplementary Appendix A.

For the model of all SRT items, participants increased their likelihood of a target structural repetition with more exposure to their L2 (Time). Greater NDW and longer MLUws were associated with more accurate repetitions on the SRT, but CD scores were not

associated with SRT performance. The pattern of results was the same in the second model, even though this model only included items closely related to CD. In sum, MLUw and NDW contributed to SRT performance but CD did not.

5. Discussion

The present study examined the longitudinal development of English-L2 productive syntax in bilingual Syrian children who

Table 4. Binomial generalized linear mixed effects models coefficient table for SRT structural accuracy scores with Time, NDW, MLUw, and CD as fixed effects and Participants and SRT Item as random intercepts.

	Estimate	Std. Error	z value	Pr(> z)
All SRT				
Intercept	.80	.26	3.04	.002 **
T2	.99	.08	12.53	< .001 ***
T3	1.53	.10	15.32	< .001 ***
NDW	.30	.05	5.44	< .001 ***
MLUw	.16	.06	2.69	.007 **
CD	.02	.05	.31	.761
Biclausal SRT				
Intercept	.70	.34	2.07	.038 *
T2	1.13	.12	9.50	< .001 ***
T3	1.44	.14	9.99	< .001 ***
NDW	.31	.08	4.05	< .001 ***
MLUw	.28	.09	3.14	.002 **
CD	.08	.08	1.06	.291

Note: * $p < .05$; ** $p < .01$; *** $p < .001$.

arrived in Canada as refugees. Using participants' narrative language samples, we measured general syntax through MLUw and complex syntax through CD. The objectives of this study were three-fold: (1) to investigate how general and complex syntax in the English-L2 developed longitudinally, (2) to investigate whether individual differences, including wellbeing and adversity factors, modulated this trajectory for refugee children from Syria and (3) to examine whether these naturalistic measures of syntactic development corresponded with participants' performance on an SRT.

5.1. Developmental trajectories of MLUw and CD over time

Our first research question asked how the longitudinal development of general syntax and complex syntax in children's English-L2 compared to each other. To address this question, we observed distributional patterns, modelled the development of bilinguals' general syntax (MLUw) and complex syntax (CD) over time, and conducted correlations between them at each time point. Results showed that MLUw and CD had different distributions in the sample – MLUw had a relatively more normal distribution than CD, meaning there was a small group of participants who had consistently higher CD than others. Regarding developmental trajectories, MLUw plateaued between T1 and T2, and only showed growth between the T2 and T3. Meanwhile, CD showed consistent linear growth from T1 to T2, and from T2 to T3. Finally, MLUw and CD were moderately correlated at T1 and T2, indicating that they measured related but not identical components of children's syntactic abilities. By contrast, at T3, MLUw and CD were highly correlated. This could be attributed to the increased production of multiclausal sentences by T3, while monoclausal sentences that were more common at T1 and T2 would vary more in length. Further analyses of the proportions of simple and complex sentences that contributed to calculating CD revealed that, even at T1, over 30% of children's sentences were complex. This proportion is slightly higher than what Paradis et al. (2017) found for younger child L2 speakers with less English exposure.

These results have implications for how interchangeable MLUw and CD are as measures of syntactic development, and whether complex syntax develops relatively early in child L2 acquisition. First, in line with studies on both monolingual and bilingual children, our findings indicate that MLUw and CD measure overlapping but not identical components of children's syntactic abilities (Castilla-Earls et al., 2021; Frizelle et al., 2018; Paradis & Kirova, 2014; Paradis, Sorenson Duncan, et al., 2022). Even though CD is not an automatically calculable measure, if multiclausal sentences and embedding are the structures of interest for researchers, MLUw would only provide a partial approximation of complex syntax. Thus, our results suggest that complex syntax as measured by CD could be distinct from a broader measure of general syntax like MLUw and point to the relatively earlier and steadier growth of complex syntax in L2 development. Second, comparisons with monolingual English age peers from Frizelle et al. (2018) show similarities in CD and proportions of complex structures produced at T1 by participants at all ages in our sample, except for the oldest group (age 11+). This bodes well for L2 syntactic skills required for schooling, in contrast with research on the protracted development of L2 lexical skills which has been a source of concern for reading development in school-age bilinguals (Lesaux et al., 2010; Soto-Corominas et al., 2020). It is relevant to consider explanations for why bilingual children might acquire complex syntax relatively rapidly in their L2. The greater linguistic and cognitive maturity of child L2 learners compared to

their younger, monolingual counterparts have been put forward as explanations (Paradis et al., 2017; Scheidnes & Tuller, 2019). First, young bilinguals have already acquired some aspects of clausal embedding in their L1, and such computational features of language are considered part of the shared representation between the two languages of bilinguals (Scheidnes & Tuller, 2019; Soto-Corominas et al., 2022). Second, children aged 4 and older have more developed cognitive skills that are used in language learning than younger children and are in a complex communicative context (school), both of which could prompt the desire and need for expressing dual propositions within one sentence more so than for younger children (Paradis et al., 2017). Furthermore, bilingual children learning a target language at school are likely exposed to more complex syntax in their input, especially through literary sources, than younger monolinguals. All these explanations could work together to account for rapid acquisition of L2 complex syntax.

5.2. Contributions of individual difference factors to MLUw and CD development

Our second research question asked which individual difference factors modulate general and complex syntax development. To address this question, we added additional predictors, e.g., sources of individual differences, to the longitudinal models generated for our first research question that examined productive syntax across time points. Overall, refugee children's outcomes were heterogeneous and variation in both MLUw and CD was predicted by many individual difference factors, suggesting that development of expressive syntax is sensitive to many extra-linguistic factors, which will be discussed in descending order of dominance ranking, followed by refugee-specific factors.

The addition of participant age in our models continued to highlight differences between what MLUw and CD measure: although both MLUw and CD increased with older age, an interaction effect between participant age and Time only emerged for MLUw. Specifically, at T1, MLUw was longer for the oldest participants compared to the youngest participants. At T2, this effect diminished and by T3, the youngest children had longer MLUw than the oldest ones. Hence, there was an older age advantage in MLUw only at the earlier stages of English-L2 acquisition, aligning with what Paradis, Sorenson Duncan, et al. (2022) found for performance on SRT.

For other sources of individual differences across time points, we found that superior non-verbal cognitive abilities and higher levels of maternal education were associated with longer MLUw and higher CD scores, consistent with other studies with bilinguals for various L2 proficiency measures (De Cat, 2021; Paradis et al., 2017; Paradis, Soto-Corominas, et al., 2022; Rojas et al., 2016; Soto-Corominas et al., 2022; Unsworth et al., 2019). Paradoxically, mothers of all education levels in our sample overwhelmingly reported using only Arabic in their interactions with their children. In fact, maternal education may measure broad socio-cultural capital in the family, which promotes advanced language development in both languages due to increased resources and overall language exposure (Bohman et al., 2010; Collins & Toppelberg, 2021; De Cat, 2021; Paradis, Soto-Corominas, et al., 2022). This finding also suggests that the influence of maternal education on English-L2 complex syntax could be mediated through interdependence in development of complex syntax across languages (Blom et al., 2021; De Cat, 2021; Paradis, Sorenson Duncan, et al., 2022), through more general

language and literacy richness at home, or both. Because all families in this study were low income and living in precarious housing situations (Paradis, Soto-Corominas, et al., 2022), which are risk factors for lowered language outcomes (De Cat, 2021; Rowe, 2018), higher maternal education seems to emerge as a protective factor in the L2 acquisition of refugee children.

Refugee-specific factors considered for this study's participants were time spent in refugee camps pre-migration and current socio-emotional wellbeing. As mentioned for Research Question #2, these variables were separated because they do not measure the same qualities in participants – time in camp and SDQ scores were weakly correlated ($r < .2$) at each time period. We found that longer time spent in camps was associated with depressed CD and MLUw. Time in camp emerged as a highly dominant predictor in our model for CD but a lower-ranked one for MLUw. The association of time in refugee camp and L2 abilities suggests that trauma and deprivation that occurred years before can still have relevance for children's language development (cf. Paradis, Soto-Corominas, et al., 2022). Wellbeing factors had small effects on MLUw and CD and were among the least dominant predictors in our models. Increased prosocial behaviors were associated with higher MLUw, while increased hyperactivity was associated with lower CD. Hyperactivity was found to be associated with performance on other English-L2 tasks for this sample of participants at T2 (Paradis, Soto-Corominas, et al., 2022). Hyperactivity or externalizing behaviors in general, interfere with executive functions necessary for learning (Graham et al., 2016). The bottom line is wellbeing factors influenced L2 expressive syntax, although their effects were relatively small. Overall, the refugee-specific factors examined in this study accounted for less variance in children's general and complex syntax acquisition than age, cognitive abilities and maternal education. A possible interpretation of this difference is that expressive L2 syntax is more resilient to the effects of refugee experiences than other L2 abilities.

5.3. Association between SRT performance and syntax in naturalistic production

The final research question of this study focused on the association between performance on SRT and measures of general (MLUw) and complex (CD) syntax from naturalistic language samples to understand if they are measuring similar constructs. Lexical diversity (NDW) from the language samples was also included, as prior research has found that lexical abilities contribute to SRT performance. To address this question, we first examined correlations before fitting binomial generalized linear mixed effects models with itemized SRT structural accuracy scores as the output. For the first model, structural accuracy scores for the entire task were used, and the second model only focused on biclausal items. Correlational analyses showed MLUw was moderately and positively correlated with total SRT scores while CD showed a small-to-moderate positive correlation coefficient. This pointed to a general tendency for children with longer MLUws and, to a lesser extent, higher CD scores to perform better on the SRT. Our follow-up modeling analyses were used to answer the more specific question of how expressive syntax and lexical diversity derived from naturalistic production contributed to variance in structural accuracy on an item-by-item basis for all SRT items and for only biclausal SRT items. Our results showed MLUw and NDW contributed significantly to both the total number of items and biclausal item models, but CD was not a significant contributor to either. MLUw and NDW specified similar variance in biclausal items, but NDW specified more than MLUw for all items.

Therefore, both the correlational and regression analyses indicated that general syntax was more associated with SRT performance than complex syntax. Our results are consistent with other studies using similar analysis techniques that found SRT to be a measure of general linguistic abilities, as both morphosyntax and lexical abilities contributed to performance (Hamann et al., 2020; Zebib et al., 2020). However, the present study is the only one to date comparing measures of *naturalistic* production with SRT performance in L2 children and youth.

One possible explanation for why CD did not predict performance on the biclausal SRT lies in the types of items represented in the SRT. Biclausal items consisted of subordinate/adjunct clauses and relative clauses. By contrast, in our language sample data, the largest proportion of complex sentences produced were sentential complements, with adjunct and relative clauses comprising a much smaller proportion (Figure 3). This mismatch between the SRT items and naturalistic data may have hindered stronger associations between CD and SRT performance in the regression analysis. Despite these null effects for CD, we cannot conclude that expressive complex syntax does not matter for SRT performance because our correlational analyses showed a relationship because MLUw does include complex sentences, and because relative clause items on SRTs provoke more errors than other items (Frizelle et al., 2017; Soto-Corominas et al., 2022). Furthermore, Torregrossa et al. (2024) found that employing a SRT where sentences were embedded in a coherent story-discourse improved performance. Perhaps this kind of SRT task would show a closer relationship to naturalistic speech, but future research is needed to know for certain.

6. Conclusions and limitations

One key finding from this study is that expressive general syntax and complex syntax in child English-L2 acquisition develop differentially but are related. Another key finding concerns L2 syntactic acquisition in refugee children compared to non-refugee children. First, the relatively high production of multiclausal sentences compared to monolingual age-peers and bilinguals from more advantaged backgrounds suggests that acquisition of complex syntax occurs quickly and steadily improves over time. Second, while individual difference factors common to all bilinguals (i.e., cognitive factors, maternal education) predicted growth in general and complex syntax overtime, time in refugee camps and current wellbeing predicted growth to a much smaller extent. Notably, time in refugee camps and current wellbeing predicted more variance in other studies examining different measures of these children's L2 abilities at Time 1 and Time 2 (Blom et al., 2021; Paradis, Sorenson Duncan, et al., 2022) than in the present study. Put together, this pattern of results suggests that L2 syntax could be more resilient than other L2 abilities to the adversity experienced by many refugee children. Future research examining the impact of wellbeing on L2 development ought to include comparisons across different linguistic domains and techniques like structural equation modeling to better understand these relations. Finally, this study contributes preliminary evidence that general syntax and lexical diversity in spontaneous language samples are more closely associated with SRT performance than CD. However, our findings were limited by using the LITMUS SRT whose items were mismatched with the frequency of complex sentence types produced in the bilingual children's narratives. Future research on the question of overlap between naturalistic production and SRT should consider adding sentential complements to an SRT.

Supplementary material. The supplementary material for this article can be found at <http://doi.org/10.1017/S1366728925100862>.

Data availability statement. The data that support the findings of this study are openly available in OSF at <https://osf.io/qygsj/>.

Acknowledgements. We would like to first thank the families for taking the time and effort to participate in this research. We also acknowledge the important contribution of all the student assistants in three cities (Edmonton, Waterloo and Toronto) who collected the data, and Rachel Tu, Maissie Hillman and Chris Burgess in Edmonton for their help with transcription and coding.

Funding statement. This research was funded by the Social Sciences and Humanities Research Council of Canada (Partnership Grant and Insight Development Grant; Chen and Paradis), for which we are grateful.

References

- Agmon, G., Pradhan, S., Ash, S., Nevler, N., Liberman, M., Grossman, M., & Cho, S. (2024). Automated measures of syntactic complexity in natural speech production: Older and younger adults as a case study. *Journal of Speech, Language, and Hearing Research*, 67(2), 545–561.
- Armon-Lotem, S., & Meir, N. (2016). Diagnostic accuracy of repetition tasks for the identification of specific language impairment (SLI) in bilingual children: Evidence from Russian and Hebrew. *International Journal of Language Communication Disorders*, 51, 715–731.
- Bartlett, J. & Sacks, V. (2019). Adverse childhood experiences are different from childhood trauma and it is critical to understand why. *Child Trends: Child Welfare*. <https://www.childtrends.org/publications/adverse-childhood-experiences-different-than-child-trauma-critical-to-understand-why>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.
- Blom, E., Soto-Corominas, A., Attar, Z., Daskalaki, E., & Paradis, J. (2021). Interdependence between L1 and L2: The case of Syrian children with refugee backgrounds in Canada and the Netherlands. *Applied PsychoLinguistics*, 42(5), 1159–1194.
- Bohman, T. M., Bedore, L. M., Peña, E. D., Mendez-Perez, A., & Gillam, R. B. (2010). What you hear and what you say: Language performance in Spanish–English bilinguals. *International Journal of Bilingual Education and Bilingualism*, 13(3), 325–344.
- Bourdon, K. H., Goodman, R., Rae, D. S., Simpson, G., & Koretz, D. S. (2005). The strengths and difficulties questionnaire: US normative data and psychometric properties. *Journal of the American Academy of Child & Adolescent Psychiatry*, 44(6), 557–564.
- Browder, C. T. (2018). Recently resettled refugee students learning English in US high schools: The impact of students' educational backgrounds. In *Educating refugee-background students: Critical issues and dynamic contexts* (pp. 17–32). Multilingual Matters.
- Brown, R. (1973). Development of the first language in the human species. *American Psychologist*, 28(2), 97.
- Cahill, P., Cleave, P., Asp, E., Squires, B., & Kay-Raining Bird, E. (2020). Measuring the complex syntax of school-aged children in language sample analysis: A known-groups validation study. *International Journal of Language & Communication Disorders*, 55(5), 765–776.
- Castilla-Earls, A., Pérez-Leroux, A. T., & Auza, A. (2021). Elicited vs. spontaneous language as methods for the assessment of grammatical development: The DEME assessment tool. *Revista de Logopedia, Foniatria y Audiología*, 41(4), 164–171.
- Chambers, J. M., & Hastie, T. J. (1992). *Statistical models in S*. Wadsworth & Brooks/Cole.
- Cleave, P. L., Girolametto, L. E., Chen, X., & Johnson, C. J. (2010). Narrative abilities in monolingual and dual language learning children with specific language impairment. *Journal of Communication Disorders*, 43(6), 511–522.
- Collins, B. A., & Toppelberg, C. O. (2021). The role of socioeconomic and sociocultural predictors of Spanish and English proficiencies of young Latino children of immigrants. *Journal of Child Language*, 48(1), 129–156.
- Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology*, 57(1), 505–528.
- De Cat, C. (2021). Socioeconomic status as a proxy for input quality in bilingual children? *Applied PsychoLinguistics*, 42(2), 301–324.
- Fazel, M., & Stein, A. (2002). The mental health of refugee children. *Archives of Disease in Childhood*, 87(5), 366–370.
- Frizelle, P., O'Neill, C., & Bishop, D. V. (2017). Assessing understanding of relative clauses: A comparison of multiple-choice comprehension versus sentence repetition. *Journal of Child Language*, 44(6), 1435–1457.
- Frizelle, P., Thompson, P. A., McDonald, D., & Bishop, D. V. (2018). Growth in syntactic complexity between four years and adulthood: Evidence from a narrative task. *Journal of Child Language*, 45(5), 1174–1197.
- Gillam, R. B., & Pearson, N. A. (2017). *Test of narrative language—second edition*. Pro-Ed.
- Goodman, R. (1997). The strengths and difficulties questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, 38(5), 581–586.
- Graham, H. R., Minhas, R. S., & Paxton, G. (2016). Learning problems in children of refugee background: A systematic review. *Pediatrics*, 137(6).
- Hamann, C., Chilla, S., Ibrahim, L. A., & Fekete, I. (2020). Language assessment tools for Arabic-speaking heritage and refugee children in Germany. *Applied PsychoLinguistics*, 41(6), 1375–1414.
- Kaltsa, M., Prentza, A., & Tsimpli, I. M. (2020). Input and literacy effects in simultaneous and sequential bilinguals: The performance of Albanian–Greek-speaking children in sentence repetition. *International Journal of Bilingualism*, 24(2), 159–183.
- Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman brief intelligence test—second edition (KBIT-2)*. American Guidance Service.
- Lenth, R. (2023). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.11.1-00001.
- Lesaux, N. K., Crosson, A. C., Kieffer, M. J., & Pierce, M. (2010). Uneven profiles: Language minority learners' word reading, vocabulary, and reading comprehension skills. *Journal of Applied Developmental Psychology*, 31(6), 475–483.
- Luchman, J. (2024). *domir: Tools to Support Relative Importance Analysis*. R package version 1.2.0.
- Mackay, E., Lynch, E., Sorenson Duncan, T., & Deacon, S. H. (2021). Informing the science of reading: Students' awareness of sentence-level information is important for reading comprehension. *Reading Research Quarterly*, 56, S221–S230.
- MacWhinney, B. (2000). *The Childes project: The database* (Vol. 2). Psychology Press.
- Marinis, T., & Armon-Lotem, S. (2015). Sentence repetition. In *Assessing multilingual children: Disentangling bilingualism from language impairment* (Vol. 13, pp. 95–124). Multilingual Matters.
- Meir, N. (2017). Effects of specific language impairment (SLI) and bilingualism on verbal short-term memory. *Linguistic Approaches to Bilingualism*, 7(3–4), 301–330.
- Meltzer, H., Gatward, R., Goodman, R., & Ford, F. (2000). *The mental health of children and adolescents in Great Britain*. The Stationery Office.
- Miller, J. F., & Chapman, R. S. (1981). The relation between age and mean length of utterance in morphemes. *Journal of Speech, Language, and Hearing Research*, 24(2), 154–161.
- Miller, J. F., & Iglesias, A. (2012). *Systematic analysis of language transcripts (SALT)*, research version 2012 [Computer software]. Middleton, WI: Salt Software, LLC.
- Mizumoto, A. (2023). Calculating the relative importance of multiple regression predictor variables using dominance analysis and random forests. *Language Learning*, 73(1), 161–196.
- National Academy of Sciences, Engineering, and Medicine. (2017). *Promoting the educational success of children and youth learning English: Promising futures* (p. e24677). National Academies Press.
- Nippold, M. A., Mansfield, T. C., Billow, J. L., & Tomblin, J. B. (2008). Expository discourse in adolescents with language impairments: Examining syntactic development. *American Journal of Speech-Language Pathology*, 17(4), 356–366.
- Paradis, J. (2011). Individual differences in child English second language acquisition: Comparing child-internal and child-external factors. *Linguistic Approaches to Bilingualism*, 1(3), 213–237.
- Paradis, J., & Kirova, A. (2014). English second-language learners in preschool: Profile effects in their English abilities and the role of home language environment. *International Journal of Behavioral Development*, 38(4), 342–349.

- Paradis, J., Rusk, B., Sorenson Duncan, T., & Govindarajan, K. (2017). Children's second language acquisition of English complex syntax: The role of age, input, and cognitive factors. *Annual Review of Applied Linguistics*, 37, 148–167.
- Paradis, J., Sorenson Duncan, T., Thomlinson, S., & Rusk, B. (2022). Does the use of complex sentences differentiate between bilinguals with and without DLD? Evidence from conversation and narrative tasks. *Frontiers in Education*, 6, 804088.
- Paradis, J., Soto-Corominas, A., Vitoroulis, I., Al Janaideh, R., Chen, X., Gottardo, A., Jenkins, J., & Georgiades, K. (2022). The role of socioemotional wellbeing difficulties and adversity in the L2 acquisition of first-generation refugee children. *Bilingualism: Language and Cognition*, 25(5), 921–933.
- Paradis, J., Soto-Corominas, A., Daskalaki, E., Chen, X., & Gottardo, A. (2021). Morphosyntactic development in first generation Arabic–English children: The effect of cognitive, age, and input factors over time and across languages. *Language*, 6(1), 51.
- Picot, G., & Sweetman, A. (2012). *Making it in Canada: Immigration outcomes and policies*. IRPP study 29. Institute for Research on Public Policy.
- Polišenská, K., Chiat, S., & Roy, P. (2015). Sentence repetition: What does the task measure? *International Journal of Language & Communication Disorders*, 50(1), 106–118.
- 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.
- R Core Team (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rojas, R., Iglesias, A., Bunta, F., Goldstein, B., Goldenberg, C., & Reese, L. (2016). Interlocutor differential effects on the expressive language skills of Spanish-speaking English learners. *International Journal of Speech-Language Pathology*, 18(2), 166–177.
- Rothman, J., Long, D., Iverson, M., Judy, T., Lingwall, A., & Chakravarty, T. (2016). Older age of onset in child L2 acquisition can be facilitative: Evidence from the acquisition of English passives by Spanish natives. *Journal of Child Language*, 43(3), 662–686.
- Rowe, L. W. (2018). Say it in your language: Supporting translanguaging in multilingual classes. *The Reading Teacher*, 72(1), 31–38.
- Sackett, G. P. (1978). *Observing behavior: Data collection and analysis methods* (Vol. 2). University Park Press.
- Scheidnes, M., & Tuller, L. (2019). Using clausal embedding to identify language impairment in sequential bilinguals. *Bilingualism: Language and Cognition*, 22(5), 949–967.
- Scott, C. M. (2009). A case for the sentence in Reading comprehension. *Language, Speech, and Hearing Services in Schools*, 40(2), 184–191.
- Scott, C. M., & Stokes, S. L. (1995). Measures of syntax in school-age children and adolescents. *Language, Speech, and Hearing Services in Schools*, 26(4), 309–319.
- Sirin, S. R., & Rogers-Sirin, L. (2015). *The educational and mental health needs of Syrian refugee children*. Migration Policy Institute.
- Squires, K. E., Lugo-Neris, M. J., Peña, E. D., Bedore, L. M., Bohman, T. M., & Gillam, R. B. (2014). Story retelling by bilingual children with language impairments and typically developing controls. *International Journal of Language and Communication Disorders*, 49, 60–74.
- Sorenson Duncan, T., & Paradis, J. (2020). Home language environment and children's second language acquisition: The special status of input from older siblings. *Journal of Child Language*, 47(5), 982–1005.
- Soto-Corominas, A., Daskalaki, E., Paradis, J., Winters-Difani, M., & Al Janaideh, R. (2022). Sources of variation at the onset of bilingualism: The differential effect of input factors, AOA, and cognitive skills on HL Arabic and L2 English syntax. *Journal of Child Language*, 49(4), 741–773.
- Soto-Corominas, A., Paradis, J., Rusk, B. V., Marinova-Todd, S., & Zhang, X. (2020). Oral language profiles of English second language learners in adolescence: Cognitive and input factors influence how they compare to their monolingual peers. *Studies in Second Language Acquisition*, 42(4), 697–720.
- Strand, S., & Demie, F. (2005). English language acquisition and educational attainment at the end of primary school. *Educational Studies*, 31(3), 275–291.
- Theunissen, M. H., de Wolff, M. S., & Reijneveld, S. A. (2019). The strengths and difficulties questionnaire self-report: A valid instrument for the identification of emotional and behavioral problems. *Academic Pediatrics*, 19(4), 471–476.
- Torregrossa, J., Listanti, A., Bongartz, C., & Marinis, T. (2024). Adding discourse to sentence repetition tasks: Under which conditions does bilingual children's performance improve? *Research Methods in Applied Linguistics*, 3(2), 100107.
- Unsworth, S., Brouwer, S., de Bree, E., & Verhagen, J. (2019). Predicting bilingual preschoolers' patterns of language development: Degree of non-native input matters. *Applied Psycholinguistics*, 40(5), 1189–1219.
- Wong, A. M. Y., Au, C. W. S., & Stokes, S. F. (2004). Three measures of language production for Cantonese-speaking school-age children in a story-retelling task. *Journal of Speech, Language, and Hearing Research*, 47(5), 1164–1178.
- Zebib, R., Tuller, L., Hamann, C., Ibrahim, L. A., & Prévost, P. (2020). Syntactic complexity and verbal working memory in bilingual children with and without developmental language disorder. *First Language*, 40(4), 461–484.