Bilingualism: Language and Cognition 22 (5), 2019, 949–967 © Cambridge University Press 2018 doi:10.1017/S1366728918000949

Using clausal embedding to identify language impairment in sequential bilinguals*

MAUREEN SCHEIDNES
Imagerie et cerveau Inserm 930, Université François Rabelais de Tours
LAURICE TULLER
Imagerie et cerveau Inserm 930, Université François Rabelais de Tours

(Received: July 17, 2017; final revision received: August 10, 2018; accepted: August 12, 2018)

Assessing language in sequential bilinguals (L2 children) for the identification of language impairment suffers from lack of appropriate standardized tools and from limited understanding of the developmental trajectories of these learners. This longitudinal study compares L2 children to children with SLI in order to better understand where these groups have overlapping performance and where they differ. An analysis of standardized test scores as well as frequency of clausal embedding and morphosyntactic errors in spontaneous speech was conducted with 22 English-speaking children (aged 6;9-12;7) acquiring French as an L2 in France and 19 monolingual French-speaking children with SLI (aged 6;5–12;11). The results revealed that L2 children used clausal embedding more often than the children with SLI, but both groups had similar rates of morphosyntactic accuracy. Facility with clausal embedding from early stages of development and continuing difficulty with morphosyntactic accuracy are argued to be characteristic of typical development in L2 children.

Keywords: sequential bilingual children, language impairment, spontaneous language, clausal embedding, morphosyntactic accuracy

1. Introduction

An increasing number of studies on simultaneous and sequential bilingual children have addressed the challenge of accurate identification of language impairment in these language learners (see Marinis, Armon-Lotem & Pontikas, 2017 and sources cited therein). The difficulty in recognizing atypical language development in bilingual school-age children is due in part to a lack of appropriate standardized tools and to a lack of knowledge about what typical bilingual developmental paths look like. Childcare professionals may interpret the stages that typical bilinguals go through as indicative of language impairment or attribute their language difficulties to the bilingual development context (Bedore & Peña, 2008; Paradis, Genesee & Crago, 2011). Furthermore, language experience variables, such as chronological age (Age), age of onset (AoO), length of exposure to the L2 (LoE) and socio-economic factors, are assumed to account for some of the documented variability in these learners, but little is known about the precise way in which these factors affect the developmental trajectories of typically compared to atypically developing bilinguals. The aim of the present paper is to contribute to the knowledge of sequential bilingual development through the longitudinal analysis of spontaneous language measures in the L2. A group of 22 sequential bilingual children without a diagnosis of language impairment (L1 English, L2 French, henceforth, L2 children) will be compared to 19 monolingual children with specific language impairment (SLI) in order to better understand which measures in spontaneous language samples are helpful in interpreting standardized test scores and which language patterns could be considered (a)typical in the L2 group. Special attention will be paid to the effect that LoE has on the L2 children's performance. The spontaneous language data presented here will lead us to argue that clausal embedding is less sensitive to LoE than other language properties.

* We would like to sincerely thank the children, families, and schools for their participation in this research. We also thank Philippe Prévost, Sandrine Ferré, Christophe dos Santos, Cécile Monjauze, Rasha Zebib, Hélène Delage, Eléonore Morin, Chloé Berruer, Claire Damourette, Mélanie Aillet, Aline Vanhaverbeke, Lucile Dupuy, Servane Galès, Sabrina Delord, Sarah Black, Nohad Abou Melhelm, and Nolwen Brunie for their help in the data collection and analysis. This research was supported by the *Agence Nationale de la Recherche* (Project ANR-08-BLAN-0328-01 for *Complexity and compensation in atypical language acquisition: A comparative approach*). Maureen Scheidnes is now at Memorial University of Newfoundland.

2. Background

2.2 Standardized assessments and problems with interpretation

There is general consensus in research on child bilinguals that standardized language assessments based on monolingual norm-referencing populations are not appropriate for the clinical evaluation (Armon-Lotem,

Address for correspondence:
Maureen Scheidnes, Department of Linguistics, Memorial University, St. John's, NL, A1B 3X9, Canada.
mscheidnes@mun.ca
Supplementary material can be found online at https://doi.org/10.1017/S1366728918000949

https://doi.org/10.1017/S1366728918000949 Published online by Cambridge University Press

2012; Bedore & Peña, 2008; Kohnert, 2010; Paradis, Genesee & Crago, 2011; Thordardottir, 2015). Yet, clinicians rely to a certain extent on standardized assessments for identifying language impairment. Seeing as standardized tests remain a staple of the assessment process, even for bilinguals, the major challenge for the clinician is how to interpret the standardized test results, knowing that many typically developing bilinguals (Bi-TD) obtain scores that are lower than those of their monolingual peers. One major difficulty in interpreting standardized test scores concerns the effect of language experience variables on language performance in bilinguals. While standardized tests use age as the benchmark for determining typical language performance, one cannot expect a school-age child whose L2 exposure has only just begun to perform like age-matched monolinguals of the target language on many aspects of language. Thus, additional methods of evaluation must be considered. These methods should allow for better interpretation of standardized test results so that clinicians can understand whether low standardized test scores from bilingual children indicate a language problem or are instead the result of test properties biased against bilingual learners. Ideally, these additional assessment methods would be neutral to the effect of language experience (e.g., Thordardottir & Brandeker, 2013; Kohnert, 2010).

2.3 Spontaneous language analysis: morphosyntactic accuracy and clausal embedding

Given the difficulty in interpreting bilingual standardized test scores against monolingual norms, researchers have suggested making spontaneous language samples an essential part of language assessment for bilinguals (Gutiérrez-Clellen & Simon-Cereijido, 2009). The question then becomes what aspects of spontaneous language should be studied. Prior direct comparisons of Bi-TD and monolingual children with SLI (Mo-SLI) have examined spontaneous language samples for the rate of accurate production of morphosyntactic properties such as tense morphology, object clitics and grammatical gender marking (Gutiérrez-Clellen, Simon-Cereijido & Wagner, 2008; Paradis & Crago, 2000, 2004; Paradis, 2004). These studies described similar performance between typically developing (TD) and clinical populations for these measures. Given this documented overlap, it may be that spontaneous language analyses that target these properties are not the best suited for detecting SLI in bilinguals, especially if the bilinguals are tested in their weaker language. However, prior work on a group of Spanish-speaking children in the US found that a model including MLU and the rate of utterances with at least one error in morphosyntax was a fair to good discriminator of Spanish-speaking children with language impairment (Simon-Cereijido & Gutiérrez-Clellen, 2007). The authors tested this model and obtained similar results on a subset of bilingual children who, according to parental ratings, had "better than limited English proficiency and somewhat frequent English use and exposure" (Simon-Cereijido & Gutiérrez-Clellen, 2007, pp. 329–330). The combination of MLU and error rates worked because some children with language impairment tended to produce shorter utterances, thereby producing fewer contexts in which they might make an error. These authors suggested that the children who were not correctly identified by error rates were relying on simple syntax, which is reflected in lower MLU. These findings suggest that monolingual and bilingual children with SLI make more morphosyntactic errors in spontaneous language than their unaffected peers, but that avoidance strategies used by some children with SLI may result in low error rates (Hamann & Tuller, 2014).

Prior research on monolingual SLI has indeed revealed avoidance strategies in affected learners that led to reduced MLU as well as lower frequency of multiclausal structures. Studies have shown that affected children make more errors than TD monolingual speakers when producing embedded clauses and also tend to avoid these structures altogether (Hamann & Tuller, 2014; Marinellie, 2004; Owen & Leonard, 2006; Reilly, Losh, Bellugi & Wulfeck, 2004; Scott, 2003; Tuller, Henry, Sizaret & Barthez, 2012). These results suggest that clausal embedding is difficult for children with SLI.

2.4 Impact of length of exposure and age on clausal embedding

In addition to targeting properties that pose problems for children with SLI, language assessments would ideally target properties that emerge early in Bi-TD. If a certain property developed somewhat uniformly and independently of LoE in Bi-TD, but was much slower to develop in bilingual children with SLI (Bi-SLI) or was prone to error or avoidance strategies, then this property would be ideal for the differential diagnosis of SLI in both simultaneous and sequential bilinguals. These properties could then discriminate between Bi-TD and Bi-SLI during all stages of development, regardless of variation in LoE. Interestingly, previous research suggests that clausal embedding is used frequently by Bi-TD and emerges early in these learners (Belletti & Hamann, 2004; Gavruseva & Lardiere, 1996; Haznedar, 2003; Lakshmanan & Selinker, 1994). For example, the L2 child Dasha (L1 Russian, L2 English) produced her first embedded clause after 2.5 months of exposure to English (Gavruseva & Lardiere, 1996). Belletti and Hamann (2004) reported that their two young simultaneous bilinguals used clausal embedding at a rate that was comparable to their monolingual peers. Although only a handful of studies on clausal embedding

https://doi.org/10.1017/S1366728918000949 Published online by Cambridge University Press

in this learner group exist, the available evidence indicates that L2 children produce these structures early in acquisition, while children with SLI tend to avoid them. Early acquisition reveals a certain facility with these structures that is translated into high frequency of use. Taken together, these results suggest that the analysis of clausal embedding in spontaneous language samples may be neutral to the effects of LoE, while still being able to identify cases of language impairment.

Indeed, recent work has begun to focus on tools that minimize the influence of input quantity on performance. These include methods such as non-language mediated processing (Kohnert, Windsor & Ebert, 2009), novel word learning (Kapantzoglou, Restrepo & Thompson, 2012) and sentence (Marinis & Armon-Lotem, 2015) and non-word repetition (Gutiérrez-Clellen & Simon-Cereijido, 2010). For example, in a study by Thordardottir and Brandeker (2013) on simultaneous bilinguals in Montreal, Bi-TD children scored significantly higher on sentence repetition and non-word repetition tasks compared to a group of Bi-SLI. These tasks discriminated between the two groups, regardless of individual differences in quantity of exposure. However, the sentence repetition task was more sensitive to exposure than the non-word repetition task. Other than the fact that the test items increased in length, syntactic complexity and number of propositions, there was little information given about the exact language properties targeted by the sentence repetition task (e.g., lexical frequency, clause type). It is possible that adjusting certain language properties in sentence repetition tasks would enhance their discriminatory power while making them less sensitive to variation in input quantity (see Marinis & Armon-Lotem, 2015). An analysis of spontaneous language measures may lead to the identification of language properties that could be targeted in a more controlled manner by sentence repetition tasks.

In addition to exposure effects, age has been shown to influence bilingual language performance. While older bilinguals have been shown to have stronger performance in vocabulary (Golberg, Paradis & Crago, 2008) and verb morphology (Paradis, 2011), they can be at a particular disadvantage when being compared to a monolingual norm. Thordardottir and Juliusdottir (2013) noted that older children have to learn more than younger ones in order to reach age-appropriate language expectations. Older monolingual children are expected to have larger vocabularies, a better mastery of irregular morphology, and use complex syntax; however, an older bilingual child may not have had enough time to acquire all of these properties to the same degree. Thus, in order to better identify (a)typical bilingual development, it is important to understand for which properties older and younger bilingual children are able to perform like age-matched peers. While accounting for the respective roles of LoE and age are crucial to this endeavor, the two variables have often been highly correlated in past research (see discussion in Stevens, 2006) and therefore difficult to tease apart.

3. Research Objectives and Questions

The objective of this paper is to better understand the similarities and differences in L2 children compared to monolingual children with SLI. More specifically, we wish to examine how these learner groups compare on standardized test performance and on measures of clausal embedding and morphosynactic errors in spontaneous speech. Moreover, we aim to explore the effect of LoE and age on these language measures. To do this, we recruited L2 children who varied in LoE, but for whom LoE and age were not highly correlated. In addition, standardized test and spontaneous language data were collected twice at an interval of 12 months, in order to measure the impact that another year of exposure had on the results.

We aim to address the following research questions:

1. How do typically developing (TD) L2 children and children with SLI compare on standardized language assessments? Does the outcome of this comparison change over a period of 12 months?
2. Do TD L2 children and monolingual children with SLI differ from each other in morphosyntactic accuracy and clausal embedding? Does the outcome of this comparison change over a period of 12 months? What is the impact of LoE and age on the language performance in the L2 children?
3. Can these particular spontaneous language measures lead to better interpretation of standardized test results for L2 children?

In order to address these research questions, the L2 and SLI will also be compared to two TD monolingual control groups: a group of 4-year-old (TD4) and a group of 8- and 11-year-olds (TD8&11). The 4-year-old controls are approximate language matches for the children with SLI and the 8- and 11-year-old controls serve as age matches to the L2 and SLI.

Moreover, language measures for which the L2 show little overlap with the SLI, but considerable overlap with the age-matched controls (TD8&11), and for which there is minimal effect of LoE will be considered as promising for discriminating between Bi-TD and Bi-SLI. On the other hand, measures for which the L2 show major overlap with the SLI, little overlap with TD monolinguals and an effect of LoE will be considered less useful in identifying Bi-SLI. The effect of LoE will be examined cross-sectionally using correlation analyses in addition to comparing longitudinal data collected twice at a 12-month interval.

https://doi.org/10.1017/S1366728918000949 Published online by Cambridge University Press

Table 1. *Presentation of child participants.*

	L2 Children	Children with SLI	TD4	TD8	TD11
Number of participants at T1	22	19	14	12	12
Number of participants at T2	22	19	–	–	–

Note: The control data collection was cross-sectional. TD = typically developing.

Table 2. *Characteristics of the L2 population (N = 22).*

	T1			T2		
Variable	Mean	SD	Range	Mean	SD	Range
Age	9;8	1;9	6;9–12;7	10;8	1;10	7;9–13;8
LoE	2;6	1;1	0;11–4;0	3;6	1;1	1;10–4;11
AoO	7;0	1;0	4–10			
WEU[a]	3.4	1.5	1–5			
SES[b]	1.6	0.8	1–3			

Notes: SD = Standard Deviation, AoO = Age of Onset, LoE = Length of Exposure, WEU = Weekly Use of English.
[a]Weekly Use of English from the parental questionnaire. Parents were asked "Approximately how many hours a week does your child use English?" and had to choose from the following 6-point scale: 0 = Never, 1 = <10 hrs, 2 = 10–20 hrs, 3 = 20–30 hrs, 4 = 30–40 hrs, 5 = 40–50 hrs.
[b]Socio-economic status was determined by the highest education level obtained by the mother: 1 = secondary, 2: undergraduate degree, 3: postgraduate degree.

4. Method

4.1 Participants

The analysis of the present paper focuses on data from 22 L2 children, 19 monolingual children with SLI, 14 4-year-old monolingual children with typical development (TD4), 12 8-year-old monolingual TD children and 12 11-year-old monolingual TD children. The groups and number of participants are summarized in Table 1 and will be presented in detail below. Data collection with the L2 children and the children with SLI was carried out twice at a 12-month interval, while the monolingual control data was cross-sectional.

L2 children

The child L2 population consists of 22 children (16 girls) from the British Isles whose parents are native English speakers and whose families immigrated to France when the children were no younger than 4 (see Table 2 for group characteristics). We reported on this same group of children in Scheidnes and Tuller (2016). These children had no prior knowledge of French before their arrival. At the time of testing, all children were living near Tours, France and attending French-speaking schools. Many of the L2 children were recruited with help from the regional school district (Académie d'Orléans-Tours) who put us in contact with schools in which students from outside of France had recently enrolled. School principals put us in touch with families whose children fit the desired language profile. Other families were recruited via word-of-mouth or advertisements in local English-speaking magazines. All families lived in small towns or farms in areas around Tours, which are inhabited almost exclusively by monolingual speakers of French.

For inclusion in the study, children had to meet the following criteria: (1) Age of Onset (AoO, i.e., age of arrival in France) later than 4;0, (2) no exposure to a third language, and (3) no parental concern about language development (except in cases of otitis media, which we return to below). Two other children met language profile requirements, but information from a parental questionnaire led us to exclude these individuals from the present study for the following reasons: (1) one had been diagnosed with moderate hearing loss between T1 and T2 and; (2) one underwent evaluation between T1 and T2 for a possible developmental disorder the results of which were not known before the end of the study. Parents completed a family questionnaire, which provided us with information concerning the child's use of English outside of school, family history of language impairment, and socio-economic status (parental education).

All children spoke French at school; however, there was some variation in the total number of hours that each child spent using English per week (Weekly English Use, WEU, in Table 2). There was little variation in terms of SES (parental education) or self-reported parental proficiency in French, in part because several children were siblings.[1] The variation in SES being limited, we will consider this group to be relatively homogenous in this respect. The family questionnaires also revealed that none of the children had a diagnosis of language impairment or were in therapy for language-related disorders. Except for cases of otitis media, parents did not express concern about their child(ren)'s development. Four children (PUS, BEA, LAD, and ALP) were reported to have had past or ongoing issues with otitis media. As this is a common problem in children, we did not consider this to be grounds for exclusion from the study. However, the child ALP was not included in the group statistical analyses because he obtained several low test scores and his parental report revealed a long history of severe otitis media that lasted into late childhood and was on-going during the study period. We did the same for the child SAJ because of very low test scores (digit span, analytic reasoning and

[1] The following groups of children are siblings: RHR & MER; PUS, WRS, & HOS; BEA & HEA; LEP & ALP; LAD & DAD; LOS, HAS, & MAS.

https://doi.org/10.1017/S1366728918000949 Published online by Cambridge University Press

Table 3. *Comparing the L2 and SLI on general characteristics.*

		Group						Significance level			
		L2[a]			SLI			*p*	95% CI[b]		*d*
	Variable	M	SD	Range	M	SD	Range		Lower	Upper	
T1	Age	9;5	1;8	6;9–12;3	9;9	1;11	6;5–12;11	.58	−18	10	0.18
	Raven's[c]	3.5	1.1	1 – 5	3.3	1.4	0 – 5	.74	−0.7	1.0	0.11
T2	Age	10;5	1;8	7;9–13;3	10;10	1;11	7;7–14;2	.52	−18	10	0.21

Notes: [a]SAJ and ALP are excluded from the group comparisons.
[b]Confidence Interval.
[c]The Raven's Matrices scores (0–5) correspond to the following percentiles: 5: ≥ 75, 4: 50 < 75, 3: 25 < 50, 2: 10 < 25, 1: 5 < 10, 0: < 5).

language, see Scheidnes & Tuller, 2016) in both languages and a family history of language problems. However, we include these children in the individual analyses as their profiles are interesting to compare with those of the other children.

As Table 2 shows, the L2 children varied in Age (chronological age at time of testing), Age of Onset (AoO, i.e., age of arrival in France) and Length of Exposure (LoE, i.e., the time the child had lived in France at the time of testing). Age and AoO were very highly correlated ($r = 0.80, p < .001$, 95% Confidence Interval (CI) [0.55, 0.92]). Due to this extreme overlap, we will consider AoO to be included within the factor Age. We therefore acknowledge that any variation attributed to the Age variable may also be influenced by AoO. Although the correlation between Age and LoE is positive, it is not significant ($r = .27$, $p = .25$, 95% CI [−0.19, 0.64]). Indeed, this group includes older children with shorter LoE, a profile type that is often missing from studies on bilingual children. The correlation between LoE and AoO does not reach significance either ($r = -.37$, $p = .11$, 95% CI [−0.70, 0.09]).

Children with SLI and control groups

A group of 19 monolingual French-speaking children with SLI (16 males, 3 females) were compared to the L2 (see Table 3 for a comparison of the L2 and SLI groups). As Table 3 shows, the L2 children and the children with SLI were matched on mean age and they did not differ in non-verbal performance (Raven's Progressive Matrices, Raven, Raven & Court, 1998).

The children with SLI were recruited from language reference centers and special language classes located in three different French cities (Tours, Angers, Poitiers), where they had been diagnosed with SLI following usual inclusionary and exclusionary criteria (Leonard, 1998). Audiograms were performed to ensure normal hearing, IQ was tested (all participants in this study have a performance IQ superior to 85), and a neuropediatric examination was carried out to rule out any acquired source for language impairment. Children with SLI who scored below the 5th percentile on Raven's Progressive Matrices were excluded unless an alternative normal nonverbal score was available (i.e., a normal WISC score, which was the case for one of the children). Thirteen of the 19 children with SLI were diagnosed at the same university teaching hospital in Tours following clinical examination and standardized tests targeting all areas of language (articulation and phonology, EEL, Chevrie-Muller, Simon & Decante, 1980; expressive lexicon and expressive morphosyntax, ELO, Khomsi, 2001; receptive lexicon, TVAP, Deltour & Hupkens, 1980; receptive morphosyntax, NSST, Lee, 1971). A cut-off for pathological language performance of −1.65 *SD* is used by clinicians at this institution, as is relatively common practice in France. Eighteen out of the 19 children had received a diagnosis of "phonologic-syntactic" SLI and one child was diagnosed as having "lexical semantic" SLI. At the time of testing, the children with SLI were attending ordinary French schools, though three of them were in a special language class, and all of them were in speech-language therapy.

Table 4. *General characteristics of the TD control groups.*

TD Control Group	N	Mean age (SD)
TD4	14	4;1 (0;1)
TD8&11	24	9;10 (1;8)

Control groups consisted of TD monolingual French-speaking children (see Table 4). Fourteen 4-year-old children (TD4) were used as approximate language matches to the SLI. A group of 24 children (12 8-year-olds and 12 11-year-olds) were used as approximate age matches (TD8&11). These two age groups were combined because their spontaneous language measures did not differ significantly and because their mean age when combined was similar to the mean age of the SLI and L2 at T1. The control data collection was cross-sectional.

https://doi.org/10.1017/S1366728918000949 Published online by Cambridge University Press

4.2 Materials

Standardized assessments

The standardized measures reported here were selected because they are used by speech-language pathologists (SLPs) in France and in the UK to assess children for language impairment. Findings from these standardized tasks were previously reported in detail for the L2 group (Scheidnes & Tuller, 2016). In the current paper, we refer to some of these previous results, but our main focus here is on the L2-SLI comparison and on spontaneous language measures. The French standardized measures included two tests from the BILO-3C (Khomsi, Khomsi, Pasquet & Parbeau-Gueno, 2007). Morphosyntax was evaluated using a sentence completion task (Sentence Compl) during which the child sees two pictures on a computer screen and hears (the beginning of) a sentence describing the first picture, and then is prompted to complete the sentence according to the second picture (ex. Here, the boy pushes the girl, and there the boy . . . [is pushed by the girl]). Phonology was evaluated using a (real) word repetition task (Word Rep). Although non-word repetition is used more frequently to reduce the effect of lexical familiarity, the BILO authors assert that many of the test items are most likely unknown to the children due to low frequency of use in French. The third task was a picture-pointing task of receptive vocabulary (ELOLA, de Agostini, Metz-Lutz, van Hout, Chavance, Deloche & Pavao-Martins, 1998, henceforth, Vocab), which is a French adaptation of the short version of the British Picture Vocabulary Scale (Dunn, Dunn, Whetton & Burley, 1997), and which was normed on 200 French children aged 4 to 12, half of which came from schools in high SES neighborhoods and half from schools in low SES neighborhoods. Notably, the children from the latter schools included children of immigrant parents who were bilingual (36/200 children). Unfortunately, the test instruction manuals do not provide information about reliability or validity, but these tools were chosen in consultation with experienced SLPs working in a diagnostic center for language impairment in children. A cut-off for pathological performance of −1.65 *SD* (percentile 5) is commonly used by clinicians in France, and indeed is explicitly given as the cut-off for the definition performance which is "too low" by the authors of the ELOLA. Again, these choices reflect current clinical practice in France and we acknowledge that ideally these cut-offs would be established empirically (Dollaghan & Horner, 2011).

For English, we used the Core Language Subtests (Core Lg. Subtests) from the CELF-4 UK version (Semel, Wiig & Secord, 2006). The cut-off score for the composite Core Lg. Score is set at −1 *SD* (85 standard score). If a child obtains a score below this cut-off, the authors of the CELF recommend further evaluation. The reliability and validity evidence for the CELF-4 are very good (and are presented in detail in Semel et al., 2006). We emphasize that both the CELF and the French standardized tests were standardized either exclusively or primarily on monolingual children, and that there are no standardized tests with norms for bilingual English–French children.

Spontaneous language measures

In addition to providing a baseline reading on the children's French, the standardized scores will be compared to spontaneous language measures to better understand to what extent the L2 children's language can be underestimated by standardized assessments and which spontaneous language measures allow for a more accurate interpretation of the L2 children's abilities. A spontaneous language sample was obtained via a structured conversation between a research assistant and each child participant. Narratives were elicited by asking for a description of a book or movie that the child had read or seen recently. Previous studies (Reed, 2005) have shown that this type of setting is favorable for eliciting complex syntax. About 60 utterances per child were transcribed and coded via the CHILDES system (MacWhinney, 2000). Each subordinate clause was coded for depth of embedding. Subordinate clauses which were not embedded in a matrix clause were coded as zero-level clauses [0], as in (1). Subordinate clauses which were embedded in a matrix clause were labeled level-1 clauses [1], as in (2). Clauses which were embedded within a level-1 clause were coded level-2 clauses [2], as in (3) and so on.

(1) [$_{CP}$ That Max lives in Paris] Level 0

(2) Lea believes [$_{CP}$ that Max lives in Paris] Level 1

(3) Lea believes [$_{CP}$ that everybody [$_{VP}$ thinks [$_{CP}$ that Max lives in Paris]]] Level 2 (from Tuller et al., 2012)

The analysis of the frequency of subordination and level of embedding in the spontaneous language samples was based on the following measures: 1) rate of subordination (Sub), 2) rate of complex utterances (Cplx), 3) clausal density (CDens), and 4) rate of deep embedding (DeepE). These measures are detailed in Table 5.

The rate of utterances with at least one error was calculated as a measure of morphosyntactic accuracy. Therefore, the higher the number, the more errors the child made, and the less accurate the child was. The following morphosyntactic error types (substitution, omission and addition) were attested and included in the overall morphosyntactic accuracy count: 1) verb morphology, 2) object clitics, 3) grammatical gender marking (see Scheidnes, 2015 for details), 4) subject pronouns, 5) pronominal reference (i.e., when the child produces a pronoun that has no identifiable referent), 6) prepositions,

https://doi.org/10.1017/S1366728918000949 Published online by Cambridge University Press

Table 5. *Calculation of clausal embedding measures.*

1.	Rate of subordination (Sub)	Percentage of total number of subordinate clauses out of the total number of verbal utterances[a]
2.	Rate of complex utterances (Cplx)	Percentage of total number of utterances with at least one subordinate clause out of the total number of verbal utterances
3.	Clausal Density (CDens)	Ratio of total number of clauses (all types, including 0-level) to total number of verbal utterances
4.	Rate of deep embedding (DeepE)	Percentage of total number of clauses embedded at level-2 or higher out of the total number of subordinate clauses

Note: [a]A verbal utterance is one that contains at least one verb, as opposed to an utterance without a verb. A clause is defined as a verb and its arguments. All verbal utterances contain at least one clause, but could in fact consist of multiple clauses.

7) adverb of negation 8) definiteness, 9) word order, and 10) complementizer errors.

4.3 Procedure

Testing took place in a quiet room at the child's school or home with trained native-speaking research assistants (native French- or English-speaking respectively). Research assistants administered tests in one language only: either French or English. The standardized BILO-3C battery is computerized, so the sentence completion and word repetition test items had been pre-recorded and were presented to the child via a laptop. During data collection, the research assistant recorded the participant's score during the tests and responses were also audio-recorded for later scoring verification. Any discrepancies that were found during scoring verification were decided by consensus after discussion with other native-speaking research assistants and researchers.

For the spontaneous language collection, the assistant began the session by asking the child to tell a story based on a wordless picture book containing six pictures (ECL-C, Khomsi & Nanty, 2001). This beginning was used only as an ice-breaker after which a narrative was elicited by asking each child to describe the events in a book or movie that the child had read or seen recently. This part of the data collection was also audio-recorded. Spontaneous language samples were transcribed by native French-speaking research assistants (some of whom had been present during data collection), linguists, and by the authors of this paper (native English speakers, near-native French-speakers). The transcriptions that were completed by non-native speakers were verified in their entirety by a native-speaking assistant or linguist (30 out of 44 files). At least ten percent of all remaining transcriptions were verified by another native speaker. When there were disagreements between two transcribers, decisions were made by committee, following group discussions.

Concerning the evaluation of morphosyntactic accuracy, grammaticality judgments were provided by research assistants and linguists who were native speakers of French. Very often these decisions were either made by committee or submitted for review by another linguist or research assistant. In cases where there was no consensus, the utterance was considered target-like, thus giving the benefit of the doubt to the child.

4.4 Statistics

The distributions in the spontaneous language data generally did not meet assumptions of normality, which are required for parametric statistical models. Therefore, robust methods, which include bootstrapping and trimmed means, were adopted for statistical analyses (see for example, Wilcox, 2003, 2005, 2010). These methods reduce problems associated with loss of power due to violations of the aforementioned assumptions while keeping the chances of committing a Type 1 error at the nominal level. The software package *R* (R Development Core Team, 2014) was used to carry out statistical tests. Means were trimmed by 20% and bootstrapping samples were either 599 or 2000, according to the recommendations given by Wilcox (2005) and Larson-Hall (2010).[2] Measures that are expressed as percentages were first arc-sine transformed before submission to statistical tests. Family-wise error was controlled for using the calculation described in Wilcox (2005, p. 302, function *mcppb20*) and the critical (corrected) *p*-value will be reported for each test. Effect sizes were calculated using Cohen's *d* for mean difference (*compute.es*, Del Re, 2013). Following Cohen (1988), *d* values of about 0.2 are considered small, ≥0.5

[2] Bootstrap defaults are as follows: For the omnibus Anova: *t1waybt* (599), for post-hoc tests: *mcppb20* (2000), for comparing two paired means: *rmmcppb* (2000), for comparing two independent means: *trimpb2* (2000), and for correlations: *corb* (599). Degrees of freedom are not associated with bootstrapping and the *t* statistic is not calculated by the functions *rmmcppb* and *trimpb2* and therefore will not be reported (Larson-Hall, 2010; Field, Miles & Field, 2012). These functions were developed for *R* by (Wilcox & Schönbrodt, 2014).

https://doi.org/10.1017/S1366728918000949 Published online by Cambridge University Press

Table 6. *French standardized test scores (raw scores in %) at T1 and T2.*

		L2			SLI			Significance level			
Time	Test	Mean	SD	Range	Mean	SD	Range	95% CI		*p*	*d*
T1	SentCpl	38	22	0–81	34	13	8–61	-0.1	.15	.71	0.2
	WRep	85	8.4	66–95	60	18	25–88	0.18	0.40	<.001	1.8
	Vocab	61	12	44–91	66	8.9	50–81	−0.13	0.02	.15	0.4
T2	SentCpl	56	17	28–97	47	18	14–83	−0.04	0.20	.23	0.5
	WRep	92	4.7	83–100	71	16	43–100	0.19	0.39	<.001	1.9
	Vocab	71	11	41–94	72	6.3	59–84	−0.05	0.06	.86	0.1

Note: SentCpl = Sentence Completion, WRep = Word Repetition, SD = Standard Deviation, CI = Confidence Interval, *d* = Cohen's *d*
Results for the L2 group are from Scheidnes and Tuller, 2016. The L2 group includes 20 children here since ALP and SAJ are excluded from group analyses.

Table 7. *French standardized test scores (z-scores) at T1 and T2.*

		L2			SLI			Significance level			
Time	Test	M	SD	Range	M	SD	Range	95% CI		*p*	*d*
T1	SentCpl	−2.7	2.8	−8.1–1.3	−3.0	2.0	−8.5–0.5	−1.4	1.9	.71	0.1
	WRep	−2.4	2.2	−8.1–0.6	−8.3	6.0	−23.2–2.1	2.6	8.5	<.001	1.3
	Vocab	−1.1	1.1	−3.1–1.3	−0.6	1.1	−4.0–1.1	−1.2	0.1	.13	0.5
T2	SentCpl	−1.0	1.7	−4.4–2.9	−2.2	2.0	−6.4–1.4	−0.1	2.2	.06	0.7
	WRep	−0.8	1.1	−3.4 – 1.0	−6.4	4.8	−19.0–0.6	3.1	7.1	<.001	1.6
	Vocab	−0.3	1.2	−3.0–1.6	−0.2	1.0	−1.8–1.8	−0.7	0.7	.92	0.1

Note: SentCpl = Sentence Completion, WRep = Word Repetition, SD = Standard Deviation, CI = Confidence Interval, *d* = Cohen's *d*.
Results for the L2 group are from Scheidnes and Tuller, 2016. The L2 group includes 20 children here since ALP and SAJ are excluded from group analyses.

are considered medium, and ≥0.8 and are considered large. Diagnostic metrics were calculated using the online calculator at www.cebm.utoronto.ca (Dollaghan & Horner, 2011).

5. Results

5.1 Standardized Tests

Comparing the L2 and SLI groups on standardized tests allows us to better understand the extent of the overlap between these two groups on tests that are used to diagnose SLI. Furthermore, this allows us to better ascertain the risk of misdiagnosis of language impairment in L2 children and to ameliorate methods of differential diagnosis in this population. The results of the sentence completion (Sentence Compl) task did in fact reveal considerable overlap between the L2 children and children with SLI in terms of raw and z-scores, especially at T1 (see Tables 6 and 7 for descriptive statistics and independent t-test results).[3]

[3] Recall that ALP and SAJ are excluded from analyses of group results.

While there were stark similarities between the L2 children and children with SLI in the standardized tests results, there were also some notable differences: The L2 children had much stronger scores on the word repetition (Word Rep) task compared to the children with SLI at both T1 and T2. In Scheidnes and Tuller (2016), we showed that the L2 children displayed significant improvement over the 12-month interval on z-scores of all three tests; however, the children with SLI only improved on Sentence Compl (SLI: Sentence Compl: $p = .04$, $d = 0.41$, 95% CI [−1.89, −0.07], Word Rep: $p = .12$, $d = 0.35$, 95% CI [−4.65, 0.58]; Vocab: $p = .21$, $d = 0.39$, 95% CI [−0.97, 0.21]). These findings suggest that while there was considerable overlap between the L2 children and children with SLI for morphosyntax and vocabulary, the L2 children showed more overall improvement over time.

5.2 Spontaneous language measures

Morphosyntactic accuracy

In this section, we compare the L2 children and children with SLI with respect to the analysis of morphosyntactic accuracy in spontaneous language. Morphosyntactic accuracy was measured using the rate of utterances

https://doi.org/10.1017/S1366728918000949 Published online by Cambridge University Press

Table 8. *Error rates (%) by group at T1 and T2.*

Time		L2	SLI	TD4
T1	Mean	27%	25%	16%
	SD	13	13	8
	Range	9–52%	7–58%	4–30%
T2	Mean	22%	18%	
	SD	13	9	
	Range	3–57%	5–33%	

Note: SD = Standard Deviation, TD4 = 4-year-old controls.
The L2 group includes 20 children here since ALP and SAJ are excluded from group analyses.

with at least one error (Error Rate). For this measure, the L2 children and children with SLI were also compared to the 4-year-old control group (TD4). The older control group (TD8&11) essentially made no errors ($M = 1.37\%$, $SD = 1.68$) and therefore, statistical analyses were not informative. We first present the group comparisons, followed by the intra-group comparisons for the longitudinal analysis. Recall that data was collected twice, at a 12-month interval (T1 and T2). Both the L2 children and the children with SLI made frequent morphosyntactic errors at T1 and T2 (see Table 8). Their error rates were higher than those of the TD4 group, but this difference reached significance at T1 only, according to a one-way ANOVA (T1: $F = 4.3$, $p = .02$; T2: $F = 1.10$, $p = .37$). Post-hoc tests showed that the L2 children and children with SLI did not differ significantly for morphosyntactic accuracy at T1 ($p = .59$, 95% CI [−0.10, 0.116], $d = 0.07$), but the L2 children made significantly more errors than the TD4 group at T1 ($p = .007$, 95% CI [0.01, 0.24], $d = 0.9$). The children with SLI made more errors than the TD4 group, but this difference was not significant at T1, unless we consider an uncorrected alpha of .05 ($p = .027$, 95% CI [−0.002, 0.21], $d = 0.78$).[4]

Turning to the longitudinal analysis of morphosyntactic accuracy, both the L2 children and children with SLI significantly reduced their mean error rates over the 12-month period (L2: $p = .037$, 95% CI [0.006, 0.13], $d = 0.34$; SLI: $p = .01$, 95% CI [0.02, 0.14], $d = 0.61$). Thus, children in both groups produced fewer errors over time.

Frequency of clausal embedding

In this section, we examine measures of clausal embedding, thus allowing us to observe how L2 children compare to children with SLI in terms of the frequency of complex syntactic structures in spontaneous language. Cross-sectional data from both control groups (TD4 group and TD8&11) were used in this analysis. We first present the group comparisons, followed by intra-group comparisons for the longitudinal analysis. We then report diagnostic accuracy metrics of the clausal embedding measures.

A series of one-way ANOVAs revealed a significant group effect for complex utterances ($F = 4.47$, $p = .02$) and clausal density at T1 ($F = 4.01$, $p = .03$). Post-hoc tests at T1 revealed a significant difference between the L2 children and children with SLI for complex utterances ($p < 0.007$, see Table S1 for outcomes of the statistical tests). Despite only one significant difference at T1, effect sizes show moderate to large differences between the L2 children and children with SLI. Furthermore, comparing boxplots in Figure 1, one can see that the overall variation in the L2 children closely resembles that of the TD8&11, thus suggesting that the L2 children are performing at age-appropriate levels. On the other hand, the children with SLI have overall performance that more closely resembles the TD4 group.

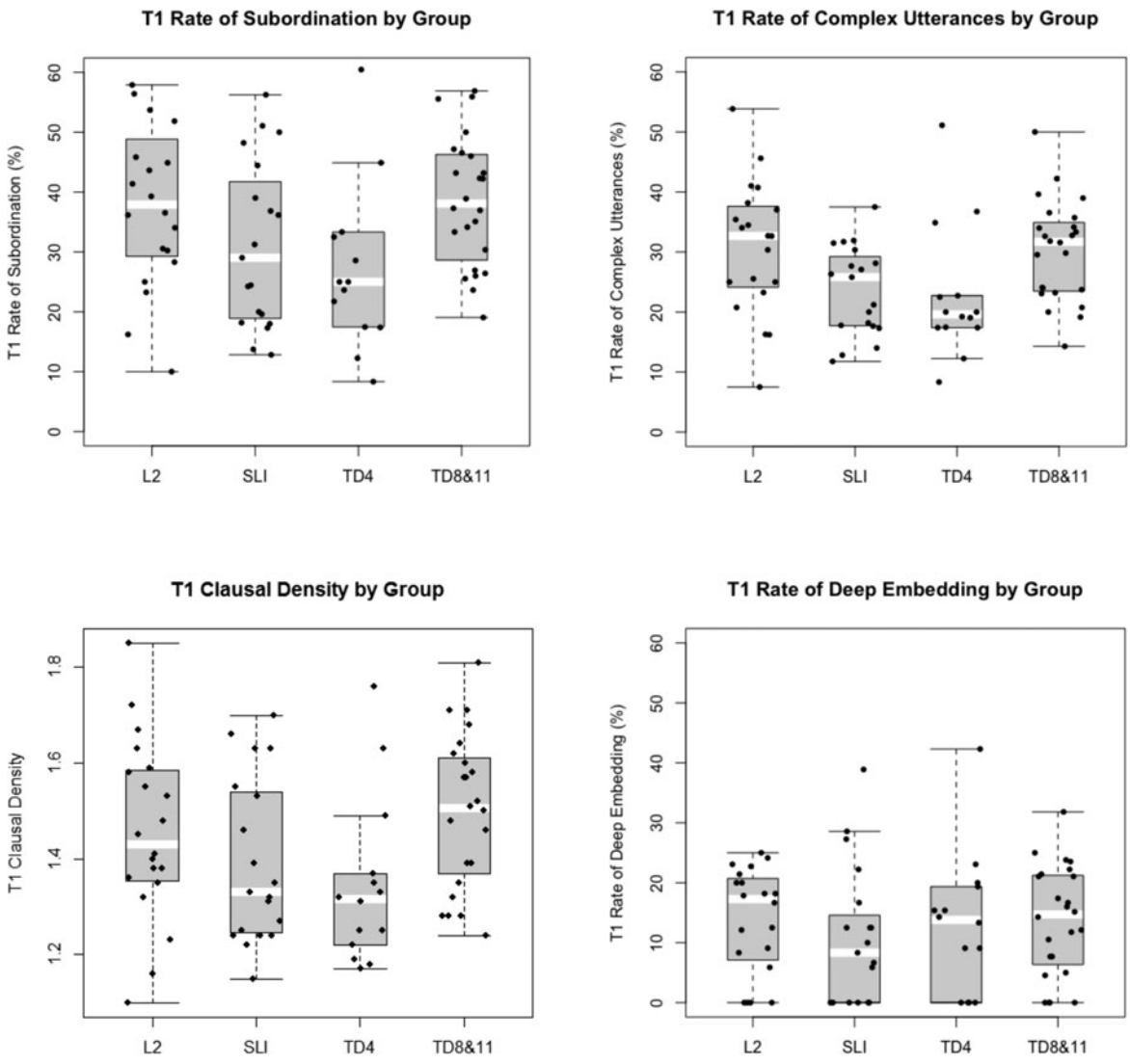
At T2, there was stronger statistical evidence suggesting that the L2 children and children with SLI differ with respect to clausal embedding. One-way ANOVAs revealed significant group effects for all four measures (complex utterances (Cplx): $F = 4.30$, $p = .03$, rate of subordination (Sub): $F = 6.07$, $p = .01$, clausal density (CDens): $F = 7.47$, $p = .005$, deep embedding (DeepE): $F = 7.24$, $p = .009$). Post-hoc tests at T2 showed significant differences between the L2 children and children with SLI for two out of the four measures (Sub: $p < 0.001$, 95% CI [0.03, 3.11], $d = 1.24$; DeepE: $p < 0.001$, 95% CI [0.12, 0.44], $d = 1.36$). L2-SLI differences were found for the other two measures; however, these did not reach significance as defined by the corrected alpha. Again, the effect sizes for all measures were medium to large ($d > 0.55$). Furthermore, the children with SLI had significantly weaker performance compared to the TD8&11, whereas there were no significant differences between the L2 and the TD8&11 at T2 (see Figure 2 and statistical results in Table S2). Although performances between the L2 children and children with SLI did not consistently differ significantly for all clausal embedding measures, the pattern observed clearly diverged from the one based on measures of morphosyntactic accuracy: the L2 children and children with SLI consistently performed similarly in terms of error rates, whereas the L2 children used embedding more often than the children with SLI.

When comparing clausal embedding frequency at T1 and T2, there was no significant improvement in either group from T1 to T2. We return to this result in the section on the impact of length of exposure (LoE).

Since there are significant group differences for clausal embedding measures, we analyzed the overall diagnostic accuracy metrics of these measures (see Tables 9 and 10). This type of analysis is generally carried out between bilingual groups (i.e., Bi-TD and Bi-SLI); however, given the present data set, the metrics could prove helpful in

[4] Critical p-value = .017.

https://doi.org/10.1017/S1366728918000949 Published online by Cambridge University Press



NOTES: Boxplots overlaid with dot charts that represent individual performance (Larson-Hall, 2010, p. 188). ALP and SAJ are excluded from L2 group analyses.

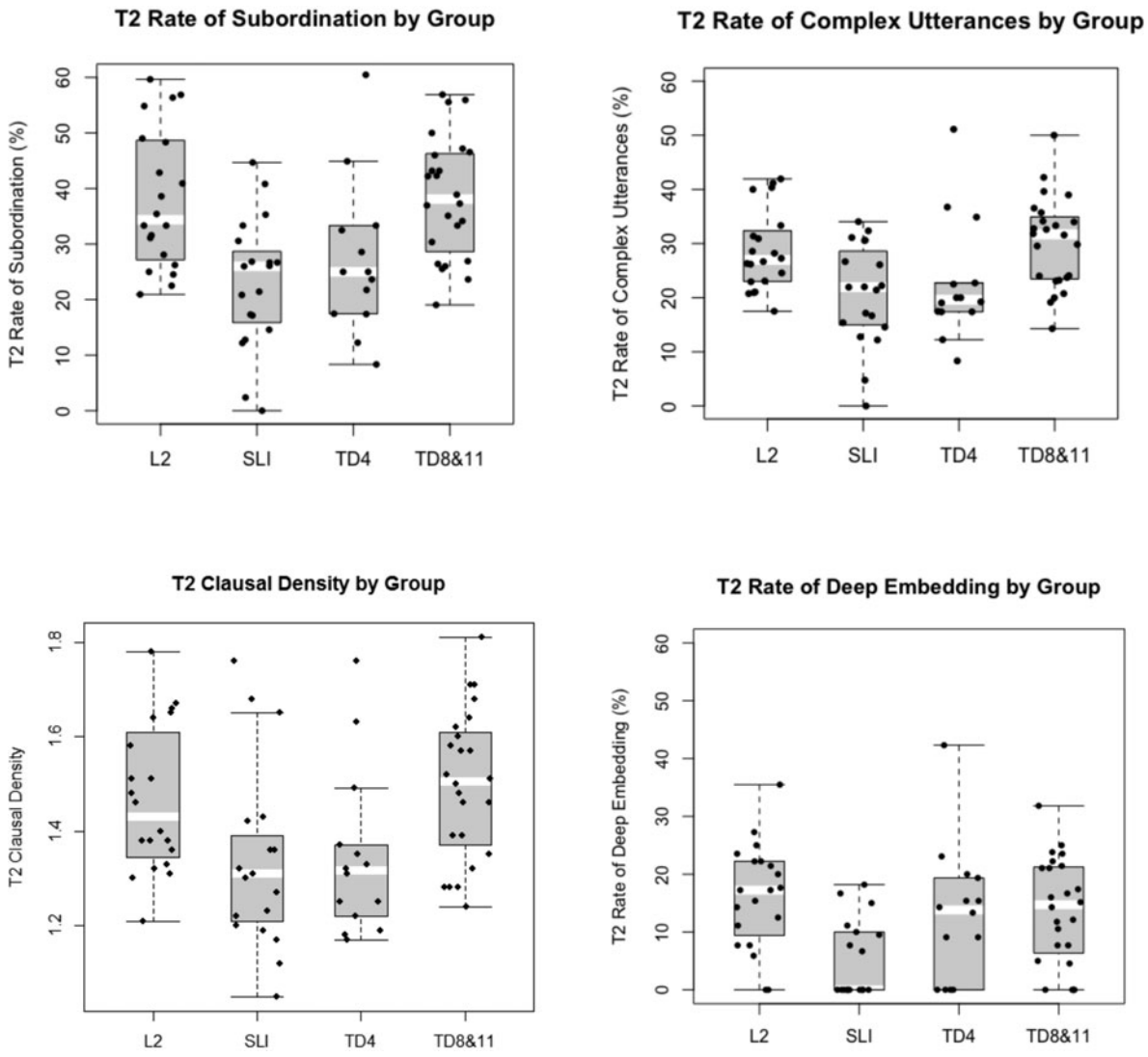
Figure 1. Group and individual clausal embedding results at T1.

understanding the extent to which clausal embedding measures alone are effective at discriminating between the two groups. We therefore report the sensitivity (true positives) and specificity (true negatives) outcomes as well as likelihood ratios (LR). In other words, the sensitivity percentages express how many children with SLI were identified as such by clausal embedding measures and the specificity percentages express how many L2 children were identified as typically developing by the same measures. We ran this analysis on 20 L2 children, excluding SAJ and ALP. According to Dollaghan and Horner (2011), positive likelihood ratios (LR+) ≥ 10 and negative likelihood ratios (LR-) ≤ 0.10 are informative, while LR+ of around 3 and LR- of around .3 are only suggestive and would require further testing. Sensitivity and specificity rates of at least 80% are considered fair (Plante & Vance, 1994). As can be seen in Tables 9–10, the clausal embedding measures generally have poor sensitivity (< 75%), especially at T1, but specificity outcomes were fair for some measures (T1: 75%-80%; T2: 70%-85%). The positive likelihood ratios are near 3 at T1 and T2, but the negative likelihood ratios are fairly uninformative (with perhaps the exception of Rate of Subordination: LR- = 0.35 and Clausal Density: LR- = 0.38 at T2).

Impact of length of exposure and age

We now turn to the role of language exposure (LoE) and age on the two spontaneous language measures (i.e., morphosyntactic accuracy (Error Rate) and clausal embedding measures). We first present the impact of LoE and age on morphosyntactic accuracy in both groups. We

https://doi.org/10.1017/S1366728918000949 Published online by Cambridge University Press



NOTES: Boxplots overlaid with dot charts that represent individual performance (Larson-Hall, 2010, p. 188). ALP and SAJ are excluded from L2 group analyses.

Figure 2. Group and individual clausal embedding results at T2.

then turn to their impact on clausal embedding. Finally, we present correlations between standardized tests and spontaneous language measures.

Tables 11–12 present the correlation results between language experience variables, standardized test scores, and spontaneous language measures in the L2 children. Tables 13–14 display a similar analysis for the children with SLI. In the L2 children, LoE was inversely and significantly correlated with Error Rate at T1, but this was not significant at T2 (see Figure 3). These results suggest that language exposure is linked to morphosyntactic accuracy, especially when children are in earlier stages of L2 development (see Scheidnes & Tuller, 2016 for a similar conclusion about the role of a critical LoE of less than 18 months for standardized test performance). These findings also indicate that L2 children with low LoE are more likely to overlap with the clinical population with respect to morphosyntactic accuracy. Age was not related to Error Rate in the L2 children, whereas this correlation was moderate to strong at both testing times in the SLI.

When examining the link between LoE and measures of clausal embedding in the L2 children, no significant correlations were found (see Tables 11–12 and Figure 4). These findings suggest that most of the L2 children with a LoE of < 18 months performed like the rest of the group. Age in the L2 children was not significantly correlated to clausal embedding measures, but these variables were significantly correlated in the children with SLI at T1 and tended towards significance at T2 (see Tables 13–14). The lack of correlation between age and embedding measures in the L2 children resembles previous findings which have shown gradual increases in such measures in TD

https://doi.org/10.1017/S1366728918000949 Published online by Cambridge University Press

Table 9. *Sensitivity, specificity, likelihood ratios and confidence intervals of the clausal embedding measures at T1.*

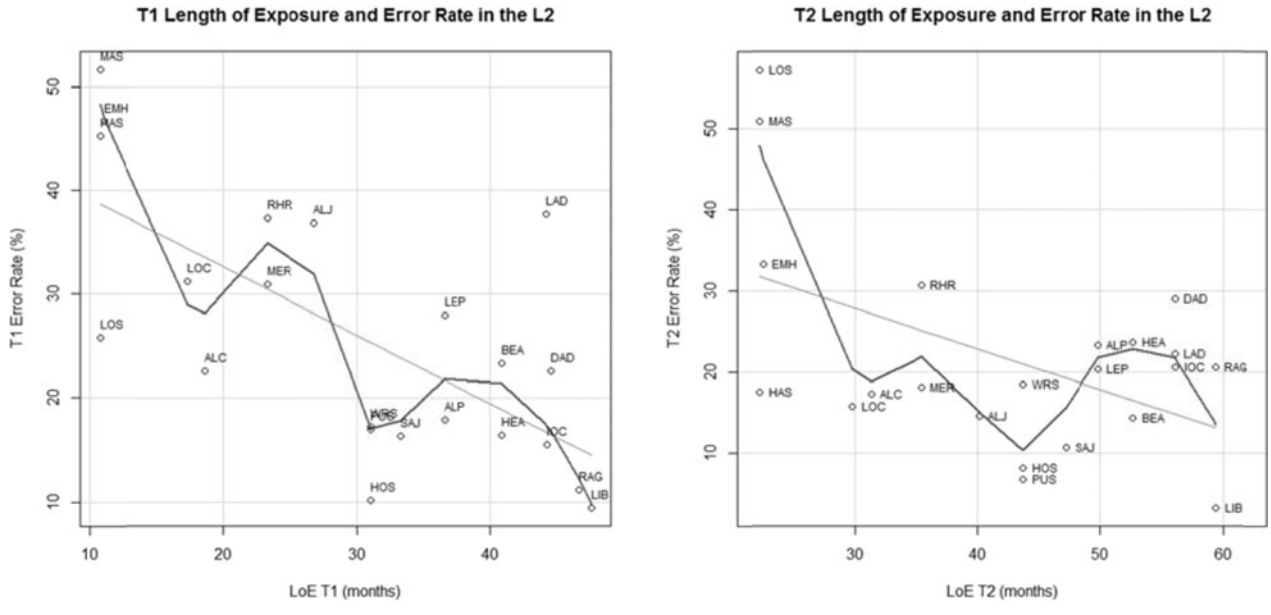
T1	Cut-off	Sensitivity	Specificity	LI+	95% CI	LI-	95% CI
Rate of Subordination (Sub)	28%	47% (9/19)	80% (16/20)	2.37	0.87–6.42	0.66	0.41 -1.06
Rate of Complex Utterances (Cplx)	22%	53% (10/19)	80% (16/20)	2.63	0.99–6.98	0.59	0.35–1.00
Clausal Density (CDens)	1.36	58% (11/19)	75% (15/20)	2.32	0.99–5.42	0.56	0.31–1.01
Rate of Deep Embedding (DeepE)	7%	47% (9/19)	75% (15/20)	1.9	0.77–4.64	0.70	0.43–1.15

Note: LI+ = positive likelihood ratio. LI- = negative likelihood ratio. Sensitivity measures the number of SLI would were identified as SLI based on the cut-off. Specificity measures the number of L2 children who were identified at typically developing based on the cut-off.

Table 10. *Sensitivity, specificity, likelihood ratios and confidence intervals of the clausal embedding measures at T2.*

T2	Cut-off	Sensitivity	Specificity	LI+	95% CI	LI-	95% CI
Rate of Subordination (Sub)	28%	73% (14/19)	75% (15/20)	2.95	1.32–6.59	0.35	0.16–0.78
Rate of Complex Utterances (Cplx)	22%	58% (11/19)	80% (16/20)	2.90	1.11–7.74	0.53	0.30–0.93
Clausal Density (CDens)	1.36	74% (14/19)	70% (14/20)	2.46	1.19–5.05	0.38	0.17–0.84
Rate of Deep Embedding (DeepE)	7%	63% (12/19)	85% (17/20)	4.21	1.40–12.63	0.43	0.23–0.80

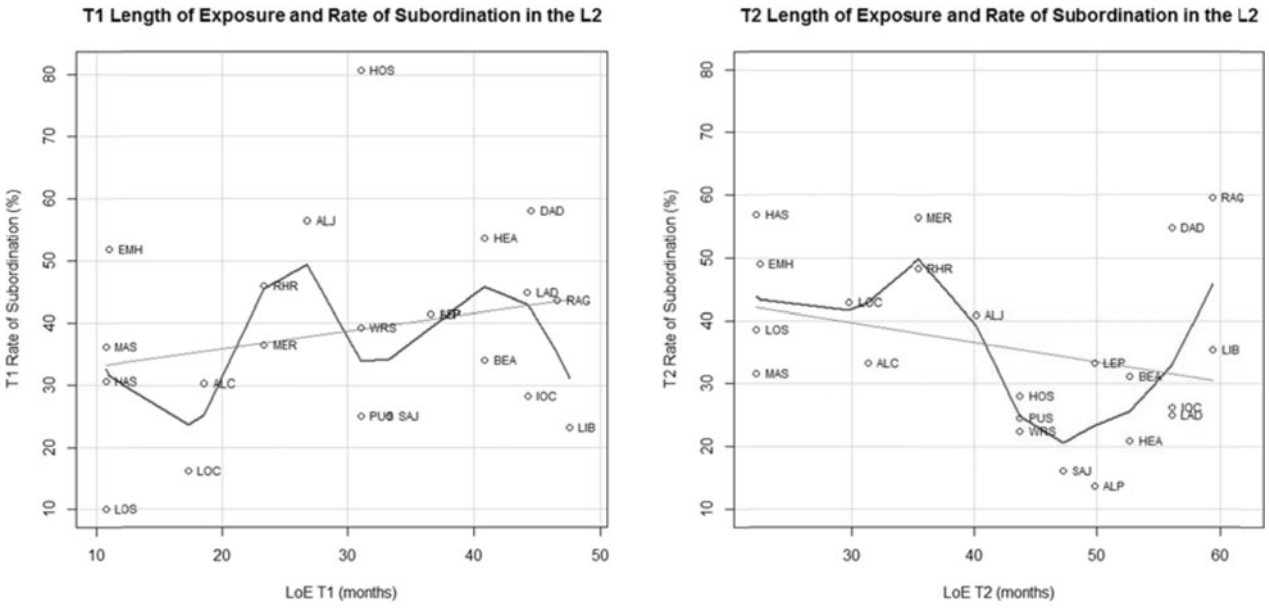
Note: LI+ = positive likelihood ratio. LI- = negative likelihood ratio. Sensitivity measures the number of SLI would were identified as SLI based on the cut-off. Specificity measures the number of L2 children who were identified at typically developing based on the cut-off.



Notes: The black line is a loess line. This figure includes ALP and SAJ for illustrative purposes.

Figure 3: Length of exposure and error rates in the L2 group ($N = 22$).

https://doi.org/10.1017/S1366728918000949 Published online by Cambridge University Press



NOTES: The black line is a loess line. This figure includes ALP and SAJ for illustrative purposes.

Figure 4. Length of exposure and rate of subordination in the L2 group ($N = 22$).

Table 11. *L2 correlation matrix at T1.*

	Sub	Cplx	CDens	DeepE	ER
Age	.13	.21	.03	−.22	−.20
LoE	.22	.18	.16	.27	−.65***
SentCpl (raw)	.39•	.41•	.32	.20	−.66***
SentCpl (z-scores)	.22	.24	.26	.13	−.54*
WRep (raw)	.32	.27	.26	.23	−.25
WRep (z-scores)	.03	.04	.10	.12	−.21
Vocab (raw)	.40•	.41	.36•	.12	−.57*
Vocab (z-scores)	.26	.22	.28	.29	−.53*

Notes: •p < .10; *p < .05; **p <.01; ***p <.001.
LoE = Length of Exposure, SentCpl = Sentence Completion, WRep = Word Repetition, Sub = Rate of Subordination, Cplx = Rate of Complex Utterances, CDens = Clausal Density, DeepE = Rate of Deep Embedding, ER = Error Rate. The L2 group includes 20 children here since ALP and SAJ are excluded from group analyses.

Table 12. *L2 correlation matrix at T2.*

	Sub	Cplx	CDens	DeepE	ER
Age	.17	.20	.17	−.07	−.05
LoE	−.32	−.08	−.37•	−.46*	−.33
SentCpl (raw)	−.18	−.12	−.15	−.22	−.09
SentCpl (z-scores)	−.32	−.24	−.26	−.28	−.004
WRep (raw)	.22	.34	.08	.07	−.24
WRep (z-scores)	.13	.28	−.01	.04	−.24
Vocab (raw)	.08	.11	−.03	.07	−.30
Vocab (z-scores)	.01	.07	−.09	.17	−.24

Notes: •p < .10; *p < .05; **p <.01; ***p <.001.
LoE = Length of Exposure, SentCpl = Sentence Completion, WRep = Word Repetition, Sub = Rate of Subordination, Cplx = Rate of Complex Utterances, CDens = Clausal Density, DeepE = Rate of Deep Embedding, ER = Error Rate. The L2 group includes 20 children here since ALP and SAJ are excluded from group analyses.

monolingual French-speaking children aged 6-, 8- and 11-years old (Tuller et al., 2012).

5.3 Interpreting standardized test scores in the L2 children using clausal embedding measures

Can the spontaneous language measures analyzed here help with the interpretation of the standardized test scores in the L2 children? We first examine any significant correlations found between the standardized tests and spontaneous language measures in the L2 children. We then discuss different language profiles observed in this group.

The only significant correlations between the standardized tests and spontaneous language measures concerned Error Rate and Sentence Completion and Error Rate and Vocab. These correlations only reached significance in the L2 children at T1, when the children with the lowest LoE made the most errors. This suggests that these measures tap into similar language abilities and that they both tend to underestimate language abilities in children with the lowest LoE (in this case, an LoE < 18 months). Interestingly, clausal embedding measures were not significantly linked to standardized test scores, which suggests that clausal embedding and standardized tests measure different language abilities.

https://doi.org/10.1017/S1366728918000949 Published online by Cambridge University Press

Table 13. *SLI correlation matrix at T1.*

	Sub	Cplx	CDens	DeepE	ER
Age	.43•	.52•	.41•	.09	−.45
SentCpl (raw)	.33	.34	.38	.06	−.47
SentCpl (z-scores)	−.11	−.17	−.13	−.12	.04
WRep (raw)	.47*	.46*	.38	.35	−.47
WRep (z-scores)	.04	.01	−.16	.12	.02
Vocab (raw)	.37	.48*	.34	.1	−.18
Vocab (z-scores)	.15	.12	.08	.06	.10

Notes: •p < .10; *p < .05; **p <.01; ***p <.001.
SentCpl = Sentence Completion, WRep = Word Repetition, Sub = Rate of Subordination, Cplx = Rate of Complex Utterances, CDens = Clausal Density, DeepE = Rate of Deep Embedding, ER = Error Rate.

Table 14. *SLI correlation matrix at T2.*

	Sub	Cplx	CDens	DeepE	ER
Age	.33	.35•	.38•	−.13	−.80***
SentCpl (raw)	.22	.16	.26	.09	−.58*
SentCpl (z-scores)	.11	.09	.13	−.03	−.22
WRep (raw)	.06	.03	.09	.03	−.36
WRep (z-scores)	−.01	−.02	.00	−.04	−.18
Vocab (raw)	.04	.01	.14	.05	−.12
Vocab (z-scores)	−.21	−.21	−.11	−.12	.32

Notes: •p < .10; *p < .05; **p <.01; ***p <.001
SentCpl = Sentence Completion, WRep = Word Repetition, Sub = Rate of Subordination, Cplx = Rate of Complex Utterances, CDens = Clausal Density, DeepE = Rate of Deep Embedding, ER = Error Rate.

As reported in Scheidnes and Tuller (2016), at T1, there are seven children (IOC, WRS, LEP, HAS, EMH, MER, SAJ) who scored at or below the −1.65 *SD* cut-off on two out of the three French standardized tests and at or below the −1 *SD* cut-off for the Core Lg. Subtests (CELF-4, English test) (see Tables S3 and S4 for the spontaneous language and standardized test scores). These types of standardized test results are difficult to interpret in bilingual children. In the current study, we found that these L2 children also had high error rates in spontaneous language relative to their age-matched monolingual peers (TD8&11), who made very few in comparison (M = 1.37%, SD = 1.68). However, the clausal embedding rates for all of these children resemble (within 1 *SD*) the TD8&11 rates, with the exception of the child SAJ.

At T2, only three L2 children have z-scores below the cut-off on two out of three tests in French and on the English assessment: HAS, SAJ, ALP (Scheidnes & Tuller, 2016). Out of these children, only HAS used clausal embedding with a similar frequency to the TD8&11. As we discussed in Section 4.1, examination of SAJ and ALP's standardized test scores in both languages and information from parental questionnaires revealed at-risk indicators (family history of language problems for SAJ and repeated and severe otitis media for ALP). The fact that these two children also had low embedding measures suggests that these low rates of embedding, along with the standardized test and questionnaire results, do in fact signal atypical development in L2 children. Interestingly, SAJ does not have a terribly high error rate in spontaneous language. She thus resembles the children with SLI who avoid clausal embedding presumably in order to limit errors (compare Figures 3 and 4). However, with respect to the L2 child HAS, who had low standardized test scores in both languages at both T1 and T2 but frequent embedding and no additional signs of an at-risk profile, his high rates of embedding suggest that he is in fact a typical learner. It appears, therefore, that clausal embedding measures provide important information about age-appropriate language abilities in children who may have been put at a disadvantage by the make-up of the standardized tests and by the comparison to monolingual norms.

Are there cases of low rates of embedding in the children who scored within norms in at least one of the languages? As the specificity rates from Tables 9 and 10 suggest, this did occur in our data, albeit in only a few cases (LOC, LOS, and LIB at T1 and PUS, WRS, and HEA at T2). The children LOC and LOS have rates of clausal embedding that are below −1 *SD* of the TD8&11 mean at T1, but not at T2. Their low rates at T1 could be due to their relatively low LoE; however, other children with similar LoE do not follow this pattern (ALC, EMH, HAS, MAS) at T1 or T2. At T1, LOC and LOS have normal scores on the English assessment and information from the family questionnaire suggests that these children are not at risk for language impairment. Therefore, the low embedding measures for these two L2 children pose a problem for the hypothesis that clausal embedding emerges early in TD L2 children and can be measured in terms of frequency in spontaneous language samples. At T1, the child LIB also has low embedding despite normal standardized test scores in both languages and an LoE of 4 years. At T2, PUS, WRS, and HEA scored within the norm on standardized tests for at least one language (Scheidnes & Tuller, 2016), but their performance on at least three out of four embedding measures fell below −1 *SD* of the TD8&11 mean (see Table S4), suggesting that they did not use clausal embedding in an age-appropriate manner. These profiles are problematic for the hypothesis that clausal embedding is easy for Bi-TD, as measured by frequency of use in spontaneous language samples; however, as the fair specificity measures show, for the vast majority of the L2 children, clausal embedding measures did not underestimate their language abilities. Therefore, this

https://doi.org/10.1017/S1366728918000949 Published online by Cambridge University Press

analysis of individual profiles suggests that the embedding measures are helpful in ruling out SLI in L2 children who have low standardized test scores but who otherwise do not present any other risk of language impairment. However, these embedding measures alone do not appear to be effective markers of SLI in monolingual children, as can be seen in the poor sensitivity rates.

Crucially, the majority of the L2 children were able to perform on a par with their monolingual age peers (TD8&11) with respect to clausal embedding, but they performed nowhere near the TD8&11 for morphosyntactic accuracy at T1 or T2. At T1, only LIB has an error rate under 10% and at T2, only HOS, LIB and PUS are under 10%. Therefore, error rates measured in such a way would risk under-estimating the language abilities of this type of bilingual.

6. Discussion

6.1 Summary

The overall objective of the present paper is to contribute to what is known about developmental trajectories in sequential child bilinguals and to better understand the extent to which they overlap or not with children with SLI on various language measures. More specifically, we wanted to know if the analysis of morphosyntactic accuracy and clausal embedding in spontaneous language would contribute to an accurate assessment of language abilities in sequential bilinguals who vary in length of exposure and age.

A group of sequential bilinguals (L2 children: L1 English, L2 French) without a diagnosis of language impairment was compared to a group of monolingual French-speaking children with SLI on standardized test scores and on spontaneous language measures. The L2 children and children with SLI had similar standardized test scores for French morphosyntax (Sentence Compl) and for morphosyntactic accuracy (Error Rate) in spontaneous language; however, the L2 used clausal embedding more frequently than the SLI. Furthermore, the frequency of clausal embedding was neutral to variation in length of exposure (LoE). Although the diagnostic accuracy results (i.e., low sensitivity) suggest that clausal embedding measures in and of themselves are not good at detecting SLI in monolinguals, the analysis of individual profiles in the L2 children suggest that clausal embedding could be used to rule out SLI in L2 children with low standardized test scores and no additional at-risk factors for language impairment.

6.2 Standardized test performance in the L2 and SLI

Our first research question concerned any potential overlap in standardized test performance in the L2 children and children with SLI. Overall, there was considerable overlap between the two groups for these tasks. However, the L2 children had stronger word repetition (Word Rep) performance than the children with SLI and the L2 children also showed more improvement from T1 to T2 than the children with SLI. The longitudinal results suggest that, with time, the L2 children have performance that overlaps less with that of the children with SLI.

6.3 Clausal embedding and morphosyntactic accuracy

With our second research question, we focused on the extent to which the L2 children and children with SLI would overlap on spontaneous language measures of morphosyntactic accuracy and clausal embedding. In terms of morphosyntactic accuracy (Error Rate), there was considerable overlap between the L2 and SLI. Furthermore, both groups produced more errors than the language-matched controls (4-year-olds: the TD4 group). However, the L2 produced embedded clauses more frequently than the SLI. Furthermore, the L2 children patterned much more closely with the age-matched controls (8- and 11-year-olds: the TD8&11group) for clausal embedding, while the SLI patterned more closely with the TD4 group. Although clausal embedding differences between the L2 children and the children with SLI were starker at T2 compared to T1, overall, the L2 children used clausal embedding in an age-appropriate manner at both T1 and T2, while the children with SLI used clausal embedding less frequently.

Although the L2 children generally produced more embedded clauses than the children with SLI, the low sensitivity rates suggest that clausal embedding measures in and of themselves are not effective at detecting SLI in monolingual children. However, diagnostic accuracy metrics are generally used in comparing monolingual children with SLI to TD monolingual children or bilingual children with SLI to TD bilingual children, so the sensitivity and specificity rates as well as the likelihood ratios reported here should be interpreted with extreme caution. Indeed, additional research on clausal embedding in monolingual and bilingual populations is warranted and could lead to a better understanding of how SLI impacts complex syntactic structures such as these. Specifically, research leading to norms for frequency of clausal embedding in both monolinguals and bilinguals could allow researchers to determine expected clausal embedding rates as a function of age. To our knowledge, such norms do not exist for French, but they would help clinicians by providing a more concrete definition of what constitutes age-appropriate clausal embedding rates.

We therefore conclude that clausal embedding measures are a stronger indicator of a typical child L2 profile compared to overall morphosyntactic accuracy.

https://doi.org/10.1017/S1366728918000949 Published online by Cambridge University Press

These findings reinforce the importance of including some sort of measure of syntactic complexity when identifying SLI in bilingual children. In this way, our results overlap with those of Simon-Cereijido and Gutiérrez-Clellen (2007), who found that a combination of MLU and error rates could distinguish between typical development and language impairment in a group of Spanish-speaking children, some of whom were also proficient in English. However, these authors found that the bilingual children in their study were similar in morphosyntactic accuracy to the children with limited exposure to English, while we found dramatic differences in this area between the L2 and the TD monolingual groups. Differences in age of acquisition, language input and dominance between the L2 children in France and the Spanish-speaking bilingual group in the US most likely contributed to these diverging results. Despite important differences between the two studies, it is interesting to note that both include descriptions of atypically developing bilinguals who used simple syntax and had low error rates. In the current study, SAJ had a lower error rate relative to other L2 children and low levels of clausal embedding. This overlap suggests that measures of syntactic complexity (e.g., MLU or embedding) could potentially be used across bilinguals in vastly different language learning contexts.

While the low sensitivity results suggest that the clausal embedding measures alone are not effective at ruling in SLI in monolinguals, the L2-SLI group differences and the neutrality of clausal embedding to LoE suggest that further investigation into the use of clausal embedding to distinguish between Bi-TD and Bi-SLI is warranted. As is often the case in research on bilingualism and language disorders, aspects of the present study were weakened by problems linked to the heterogeneity in language performance in L2 children and children with SLI. Although spontaneous language samples are recommended for use in clinical settings, this data collection method contributes to the problem of individual variation, seeing as each child is relatively free to choose from different lexical, morphosyntactic, and structural items to complete the task. However, the flexibility of spontaneous language also gives the researcher the opportunity to observe a wider array of language abilities, as was the case for the clausal embedding measures in the present study. Moving forward, these findings can be used to improve upon more efficient clinical tools, such as sentence repetition. As demonstrated by Thordardottir and Brandeker (2013), sentence repetition can discriminate between Bi-TD and Bi-SLI. Furthermore, sentence repetition can be designed to target those language properties that are more neutral to language exposure, thereby reducing biases in the language assessment of bilinguals. If we apply the findings of the current study, it would be interesting to evaluate sentence repetition using simple lexical and morphological items and targeting the depth of embedding. Scoring approaches could be designed to target the repetition of a certain structure type (e.g., relative clauses) independently of other errors (see Marinis & Armon-Lotem, 2015). This may lead to an improvement in the diagnostic metrics reported here (see for example, Tuller, Hamann, Chilla, Ferré, Morin, Prevost, dos Santos, Abed Ibrahim & Zebib, 2018, and sources cited therein).

6.4 Effect of LoE and age

Our third research question focused on the effect of various language experience factors (notably, length of exposure, LoE, and age). One major difficulty in distinguishing between typically and atypically developing bilinguals concerns the impact these factors have on development. However, if clinicians know that a certain structure is less sensitive to LoE, then this structure can be targeted in bilingual language assessment, even in early stages of acquisition. The results showed that the L2 children used embedding with the same frequency as the TD8&11, regardless of low LoE (e.g., EMH, HAS, and MAS, who had a LoE of 0;11 and ALC who had a LoE of 1;7, had rates of subordination that were within 1 *SD* of the TD8&11 mean) and there were no significant differences for rates of embedding from T1 to T2. These findings are similar to those of previous case studies on L2 children that have reported early emergence of embedded clauses (e.g., Gavruseva & Lardiere, 1996). Given the absence of a significant correlation between LoE and clausal embedding and the fact that most L2ers with low LoE used embedding frequently, clausal embedding measures could be a promising candidate for assessing bilinguals, even in early stages of development, after just one year of exposure to the L2.

These results are in line with a recent proposal to consider recursion as a focal point in theoretical and applied work on SLI and bilingualism (Roeper, 2012). In spite of the complexity involved in embedding one clause within another, it seems that L2 children are able to produce embedded clauses as soon as they have acquired enough lexical items. This early emergence and age-appropriate use of clausal embedding in the L2 supports a bilingual language architecture of the type described in MacSwan (2000), in which one computational module underlies both languages. One computational module would explain why the operation involved in embedding one clause within another appears to come for free for the L2 children in their L2.

Compared to clausal embedding, morphosyntactic accuracy is related to LoE in a different way. Error rates were significantly correlated with LoE at T1. Furthermore, the rate of erroneous utterances improved significantly from T1 to T2. Measures based on morphosyntactic accuracy could therefore introduce a testing bias that

https://doi.org/10.1017/S1366728918000949 Published online by Cambridge University Press

would put TD bilinguals, especially those with less exposure to the target language, at a disadvantage and at a greater risk for misdiagnosis of language impairment. The correlations between error rates and some standardized test scores suggest that these measures are evaluating similar language abilities that appear to be sensitive to language experience and not necessarily to typical or atypical development in bilinguals.

With respect to the age factor and embedding, again, the L2 children resembled the monolingual controls, in that significant correlations between age and embedding measures are generally not observed across smaller age ranges, but embedding increases gradually over longer periods of time (Tuller et al., 2012).

6.5 Interpreting standardized test scores in the L2 children using clausal embedding measures

The description of the individual profiles showed that the clausal embedding measures were useful in ruling out SLI in L2 children who had low standardized test scores in both languages and no additional at-risk factors for language impairment. However, these clausal embedding measures alone are not sufficient for clinical diagnostics, as shown by the poor sensitivity rates and generally uninformative negative likelihood ratios. While standardized test scores based on monolingual norms must be interpreted with great caution when used with bilinguals, these test scores nevertheless provided relevant information concerning the level of age-appropriate language in each child. Furthermore, information from family background questionnaires was also vital in the description of at-risk profiles. Indeed, recent work has demonstrated the discriminatory power that this type of tool can have and is especially important in multilingual contexts because they target information about language experience variables and variables related to risk factors for SLI (Gutiérrez-Clellen & Kreiter, 2003; Paradis, Emmerzael & Duncan, 2010; Restrepo, 1998; Tuller, Abboud, Ferré, Fleckstein, Prévost, dos Santos, Scheidnes & Zebib, 2015).

We therefore propose to consider that typically developing bilinguals can be expected to produce age appropriate levels of clausal embedding and that these measures can be used, along with standardized test scores and family questionnaire data, to better distinguish between typically and atypically developing bilinguals. However, morphosyntactic accuracy, as measured by an error rate in spontaneous language, does not appear to discriminate between typically and atypically developing bilinguals, given the considerable overlap between the L2 and SLI for this measure. Furthermore, the L2 were consistently and dramatically less accurate in morphosyntax compared to their monolingual age-matched peers.

7. Conclusion and limitations

In this study, we explored the idea that certain spontaneous language measures (i.e., error rates and clausal embedding) provide a more complete picture of the L2 children's language abilities. Although individual variation was high and sample sizes were small, statistical analyses revealed that the L2 children used clausal embedding more frequently than the children with SLI. Crucially, the L2 children patterned with (approximate) age-matched controls on measures of clausal embedding. The diagnostic accuracy metrics of the clausal embedding measures revealed weak sensitivity; however, such analyses are generally reported for comparisons between monolingual children with SLI and monolingual TD children on the one hand and between bilingual children with SLI and bilingual TD children on the other. Furthermore, the fact that the L2 children used clausal embedding in an age-appropriate way regardless of LoE suggests that focusing on this language property, perhaps in a more controlled task, may allow for better delineation of typical from atypical language development in L2 children.

Finally, we would like to highlight that the clausal embedding results diverged remarkably from those of morphosyntactic accuracy. Indeed, there was considerable overlap between the L2 and the SLI with respect to the error rate and the L2 children did not perform like the age control group for this measure. Error rates were high at T1 and remained high at T2 for almost all L2 children, even for those with LoE > 4;0 and for those with age-appropriate French standardized scores. We therefore conclude that clausal embedding emerges early and is used frequently in TD L2 children. Furthermore, high rates of errors relative to TD monolinguals are indicative of typical development in sequential bilinguals, even after several years of L2 exposure.

Supplementary material

To view supplementary material for this article, please visit https://doi.org/10.1017/S1366728918000949

References

Armon-Lotem, S. (2012). Introduction: Bilingual children with SLI – the nature of the problem. *Bilingualism: Language and Cognition*, *15*, 1–4.

Bedore, L. M., & Peña, E. D. (2008). Assessment of bilingual children for identification of language impairment: Current findings and implications for practice. *International Journal of Bilingual Education and Bilingualism*, *11*, 1–29.

Belletti, A., & Hamann, C. (2004). On the L2/bilingual acquisition of French by two young children with different source languages. In P. Prévost & J. Paradis (eds.), *The*

https://doi.org/10.1017/S1366728918000949 Published online by Cambridge University Press

acquisition of French in different contexts, pp. 147–174. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Chevrie-Muller, C., Simon, A. M., & Decante, P. (1980). *Epreuves pour l'examen de langage*. Paris: Editions du Centre de Psychologie Appliquée.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. L. Erlbaum Associates.

De Agostini, M., Metz-Lutz, M. N., van Hout, A., Chavance, M., Deloche, G., & Pavao-Martins, I. (1998). Oral language evaluation battery of aphasic children (ELOLA): A French standardization (4-12 years). *Revue de Neuropsychologie*, *3*, 319–368.

Del Re, A. C. (2013). *compute.es: compute effect sizes* [R Package]. Retrieved from http://cran.r-project.org/web/packages/compute.es

Deltour, J. J., & Hupkens, D. (1980). *Test de vocabulaire actif et passif (t.v.a.p.)*. Braîne le Château, Belgium: Editions Applications des Techniques Modernes.

Dollaghan, C. A., & Horner, E. A. (2011). Bilingual language assessment: A meta-analysis of diagnostic accuracy. *Journal of Speech, Language & Hearing Research*, *54*, 1077–1088.

Dunn, L. M., Dunn, L. M., Whetton, C., & Burley, J. (1997). *British picture vocabulary scale*. London, UK: GL Assessment.

Field, A. P., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Los Angeles: Sage.

Gavruseva, E., & Lardiere, D. (1996). The emergence of extended phrase structure in child L2 acquisition. In A. Stringfellow, D. Cahana-Amitay, E. Hughes, & A. Zukowski (eds.), *BUCLD 20 Proceedings*, pp. 225–236. Somerville, MA: Cascadilla Press.

Golberg, H., Paradis, J., & Crago, M. (2008). Lexical acquisition over time in minority first language children learning English as a second language. *Applied Psycholinguistics*, *29*, 41–65.

Gutiérrez-Clellen, V. F., & Kreiter, J. (2003). Understanding child bilingual acquisition using parent and teacher reports. *Applied Psycholinguistics*, *24*, 267–288.

Gutiérrez-Clellen, V. F., & Simon-Cereijido, G. (2009). Using language sampling in clinical assessments with bilingual children: Challenges and future directions. *Seminars in Speech and Language*, *30*, 234–245.

Gutiérrez-Clellen, V. F., & Simon-Cereijido, G. (2010). Using nonword repetition tasks for the identification of language impairment in Spanish-English-speaking children: Does the language of assessment matter? *Learning Disabilities Research & Practice*, *25*, 48–58.

Gutiérrez-Clellen, V. F., Simon-Cereijido, G., & Wagner, C. (2008). Bilingual children with language impairment: A comparison with monolinguals and second language learners. *Applied Psycholinguistics*, *29*, 3–19.

Hamann, C., & Tuller, L. (2014). Genuine versus superficial relatives in French: The depth of embedding factor. *Rivista Di Grammatica Generativa: Research in Generative Grammar (RGG)*, *36*, 47sear.

Haznedar, B. (2003). The status of functional categories in child second language acquisition: Evidence from the acquisition of CP. *Second Language Research*, *19*, 1–41.

Kapantzoglou, M., Restrepo, M. A., & Thompson, M. S. (2012). Dynamic assessment of word learning skills: Identifying language impairment in bilingual children. *Language, Speech, and Hearing Services in Schools*, *43*, 81–96.

Khomsi, A. (2001). *Evaluation du langage oral*. Paris: Editions du Centre de Psychologie Appliquée.

Khomsi, A., Khomsi, J., Pasquet, F., & Parbeau-Gueno, A. (2007). *Bilan informatisé de langage oral au cycle III et au collège (BILO3C)*. Paris: Centre de Psychologie Appliquée.

Khomsi, A., & Nanty, I. (2001). *Evaluation des compétences langagières au collège (ECL-C)*. Paris: ECPA.

Kohnert, K. J. (2010). Bilingual children with primary language impairment: Issues, evidence and implications for clinical actions. *Journal of Communication Disorders*, *43*, 456–473.

Kohnert, K., Windsor, J., & Ebert, K. D. (2009). Primary or "specific" language impairment and children learning a second language. *Brain and Language*, *109*, 101–111.

Lakshmanan, U., & Selinker, L. (1994). The status of CP and the tensed complementizer *that* in the developing L2 grammars of English. *Second Language Research*, *10*, 25–48.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using R*. Florence, Kentucky: Taylor & Francis Group. Retrieved from http://cw.routledge.com/textbooks/9780805861853/guide-to-R.asp

Lee, L. L. (1971). *Northwestern syntax screening test (NSST)*. Evanston: Northwestern University Press.

Leonard, L. B. (1998). *Children with specific language impairment*. Cambridge, MA: MIT Press.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Macswan, J. (2000). The architecture of the bilingual language faculty: Evidence from intrasentential code-switching. *Bilingualism; Cambridge*, *3*(*1*), 37–54.

Marinellie, S. A. (2004). Complex syntax used by school-age children with specific language impairment (SLI) in child-adult conversation. *Journal of Communication Disorders*, *37*, 517–533.

Marinis, T., & Armon-Lotem, S. (2015). Sentence repetition. In S. Armon-Lotem, J. de Jong, & N. Meir (eds.), *Assessing multilingual children: Disentangling bilingualism from language impairment*. Bristol, UK: Multilingual Matters.

Marinis, T., Armon-Lotem, S., & Pontikas, G. (2017). Language impairment in bilingual children. *Linguistic Approaches to Bilingualism, 7*, 265–276.

Owen, A. J., & Leonard, L. B. (2006). The production of finite and nonfinite complement clauses by children with specific language impairment and their typically developing peers. *Journal of Speech, Language, and Hearing Research, 49*, 548–571.

Paradis, J. (2004). The relevance of specific language impairment in understanding the role of transfer in second language acquisition. *Applied Psycholinguistics*, *25*, 67–82.

Paradis, J. (2011). Individual differences in child English second language acquisition: Comparing child-internal and child-external factors. *Linguistic Approaches to Bilingualism*, *1*, 213–237.

https://doi.org/10.1017/S1366728918000949 Published online by Cambridge University Press

Paradis, J., & Crago, M. (2000). Tense and temporality: A comparison between children learning a second language and children with SLI. *Journal of Speech, Language, and Hearing Research: 43*, 834–847.

Paradis, J., & Crago, M. (2004). Comparing L2 and SLI grammars in child French: Focus on DP. In P. Prévost & J. Paradis (eds.), *The acquisition of French in different contexts: Focus on functional categories*, pp. 89–107. Amsterdam / Philadelphia: John Benjamins.

Paradis, J., Emmerzael, K., & Duncan, T. S. (2010). Assessment of English language learners: Using parent report on first language development. *Journal of Communication Disorders*, *43*, 474–497.

Paradis, J., Genesee, F., & Crago, M. (2011). *Dual language development and disorders: A handbook on bilingualism and second language learning* (2nd ed.). Baltimore, MD: Brookes Publishing.

Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-base approach. *Language, Speech, and Hearing Services in the Schools*, *25*, 15–24.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Raven's coloured matrices*. London: H. K. Lewis.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Reed, V. A. (2005). Adolescents with language impairment. In V. A. Reed (eds.), *An introduction to children with language disorders*, pp. 168–219. Boston, MA: Pearson Press.

Reilly, J., Losh, M., Bellugi, U., & Wulfeck, B. (2004). "Frog, where are you?" Narratives in children with specific language impairment, early focal brain injury, and Williams syndrome. *Brain and Language*, *88*, 229–247.

Restrepo, M. A. (1998). Identifiers of predominantly Spanish-speaking children with language impairment. *Journal of Speech, Language, and Hearing Research*, *41*, 1398–1411.

Roeper, T. (2012). Minimalism and bilingualism: How and why bilingualism could benefit children with SLI. *Bilingualism: Language and Cognition*, *15*, 88–101.

Scheidnes, M. (2015). Gender marking strategies in L2 children and monolingual children with SLI. In C. Hamann & E. Ruigendijk (Eds.), *Language acquisition and development: Proceedings of Gala 2013* (pp. 416–428). Newcastle upon Tyne: Cambridge Scholars Publishing.

Scheidnes, M., & Tuller, L. (2016). Assessing successive bilinguals in two languages: A longitudinal look at English-speaking children in France. *Journal of Communication Disorders*, *64*, 45–61.

Scott, C. (2003). Language as a variety: An analysis of clausal connectivity in spoken and written language of children with language learning disabilities. Oral communication presented at the *SRCLD*. Madison, WI.

Semel, E., Wiig, E., & Secord, W. (2006). *Clinical Evaluation of Language Fundamentals (CELF4- UK)* (Fourth UK). London: Harcourt Assessment.

Simon-Cereijido, G., & Gutiérrez-Clellen, V. F. (2007). Spontaneous language markers of Spanish language impairment. *Applied Psycholinguistics*, *28*, 317–339.

Stevens, G. (2006). The age-length-onset problem in research on second language acquisition among immigrants. *Language Learning, 56*, 671–692.

Thordardottir, E. (2015). Proposed diagnostic procedures for use in bilingual and cross-linguistic contexts. In S. Armon-Lotem, J. de Jong, & N. Meir (eds.), *Assessing multilingual children: Disentangling bilingualism from language impairment*, pp. 329–358. Bristol, UK: Multilingual Matters

Thordardottir, E., & Brandeker, M. (2013). The effect of bilingual exposure versus language impairment on nonword repetition and sentence imitation scores. *Journal of Communication Disorders*, *46*, 1–16.

Thordardottir, E., & Juliusdottir, A. G. (2013). Icelandic as a second language: A longitudinal study of language knowledge and processing by school-age children. *International Journal of Bilingual Education and Bilingualism*, *16*, 411–435.

Tuller, L., Hamann, C., Chilla, S. Ferré, S., Morin, E., Prevost, P., dos Santos, C., Abed Ibrahim, L. & Zebib, R. (2018). Identifying language impairment in bilingual children in France and in Germany, *International Journal of Language & Communication Disorders*, *53*, 888–904.

Tuller, L., Abboud, L., Ferré, S., Fleckstein, A., Prévost, P., dos Santos, C., Scheidnes, M., & Zebib, R. (2015). Specific language impairment and bilingualism: Assembling the pieces. In C. Hamann & E. Ruigendijk (eds.), *Proceedings of GALA 2013*.

Tuller, L., Henry, C., Sizaret, E., & Barthez, M.-A. (2012). Specific language impairment at adolescence: Avoiding complexity. *Applied Psycholinguistics*, *33*, 161–184.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego, CA: Elsevier Science.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). Burlington, MA: Elsevier Inc.

Wilcox, R. R. (2010). *Fundamentals of modern statistical methods: Substantially improving power and accuracy* (2nd ed.). New York: Springer.

Wilcox, R. R., & Schönbrodt, F. D. (2014). *The WRS package for robust statistics in R* (version 0.24). Retrieved from http://r-forge.r-project.org/projects/wrs/

https://doi.org/10.1017/S1366728918000949 Published online by Cambridge University Press