Simone E. Pfenninger*

About the INTER and the INTRA in age-related research: Evidence from a longitudinal CLIL study with dense time serial measurements

<https://doi.org/10.1515/lingvan-2020-0028>

Abstract: This is the first longitudinal study to explore the best time and timing for regular versus bilingual language exposure in (pre)primary programs, using multiple measures over time so as to focus on fluctuations, trends and interactions in individual data as well as intra-individual variation over time. We studied children who had received 50/50 bilingual instruction in German and English (so-called ‘partial CLIL’ programs) as well as children in ‘minimal CLIL’ programs with almost uniquely monolingual German instruction (90% German, 10% English). Results show that, like other individual differences (ID) variables, the age factor behaves like a dynamic entity that changes over time and affects L2 literacy development differentially at different times. Furthermore, while an early age of first bilingual language exposure has no effect on the L2 development for the children in the minimal CLIL program, early-AO bilinguals in the partial CLIL program (age of first exposure 5) outperform the older-AO bilingual group (age of first exposure 7 and 9) in terms of accuracy and (syntactic and morphological) complexity but not in terms of lexical richness and fluency.

Keywords: SLA, age factor, CLIL, complex dynamic systems, young learners, individual differences

1 Introduction

The prevailing approach to SLA up to the beginning of this century was to focus on product-based explanations of SLA (Lowie and Verspoor 2015), which is reflected, among other things, in the assumption that the variation in interlanguage is either rather systematic or completely random and can be relegated to “(white) noise”. In particular, it was non-systematic, free intra-individual variation that had often been too readily dismissed as noise or measurement error or attributed to “outliers”. Recent years, however, have seen attempts to view this type of individual variation as a prerequisite of development (see Verspoor et al. this special issue).

In order to analyze non-systematic, free intra-individuals as a source of information in my work on the age factor in SLA, we studied Swiss children who had received 50/50 bilingual instruction in German and English (so-called ‘partial Content and Language Integrated Language CLIL’ programs) as well as children in ‘minimal CLIL’ programs with almost uniquely monolingual German instruction (programs with largely – 90% – German instruction vs. 10% English). The main idea of CLIL is that proficiency will be developed in both the non-language subject and the language in which it is taught. In the cross-sectional part of the study at the end of primary education (age 12), we assessed the German and English writing skills of 251 students who varied in their age of first CLIL instruction onset (5, 7, 9 or 11), 231 of whom were from German-speaking homes (new to English), while 20 were from English-speaking homes (new to German). For 91 of them data collection occurred four times annually over eight school years (ages 5–12), *via* narrative and argumentative essays.

The aim of the paper is twofold. Firstly, to analyze the optimal age of CLIL instruction onset and then to explore the variability of the age factor. The fact that very few variables remain the same over time and the inter-individual structure cannot be generalized to the level of variation within each subject (see Bülow and Vergeiner this special issue) led to the longitudinal component of this study with multiple measures over time,

*Corresponding author: Simone E. Pfenninger, University of Salzburg, Salzburg, Austria, E-mail: simone.pfenninger@sbg.ac.at.
<https://orcid.org/0000-0002-0433-4812>

focusing on fluctuations in individual data and intra-individual variation. Thus, in this research design, variation lies in the time-series (L2 fluctuation observed in dense time serial measurements). For the data analysis, linear and logit mixed models and generalized additive mixed-effects regression models (GAMM) were used, with a view to tapping age effects on L2 progress *within* programs (rather than *between* programs). Data collection started in 2008 and occurred four times annually until 2016.

Micro-development studies such as this one, which require dense data collection intervals to focus on studying change as it occurs in the data that cover the entire period during which development is studied, are extremely rare in SLA in general (Lowie 2017; Van Dijk et al. 2011) and non-existent in age-related research. According to Larsen-Freeman (2009), “language performance and development are complex, nonlinear, dynamic, socially situated processes” (p. 588), and when we add writing or literacy to the mix, the complexity of any related research endeavor is compounded even further.

From a practical perspective, it is important for educators and parents to know about the optimal age of CLIL instruction onset so as to be in a position of choice – whether or not and *when* to enroll children in CLIL programs. Finally, the optimal starting age question is of interest in immersive contexts, since intensive CLIL programs constitute a hybrid between naturalistic L2 acquisition (where we find a clear ‘earlier = better’ trend) and instructional foreign language (FL) contexts (where usually no age effects in favor of early starters are found), as discussed in the following.

2 Theoretical background

SLA has had a long-standing research tradition regarding intra-individual variation in interlanguage amongst L2 learners (see e.g., Hyltenstam 1977; Sorace 2005; Tarone 1983) and has always taken a “vibrant interest in the controversial notion of free or random variation” (Ortega 2014: 195). According to Ellis (1985), intra-individual variation – or “interlanguage variability” as he calls it – can be categorized into systematic (situational, contextual or psycholinguistic) variation versus non-systematic (performance or free) variation. Non-systematic, free variation – sometimes referred to as “random variability” – relates to the existence of two or more forms in the learner’s mind, which they use to realize the same range of meanings. It is this type of intra-individual variation that has been suggested to be the key to understanding the vertical dimension of interlanguage, i.e. language developments, particularly in the realm of complexity theories (e.g., Larsen-Freeman and Cameron 2008; Lowie and Verspoor 2015, 2019; Lowie et al. 2014; van Dijk et al. 2011 van Geert and Verspoor 2015).

Complex Dynamic System Theory (CDST) is a (meta)theory of change as well as a relational theory with numerous principal characteristics, among which the following are most relevant to this study (see e.g., Larsen-Freeman 2015):

- (1) Learning as a (non-linear) process rather than a product (there is no stasis, only change);
- (2) The meaningfulness of intra-learner variation: increased variability coincides with a developmental jump i.e., a large amount of variability signals that the learner is apparently trying things out and that the subsystem under consideration is unstable.

Not only does this mean that studying performance at one point in time may provide an inaccurate or at least an incomplete picture of language development; recent trends tend to see individual differences (ID) variables as dynamic entities that change over time and may affect development differentially at different times (Dörnyei et al. 2015). Of course, not all factors are equally variable, and the variability may depend on the time scale. Motivated behavior, for instance, differs across both shorter and longer timescales, from seconds (MacIntyre and Serroul 2015) to the lifespan (Kormos and Csizér 2008). The question arises, then, if findings obtained in age-related research are not representative for a longer period of time either. What we know is that while the age factor may matter in L2 acquisition, it is a relatively weak predictor of L2 outcome in formal, instructional contexts. Numerous age-related classroom studies have demonstrated that a number of variables are much stronger than starting age for a range of L2 proficiency dimensions, notably intensity of instruction – albeit so

far only in snapshot studies, cross-sectional studies or longitudinal studies with 1–3 measurements (e.g., Jaekel et al. 2017; Muñoz 2006; Pfenninger and Singleton 2019). The current study expands this research scope, as indicated, by investigating the micro-development of age effects in different contexts of early L2 learning in school.

It should be mentioned that a number of CDST-inspired longitudinal (case) studies have examined the development of complexity in L2 writing, focusing on individual patterns of change over time, studying, *inter alia*, possible interactions between different dimensions of L2 (complexity) development (e.g., Spoelman and Verspoor 2010). CDST supporters suggest that “the learner may initially use many different forms rather randomly but become increasingly sensitive to using the most conventional forms in a certain context. Initially this more balanced use may be disturbed by conditions like stress, but in the course of time the stability of the language system is likely to increase” (Lowie and Verspoor 2015: 76). Related to this, more variability seems to indicate more L2 growth. However, in CDST-inspired publications, almost nothing is said about the critical importance of instruction, particularly for the development of writing abilities (but see Verspoor and Smiskova 2012). From a dynamic usage-based perspective, however, it is assumed that frequency of input (Ellis 2002) is a crucial factor in language development.

3 Aim and scope

This study approaches the age factor from two main perspectives: (1) investigating the relative contribution of individual factors that affect language learning (inter-individual variation); (2) seeking to gain insight into the actual developmental process (intra-individual variation).

The following research questions are addressed:

- (1) Can age of first CLIL exposure predict L2 writing development in children who are educated in bilingual schools?
- (2) How does AO predict variability in change over time? E.g., is the progress of individuals of certain characteristics (e.g., early starters) more uniform overall?

The ultimate aim of the study is thus not only to shed more light on the impact of AO on L2 development in writing but also to show how we can capture this process. As Byrnes (2009) points out, L2 writing research benefits from some of the insights contributed by dynamic systems theory, as it takes an emergentist position, which means it is interested in instabilities, variation, and discovering larger developmental trajectories. We expect to find different developmental patterns for different learners, in line with previous CDST-related studies (e.g., Larsen-Freeman 2006). We also do not expect the group results to be representative for the individual learners, considering the findings of several previous CDST studies on L2 development (Lowie and Verspoor 2019).

We use the notion of ‘age of first CLIL exposure’ as referring to the age when a child first begins receiving intensive, systematic, and maintained exposure to their new language. It is important to bear in mind that this study does not focus on *between*-group comparisons across CLIL programs (comparing, e.g., partial CLIL with minimal CLIL) but on *within*-group and *between*-learner analyses of age effects.

Finally, due to the limits of this short article, this study is not about certain features of the language system (e.g., complexity in L2 writing) or the role played by micro social variables (such as interaction with other learners), but about the impact of a specific macro contextual variable (AO) on learning and performance in L2 writing. As such, the study presented here should be considered in conjunction with the analyses of the same learners and dataset presented in Pfenninger (2020), in which the goal is to elucidate what *causes* significant L2 growth, and how L2 writing – and oral language – development are mediated by a complex, dynamic constellation of individual and social factors.

4 Participants and procedure

In the cross-sectional part, 251 students who varied in their age of first CLIL instruction onset (5, 7, 9, or 11) were recruited at the end of primary education (age 12). 146 of them were in partial CLIL (PAC) classes (see Table 1 below), while 105 of them came from six minimal CLIL (MIC) classes in Switzerland, where students started learning English at different ages: 54 of them were early starters (AO 7 henceforth earlyMIC), while 51 were late starters (AO 11; henceforth lateMIC). The MIC participants' mean age at testing was 12.6 (range 11–14). In these samples, we observed children only from German-speaking homes.

Four groups of learners in the PAC program ($n = 91$) formed the 'focal group', i.e., the basis of the longitudinal part of this study: a group of 25 Swiss learners from monolingual German-speaking homes with age of first CLIL exposure 5 (length of German/English PAC until the end of primary: eight years), a group of 24 Swiss learners with starting age 7 (length of PAC: 6 years), a group of 22 Swiss learners with starting age 9, and a group of 20 children from monolingual English homes with starting age 5 (length of PAC: eight years). Table 1 displays information about these subjects.

Children received between 28 lessons (Grades 1–3) and 30 lessons (Grades 4–6) per week, and they spent between 6.5 and 11 h per day at school. In addition, they received traditional English/German-as-a-second language instruction: 6 h each in Grade 1, 5 h each in Grades 2 and 3, and 4 h each in Grades 4–6 (for a more detailed description, see Pfenninger 2020).

The English-speaking children (Group 4 in Table 1) served as a benchmark against which the performance of the bilingual preschool children could be compared. Following Festman (2018), I refer to them as 'international students' (i.e., earlyPAC-int) – children born outside of Switzerland who had grown up with another language abroad (English) which was still spoken in their international families in the new context (i.e., living and working in Switzerland). These children were in the same classes as the children in Groups 1–3. For them, German was both their L2 and the language spoken locally and their parents were native speakers of English. This enables us to assess the extent to which English home exposure impacts on literacy development in English and German and the extent to which their patterns of development are confluent with those of the other three groups.

Finally, all the children were matched for socio-economic status (and similar home literacy environments), considering that home literacy is a significant factor in early literacy development (see Kovelman et al. 2008).

5 Tasks and procedure

While the 251 students in the cross-sectional design were asked to write an English narrative essay (topic: the plot of their favorite movie, book or TV series) at the end of primary school, the 91 students in the longitudinal design each wrote one English essay per term (i.e., four times a year). Only formally written texts were included (i.e., official assignments rather than journals or personal reflection, see Penris and Verspoor 2017). The participants' texts were matched for genre (argumentative vs. narrative essays) and topic. Note that two dimensions of writing are at play here (Manchón 2011): learning to write in another language and using writing

Table 1: Focal subjects participating in the longitudinal part.

Group	Number of subjects (sex)	L1 (home language)	L2	Age of first bilingual exposure	Age of onset of French learning (2 h/week)	No. of trajectories	Total no. of measurements (four times/year)
1 earlyPAC	25 (15 F)	German	English	5	11	32	800
2 midPAC	24 (12 F)	German	English	7	11	24	576
3 latePAC	22 (14 F)	German	English	9	11	16	352
4 earlyPAC-int	20 (11 F)	English	German	5	11	32	640

to learn the target language. Finally, I am aware of the fact that by looking at a limited number of measures of, e.g., complexity only, progress on other components or along other dimensions may be lost, leading to overgeneralizations and possibly inaccurate conclusions about the course of L2 complexity development (see e.g., Bulté and Housen 2020).

6 Data analysis

Children's essays were transcribed using the CLAN program and CHILDES (McWhinney 2000), as well as additional standard guidelines for transcribing bilingual children's speech (see e.g., Petitto and Kovelman 2003). Once the transcripts were completed, coders with expertise in linguistics, who were also bilingual German-English speakers, coded the children's speech, using the koRpus package in R (version 0.11-5). Transcripts were coded for the following aspects:

- Morphosyntactic complexity and fluency:
 - Mean length of utterance (MLU, i.e., number of morphemes per word)
 - Text length in tokens i.e. word count (W)
 - Clauses per T-Unit (C/T)
- Lexical diversity:
 - Measure of Textual Lexical Diversity (MTLD) (McCarthy and Jarvis 2010), which is independent of text length;
- Accuracy: total number of utterances produced correctly, total number of utterances that contained morphosyntactic and/or semantic errors (token errors and type errors).

For each text (topic), the score of the group was averaged, and these averages were expected to improve over time. Average text length was 711 words and varied between 12 and 1588 words per text, with a gradual increase toward the later samples. After correcting for the increasing trend, none of the topics deviated from the expected score, and there was no reason to delete any of the topics from the data set.

GAMM analyses were performed using the mgcv R package (Wood 2006) and results were plotted using the packages ggplot2 and itsadug (van Rij et al. 2016). I fitted separate smooths to the trajectories of the four groups, and I used model comparison and difference smooths to see whether they were different. Cross-sectional analyses at the end of primary school were conducted using mixed-effects models with hierarchical random effects for 'group', and crossed random effects for classes, subjects and items respectively, using the lme4 package (version 1.1–21) in R (version 3.6.0; R Development Core Team 2019). The final models included fixed effects for age of onset (AO); random effects (intercepts) were added to account for group-to-group differences that induce correlation among scores for students within a CLIL program.

7 Results

As a first step, mixed models were specified in order to establish whether there were any AO-related differences (a) between the two AO groups in the two MIC programs, namely midMIC (AO = 7) and lateMIC (AO = 9) and (b) between the three AO groups in the PAC program, namely earlyPAC (AO = 5), midPAC (AO = 7) and latePAC (AO = 9) (the descriptive statistics, i.e. means, standard deviations and confidence intervals, are shown in Table A in the Supporting Information S1). Tables 2 and 3 show that there were no such differences for the MIC groups, while some of the written measures (text length, clauses per T-unit and accuracy) reached significance in favor of an earlier AO in the PAC program.

It has to be noted, however, that while there were no significant AO-related differences in the PAC program in terms of certain measures (MLU and lexical richness), earlyPAC and midPAC always fell outside the CI of the bilingual participants (earlyPAC-int) and the latePAC; the latter lagged behind significantly (see Table A in the Supporting Information S1 as well as Figures 1 and 2).

Table 2: Linear mixed-effects regression models for the investigated dependent variables at the end of primary school in the **MIC program** (fixed effect estimates for AO).

	Fixed effect: AO		
	<i>Estimate ± SE</i>	<i>t</i>	Main effect <i>p</i>
MLU	−0.19 ± 0.52	−0.36	0.708
Fluency (text length)	−2.45 ± 2.51	−0.98	0.172
Clauses per T-Unit	−0.11 ± 0.17	−0.67	0.443
Lexical richness (MTLD)	−11.19 ± 8.37	−1.34	0.285
Accuracy	−0.08 ± 0.45	−0.17	0.865

Table 3: Linear mixed-effects regression models for the investigated dependent variables at the end of primary school in the **PAC program** (fixed effect estimates for AO).

	Fixed effect: AO		
	<i>Estimate ± SE</i>	<i>t</i>	Main effect <i>p</i>
MLU	−1.71 ± 1.45	−1.19	0.105
Fluency (text length)	−15.00 ± 4.41	−3.46	<0.0001**
Clauses per T-Unit	−0.80 ± 0.47	−1.71	0.043*
Lexical richness (MTLD)	−7.58 ± 6.38	−1.58	0.105
Accuracy	−3.31 ± 0.27	−12.44	<0.0001**

Asterisks indicate significance of * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

There are two problems with such a snapshot analysis: first, it turned out that the findings are not representative for a longer period of time, as measurements *before* the focal point (i.e., at the end of primary school) showed a different picture for all the measures: for fluency, for example, the observed age effect at the end of primary school (see above) was absent two months earlier ($\beta = -83.50 \pm 88.88$, $t = -0.94$, $p = 0.227$) but appeared four months earlier ($\beta = -16.33 \pm 3.61$, $t = -2.78$, $p < 0.001$). Second, such an analysis does not say anything about the nature of the children's L2 development or the L2 growth of the various groups. In order to gain insight into the actual developmental process, GAMMs were specified (see codes in the Supporting Information) for each of the written measures (see e.g., Tables 4 and 5 for fluency and accuracy respectively).

For accuracy (see Table 5), complexity and MLU, both midPAC and latePAC differed significantly from the earlyPAC in terms of height and shape of their L2 trajectories (see Supporting Information). A different picture emerged for fluency (see Table 4) and lexical richness, where the earlyPAC and midPAC showed no significant

Figure 1: Written fluency across AO groups and contexts at the end of primary school.

Figure 2: Written lexical richness across AO groups and contexts at the end of primary school.

difference in their L2 development (for a global measure of oral and written performance using summated z-scores, see Pfenninger in press).

In order to take *all* the measurements into account (not only the last 24 and 16 measurements respectively) the smoothing technique allows us to visualize the general developmental pattern (Figure 3).

The fact that the difference smooths for ‘subject’ were significant in the GAMM analyses suggests that the trajectories of the individuals were indeed different i.e., that there was significant intra-individual variation. In Figure 4, the individual growth curves for the development of the four PAC groups are plotted for lexical richness according to the original longitudinal design; each line represents the development of one student.

Figure 4 illustrates that it is not until we trace learners individually over time that we see clear differences, not only in L2 used at the end, but also in the acquisitional process leading to such use. It is noteworthy at this point that none of the 91 participants showed the mean pattern of the group. What is more, the learners did not initially show more variability. To the contrary, variability increased with time, and so did the likelihood of significant L2 growth. As Figure 5 below shows, all 91 learners made significant improvement in the last 2.5 years of primary school, starting from age 10. Hence, it is not surprising that biological age as well as its interaction with time was significant in all analyses; i.e., there was a significant non-linear interaction between these factors, plotted in Figure 5 for lexical richness.

Time in measurements (x-axis) is plotted against age in years (y-axis), while the third dimension (i.e., the contour lines) represents lexical richness (MTLD) in scores. The contour lines connect points (i.e., combinations between time and age) of similar scores. Confidence intervals for the contour lines are shown in green (lower bound) and red (upper bound). We can see a steeper slope and sharper peak for older children, irrespective of their AO (see also Figure 4 above).

Table 4: Summary of the results of the GAMM model with respect to written fluency.

<i>Parametric coefficients:</i>	Estimate	Std.error	t-value	Pr(> t)
(Intercept)	339.77	18.86	35.09	<0.001***
<i>Smooth terms:</i>	Edf	Ref.df	F	p-value
s(Time)	5.13	6.18	79.07	<0.001***
s(Time):lsMidPAC	2.00	2.00	1.98	0.054
s(Time):lsLatePAC	4.23	4.93	16.68	<0.001***
s(Time, Subject)	12.49	16.84	4.49	<0.001***

Asterisks indicate significance of * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5: Summary of the results of the GAMM model with respect to written accuracy.

<i>Parametric coefficients:</i>	Estimate	Std.error	t-value	Pr(> t)
(Intercept)	19.88	0.51	39.27	<0.001***
<i>Smooth terms:</i>	Edf	Ref.df	F	p-value
s(Time)	8.40	8.88	23.84	<0.001***
s(Time):lsMidPAC	6.19	7.51	29.13	<0.001***
s(Time):lsLatePAC	5.03	5.95	19.45	<0.001***
s(Time, Subject)	28.00	28.00	20.94	<0.001***

Asterisks indicate significance of * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

As a next step, visual methods for significance testing allow us to see where and in what way the trajectories differ, i.e., we can check for overlap or lack of overlap at different points. Figures 6–10 specify the levels of ‘group’ that the difference smooth is based on with corresponding pointwise confidence intervals (minus random effects). When the shaded confidence band does not overlap with the x-axis (i.e., the value is significantly different from zero), this is indicated by a red line on the x-axis (and vertical dotted lines) – in other words, this is where we find significant differences between the two groups.

Figures 9 and 10 reveal that the trajectories between the two groups who started CLIL earliest (at five), namely earlyPAC and earlyPAC-int, differ significantly almost throughout their entire development for

Figure 3: AO effects across five L2 measures.

Figure 4: Individual growth curves for the development of the four PAC groups for lexical richness.

complexity (CT) and accuracy. The opposite picture emerges for mean length of utterance, lexical richness, and fluency: there are only differences in the last two ears (in favor of the international students). EarlyPAC and midPAC almost always differ at the beginning of their L2 development, when the midPAC lag behind, as well as at the end, when the midPAC not only catch up but outperform the earlyPAC. By contrast, there are significant differences between the late starters (latePAC) and the other groups almost all the way through for all measures.

Figure 5: Visualization of the GAMM model for the effects of time and age.

Figure 6: Difference smooths for MLU.

Figure 7: Difference smooths for fluency.

Figure 8: Difference smooths for lexical richness.

Figure 9: Difference smooths for complexity.

Figure 10: Difference smooths for accuracy.

8 Discussion and conclusion

The goal of the current study, was, on the one hand, to investigate the weight of the age factor as it affects L2 attainment at the end of bilingual pre/primary schools. The findings showed that age effects are sensitive to, and thus mediated by, learning contexts. Depending on the intensity of the CLIL program (minimal CLIL vs. partial CLIL), an earlier AO might lead to better outcomes (in the PAC program), i.e., PACs did not show comparable effect structure as children in regular FL programs across the board. Such findings have important implications for multilingual education when decisions are made about (1) early teaching *of* different languages and (2) early instruction *through* different languages in primary school. However, while mixed-model designs can neutralize individual variation within the group by including it as a random effect in the analysis to allow for adequate generalizations in the frequency domain, such generalizations are not warranted in the time domain (for a discussion of this, see Lowie and Verspoor 2015: 69). My second goal was thus to perform an analysis that meets the criteria for an appropriately fine-grained investigation of AO effects – a novel undertaking in age-related research.

Results of the GAMM analysis showed that early-AO bilinguals in the *partial* CLIL program (age of first exposure 5 and 7) outperformed the later-AO bilingual group (age of first exposure 9) in terms of written lexical richness (measure of textual lexical diversity) and fluency (text length in tokens and types) – but not in terms of morphosyntactic complexity and fluency (mean word length, mean length of utterance, verb phrases per T-Unit, and clauses per T-Unit). Importantly, however, there was no significant difference between children who began the CLIL program at the age of five and children with AO 7; the children with AO 9 stood out from the rest with respect to overall height as well as shapes of trajectories: their L2 development looked markedly different across the board and they were not able to catch up with the other groups by the end of primary school.

The finding that a slightly later age of CLIL onset can be as effective as very early CLIL gives educators and parents choices with respect to when they begin L2 instruction using bilingual education without compromising outcomes. What is more, a later start in CLIL can also optimize resources (e.g., a middle or late CLIL

implementation is more cost-effective than an early one). The fine-grained GAMM analysis also revealed evidence of a clear discontinuity in the age curve for some measures.

On the basis of the longitudinal data in the visualizations, several observations can be made. First, it is apparent that the development of writing skills in the first grades of (pre)primary education differs between students; for instance, some students hardly appear to learn to spell correctly, whereas others show a clear development in accuracy scores (for their oral performance, see Pfenninger in press). Second, the difference between AO groups in L2 development scores seems to decrease. Third, a particular contribution of the current study is that it constitutes another demonstration of the heterogeneity of within-person covariance structures – and due to a sufficient sample size, it allows direct comparisons of within-person and between-person factor structures (see Pfenninger & Bülow, this special issue). Specifically, the patterns observed at the level of the group do not seem to be representative for those displayed by the individual learners, which appears to confirm the assertion in CDST-related studies that one should be cautious when drawing conclusions on the basis of single learners (Lowie and Verspoor 2019). Fourth, the results refute the hypothesis that free variation occurs during an early stage of development and then disappears as learners develop better organized L2 systems (e.g., Ellis 1985). One explanation for this – at least with respect to e.g. lexical richness – is that as non-systematic variability decreases with time, systematic variability – e.g., stylistic range – increases (see Pfenninger in press for a more detailed analysis of this). Nevertheless, the observed connection between significant periods of L2 growth and variability corroborates the studies that have shown that increased variability coincides with a developmental jump (e.g., van Dijk et al. 2011): “if there is no variability, there can be no development” (Lowie and Verspoor 2015: 76).

Finally, the study illustrates that the developmental process of the different AO groups can only be approximated by extended time series and cannot be inferred by measurements at one or two points in time. Although age effects might not be readily considered as variable a factor as e.g., motivation, it turned out that measurements at different moments in time gave a different picture. In other words, conclusions about the eventual attainment are strongly dependent on the coincidental time of the measurement (see also Lowie and Verspoor 2015). However, due to the complex status of age as a “macro-variable” (Montrul 2008: 1) associated with a myriad of factors, it needs to be investigated in a next step whether this variability is due to other co-varying environmental and socio-affective factors. The narrow focus on accuracy, fluency, complexity and vocabulary traits in isolatable domains such as lexis and grammar does not capture defining aspects of advanced levels of ability, particularly the “socially embedded and situationally motivated nature of language use that addresses a vast array of concerns in human life” (Ortega and Byrnes 2008, p. 282). An important direction for future research will thus be to integrate quantitative data with qualitative data in order to better understand the process of individual learners’ development, e.g., significant increases and decreases in the L2 trajectories. To this end, the participants and their parents and teachers have filled in language awareness questionnaires, parental questionnaires and teacher questionnaires (see Pfenninger 2020). Such an approach will focus not only on the process itself and on quantification of change but also on the underlying environmental, biological, or psychological *reasons* for change.

References

- Bulté, Bram & Alex Housen. 2020. A DUB-inspired case study of multidimensional L2 complexity development: Competing or connected growers? In W. Lowie, M. Michel, A. Rousse-Malpat, M. Keijzer & R. Steinkrauss (eds.), *Usage-based dynamics in second language development: In celebration of Marjolijn Verspoor*, 50–86. Bristol, UK: Multilingual Matters.
- Byrnes, Heidi. 2009. Emergent L2 writing ability in a curricular context: A longitudinal study of grammatical metaphor. *Linguistics and Education* 20. 50–66.
- Dörnyei, Zoltan, Peter MacIntyre & Alastair Henry. 2015. *Motivational dynamics in language learning*. Bristol: Multilingual Matters.
- Ellis, Rod. 1985. Sources of variability in interlanguage. *Applied Linguistics* 6(2). 118–131.
- Ellis, Nick. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* 24. 143–188.

- Festman, Julia. 2018. Vocabulary gains of mono- and multilingual learners in a linguistically diverse setting: Results from a German-English intervention with inclusion of home languages. *Frontiers in Communication*. <https://doi.org/10.3389/fcomm.2018.00026> (accessed 29 May 2019).
- Hyltenstam, Kenneth. 1977. Implicational patterns in interlanguage syntax variation. *Language Learning* 27. 383–411.
- Jaekel, Nils, Michael Schurig, Merle Florian & Markus Ritter. 2017. From early starters to late finishers? A longitudinal study of early foreign language learning in school. *Language Learning* 67(3). 631–664.
- Kormos, Judit & Kata Csizér. 2008. Age-related differences in the motivation of learning English as a foreign language: Attitudes, selves, and motivated learning behavior. *Language Learning* 58(2). 327–355.
- Kovelman, Ioulia, Stephanie A. Baker & Laura-Ann Petitto. 2008. Age of first bilingual language exposure as a new window into bilingual reading development. *Bilingualism: Language and Cognition* 11(2). 203–223.
- Larsen-Freeman, Diane. 2006. The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics* 27. 590–619.
- Larsen-Freeman, Diane. 2009. Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics* 30(4). 579–589.
- Larsen-Freeman, Diane. 2015. Complexity theory: The lessons continue. In Lourdes Ortega & ZhaoHong Han (eds.), *Complexity theory and language development. In celebration of Diane Larsen-Freeman*, 11–50. NY and Amsterdam: Benjamins.
- Larsen-Freeman, Diane & Lynne Cameron. 2008. *Complex systems and applied linguistics*. Oxford: OUP.
- Lowie, Wander. 2017. Lost in state space? Methodological considerations in Complex Dynamic Theory approaches to second language development research. In Lourdes Ortega & ZhaoHong Han (eds.), *Complexity theory and language development. In celebration of Diane Larsen-Freeman*, 123–141. NY and Amsterdam: Benjamins.
- Lowie, Wander & Marjolijn Verspoor. 2015. Variability and variation in second language acquisition orders: A dynamic reevaluation. *Language Learning* 65. 63–88.
- Lowie, Wander & Marjolijn Verspoor. 2019. Individual differences and the ergodicity problem. *Language Learning* 69(S1), 184–206.
- Lowie, Wander, Rika Plat & Kees de Bot. 2014. Pink noise in language production: A nonlinear approach to the multilingual lexicon. *Ecological Psychology* 26(3). 216–228.
- MacIntyre, Peter & Alicia Serroul. 2015. Motivation on a Per-Second Timescale: Examining approach–avoidance motivation during L2 task performance. In Zoltan Dörnyei, Peter MacIntyre & Alastair Henry (eds.), *Motivational dynamics in language learning*, 109–138. Bristol: Multilingual Matters.
- Manchón, Rosa M. (ed.). 2011. *Learning-to-write and writing-to-learn in an additional language*. Philadelphia/Amsterdam: John Benjamins.
- McCarthy, Philip M. & Scott Jarvis. 2010. MTL, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 42. 381–392.
- McWhinney, Brian. 2000. *The CHILDES project: Tools for analyzing talk*, 3rd edn. Mahwah, NJ: Lawrence Erlbaum.
- Montrul, Silvina. 2008. *Incomplete acquisition in bilingualism: Re-examining the age factor*. Amsterdam: John Benjamins.
- Muñoz, Carmen (ed.). 2006. *Age and the rate of foreign language learning*. Clevedon: Multilingual Matters.
- Ortega, Lourdes. 2014. Trying out theories on interlanguage: Description and explanation over 40 years of L2 negation research. In ZhaoHong Han & Elaine Tarone (eds.), *Interlanguage: Forty years later*, 173–202. Amsterdam: John Benjamins.
- Ortega, Lourdes & Heidi Byrnes (eds.). 2008. *The longitudinal study of advanced L2 capacities*. London, New York: Routledge.
- Penris, Wouter & Marjolijn Verspoor. 2017. Academic writing development: A complex, dynamic process. In Simone E. Pfenninger & Judit Navracsis (eds.), *Future research directions for applied linguistics*, 215–242. Bristol: Multilingual Matters.
- Petitto, Laura-Ann & Ioulia Kovelman. 2003. The Bilingual Paradox: How signing–speaking bilingual children help us resolve bilingual issues and teach us about the brain’s mechanisms underlying all language acquisition. *Learning Languages* 8. 5–18.
- Pfenninger, Simone E. Emergent bilinguals in a digital world: A dynamic analysis of long-term L2 development in (pre)primary school children. *International Review of Applied Linguistics IRAL*, In press.
- Pfenninger, Simone E. & David Singleton. 2019. Starting age overshadowed: The primacy of differential environmental and family support effects on L2 attainment in an instructional context. *Language Learning* 69(S1). 207–234.
- Pfenninger, Simone E. 2020. The dynamic multicausality of age of first bilingual language exposure: Evidence from a longitudinal CLIL study with dense time serial measurement. *The Modern Language Journal* 104(3). 662–686.
- R Core Team. 2019. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: <http://www.R-project.org/>.
- Sorace, Antonella. 2005. Selective optionality in language development. In Leonie Cornips & Karen P. Corrigan (eds.), *Syntax and variation: Reconciling the biological and the social*, 55–80. Amsterdam & Philadelphia: John Benjamins.
- Spoelman, Marianne & Marjolijn Verspoor. 2010. Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. *Applied Linguistics* 31(4). 532–553.
- Tarone, Elaine. 1983. On the variability of interlanguage systems. *Applied Linguistics* 15. 285–295.
- Van Dijk, Marijn, Marjolijn Verspoor & Wander Lowie. 2011. Variability and DST. In Marjolijn Verspoor, Kees de Bot & Wander Lowie (eds.), *A dynamic approach to second language development: Methods and techniques*, 55–84. Amsterdam: John Benjamins.
- van Geert, Paul & Marjolijn Verspoor. 2015. Dynamic systems and language development. In Brian McWhinney & William O’Grady (eds.), *The handbook of language emergence*, 537–556. Oxford: Wiley-Blackwell.

- van Rij, Jacolien, Martijn Wieling, Harald R. Baayen & Hedderik van Rijn. 2016. *itsadug*: Interpreting time series and autocorrelated data using GAMMs. R package version 2.2.
- Verspoor, Marjolijn & Hana Smiskova. 2012. Foreign language development from a dynamic usage-based perspective. In Rosa M. Manchón (ed.), *L2 writing development: Multiple perspectives*, 17–46. Berlin: Walter de Gruyter.
- Wood, Simon. 2006. *Generalized additive models: An introduction with R*. Boca Raton, FL: CRC Press.

Supplementary Material: The online version of this article offers supplementary material (<https://doi.org/10.1515/lingvan-2020-0028>).