A LONGITUDINAL STUDY OF ADVANCED LEARNERS’ LINGUISTIC DEVELOPMENT
BEFORE, DURING, AND AFTER STUDY ABROAD

Kevin McManus, Pennsylvania State University, USA, kmcmanus@psu.edu
Rosamond Mitchell, University of Southampton, UK, r.f.mitchell@soton.ac.uk
Nicole Tracy-Ventura, West Virginia University, USA, nicole.tracyventura@mail.wvu.edu

ABSTRACT
The current study investigated advanced L2 learners’ linguistic development before, during, and after a nine-month stay abroad, the extent to which contextual changes (home-abroad-home) influenced the nature and magnitude of development, and the ways in which relationships among different linguistic elements changed over time. Participants were 56 university learners majoring in French (n = 29) and Spanish (n = 27), who spent an academic year abroad in the middle of a four-year BA degree program. Oral data were collected six times over 21 months to trace development and change in complexity, accuracy, fluency, and lexis (CAFL). Results showed ongoing improvements over time on most measures, including accuracy. Correlations indicated long-term relationships between fluency and vocabulary only and that accuracy-complexity relationships emerged in instructed home contexts only. These findings suggest that the affordances of home and abroad contexts can shape learners’ linguistic development and use differently. The role of pre-departure linguistic ability is discussed as critical to understanding the nature and extent of L2 linguistic development in study abroad.
INTRODUCTION

Study abroad (SA) research that documents learners’ linguistic abilities before, during, and after study abroad makes an important contribution to understanding how context can shape L2 development (Mitchell, Tracy-Ventura & McManus 2017; Pérez-Vidal 2014, 2017). For example, while study abroad appears to benefit oral fluency and vocabulary skills to a greater extent than at-home instructed contexts (e.g., Segalowitz & Freed 2004; Pizziconi 2017), the reverse has often been found for accuracy and syntactic complexity skills (e.g., DeKeyser 2017; Isabelli 2007). This patterning of results has been explained by social/interactional differences between classroom and SA contexts and, to some extent, learners’ pre-departure linguistic abilities and/or length of prior classroom study (for recent reviews, see Isabelli-Garcia et al. 2018; Taguchi & Collentine 2018). In contrast to some theories of L2 learning that prioritize meaningful language use and interaction (e.g., Duff 2018) or linguistic knowledge (e.g., Rothman & Slabakova 2019), Skill Acquisition Theory (DeKeyser 2017) assigns critical roles to both: L2 development is dependent on linguistic knowledge and opportunities for language use that rehearse that knowledge. A mismatch between the linguistic knowledge base and opportunities for language use is understood to be problematic for L2 development. Specifically, opportunities for language use that draw on a limited linguistic knowledge base likely lead to limited L2 development (DeKeyser 1997, 2017).

To date, however, the role of prior linguistic knowledge in understanding the nature and extent of linguistic development during SA is not well understood. One reason for this is that the participants most often investigated in SA research are North American college students in short-term programs who study abroad with little prior language instruction (Dietrich 2018; Sanz & Morales-Front 2018). These learners tend to have limited classroom-acquired linguistic
knowledge base (declarative knowledge) available for proceduralization/automatization through language use. As a result, these learners must rely on other means to establish and proceduralize new linguistic knowledge. In contrast, more proficient/advanced learner populations who are already equipped with a large amount of declarative knowledge are likely to make fuller use of the opportunities available abroad to rehearse existing linguistic knowledge. A focus on more proficient/experienced learner populations has the potential to advance our understanding of how context facilitates those later stages of acquisition. Therefore, longitudinal research is critically needed with advanced-level classroom learners in both SA and instructed contexts in order to better understand how SA shapes L2 development.

This study addresses this gap in two ways. First, it provides a systematic longitudinal account of spoken L2 development in two languages, French and Spanish. The study took place over 21 months, including six total data collection waves: three at-home (Presojourn, Postsojourn-1 and Postsojourn-2) and three abroad, allowing for a comparison across two learning contexts. The inclusion of three data collection waves while participants were abroad made it possible to examine the trajectory of development within the stay abroad, a variable not included in much previous SA research (for reviews, see Llanes 2011; Pérez-Vidal 2014, 2017).

Second, this study focuses on advanced instructed learners who had already built up a large amount of declarative knowledge over many years of classroom instruction. The change in context for them was anticipated, and they were motivated to take full advantage of authentic target language interactions. Consequently, this study has the potential to further our understanding of interactions between substantial prior linguistic knowledge and contextual influences (the classroom vs SA).
Consistent with previous SA research investigating linguistic development, we adopted the CAFL framework (complexity, accuracy, fluency, lexis) to conceptualize L2 development (Housen, Kuiken & Vedder, 2012). While CAFL is primarily a descriptive framework, its findings can usefully be interpreted with reference to relevant theoretical accounts of L2 development, including Skill Acquisition Theory (DeKeyser 2017).

In the following sections, we begin by reviewing previous research on L2 linguistic development during SA, paying close attention to the development of different CAFL elements, as well as relationships between them. We then review theoretical accounts to explain these developmental trajectories.

LINGUISTIC DEVELOPMENT IN STUDY ABROAD

Research broadly confirms the widespread belief that SA benefits oral fluency (Juan-Garau 2018; Mora and Valls-Ferrer 2012; Segalowitz and Freed 2004; Valls-Ferrer and Mora 2014). Pretest-posttest SA comparisons of speed fluency (e.g., speech rate) have repeatedly shown major improvement (Du 2013), both for short stays lasting three to four weeks (Llanes and Muñoz 2009) and longer stays of one semester (Segalowitz and Freed 2004). However, Llanes and Serrano (2011) found ceiling effects for fluency when comparing two-month and three-month SA programs: an extra month abroad appeared to afford no additional speed fluency benefits (see also Huensch & Tracy-Ventura 2017; Lara, Mora and Pérez-Vidal 2015). In addition, research suggests that length of prior classroom study and pre-departure proficiency play an important role in explaining both the rate of fluency gains during SA and their maintenance afterwards: while both low and high proficiency learners show fluency
improvements, gains are more rapid and longer lasting for more proficient learners (Valls-Ferrer and Mora 2014; Juan-Garau 2018; Leonard & Shea 2017).

SA has also been shown to improve learners’ vocabulary knowledge (Briggs 2015, Serrano, Tragant and Llanes 2012, Zaytseva, Pérez-Vidal and Miralpeix 2018, Tracy-Ventura, Mitchell & McManus 2016). Some research suggests that length of SA appears important in explaining the size and diversity of learners’ vocabularies. For example, Ife, Vives Boix, and Meara (2000) showed better performance on word association and translation tests for learners who studied abroad for two semesters rather than one (see also Serrano, Llanes and Tragant 2011 and Serrano et al. 2012). But, consistent with oral fluency, ceiling effects are detectable for stays longer than three-to-four months (Tracy-Ventura 2017). In addition, comparisons between SA and at-home learners have shown that the SA context offers greater advantages for vocabulary development than at-home contexts (Pérez-Vidal et al. 2015). Regarding the impact of pre-departure proficiency, Leonard and Shea (2017) found that more advanced learners made greater gains in both lexical diversity and lexical sophistication during SA.

Despite these generally encouraging accounts of fluency and lexis improvement during SA, research with low-to-intermediate proficiency participants generally indicates few SA advantages for morphosyntactic accuracy (e.g., Collentine 2009; Isabelli-Garcia 2010; DeKeyser 2010, see Howard & Schwieter 2018 for review). Recent studies with more proficient learners, however, indicates accuracy improvement (Leonard & Shea, 2017; Polat and Kim 2014, Rees and Klapper 2007, Serrano et al. 2012), with accuracy gains sustained after SA (e.g., Juan-Garau 2014; Pérez-Vidal 2017). As for syntactic complexity, even though development during SA seems limited (Lara, Mora & Pérez-Vidal 2015; Mora & Valls-Ferrer 2012; Pérez-Vidal & Barquin 2014), Leonard and Shea’s (2017) experienced participants (with 6 years’ prior study of
L2 Spanish) showed significant development (length of t-unit, subordinate clauses per t-unit), suggesting that the complexity of learners’ utterances is potentially amenable to change during SA among more experienced/proficient L2 learners.

One important limitation of SA research focusing on linguistic development is that we know very little about how different linguistic elements develop in relation to each other. This is because research has predominantly examined how SA contributes to the development of individual linguistic features or areas of L2 performance. To date, two studies have explored relationships among CAFL measures after SA. Both included advanced learners with data collected (i) at the beginning and end of a three-month stay abroad (Leonard & Shea 2017) and (ii) after a three-month stay abroad (Mora & Valls-Ferrer 2012), and indicate some relationships between linguistic elements. For example, Leonard and Shea’s (2017) analysis of CAFL relationships showed pre-SA relationships (positive correlations) among fluency, lexical complexity, and accuracy, though syntactic complexity did not correlate with any other CAFL element. At post-test, participants showed significant gains on all the measured CAFL dimensions. However, no significant relationships were found among the various individual CAFL gain scores. The authors suggest that, in the short term, CAFL elements might develop independently of each other (e.g., short term improvement in fluency is independent from accuracy), potentially indicating that CAFL relationships emerge only slowly, driven by extended language use. However, pre-post gain scores are relatively broad and imprecise measures of change that cannot show the trajectories of CAFL development and relationships among them. To better understand these issues, studies are needed that document in more detail the emergence and development of CAFL relationships over the course of SA.
In summary, research documenting linguistic development during SA indicates notable benefits for fluency and vocabulary, but findings are less clear for accuracy and syntactic complexity. As noted in previous reviews (see Collentine 2009; Howard & Schweiter 2018), our evidence base about linguistic development resulting from SA is primarily based on low-proficiency North Americans taking part in short-term SA programs (although longstanding exceptions exist, especially involving European learners, e.g. Perez-Vidal 2014). One limitation of short stays abroad using before-after designs is that these studies have not sought to investigate development over the course of or much after SA, which necessarily limits our understanding about the ways in which different learning contexts contribute to linguistic development (see Ren 2018; Sanz & Morales-Front 2018).

THEORIZING LINGUISTIC DEVELOPMENT IN SA
A major challenge for L2 research is how to explain the patterns of linguistic development resulting from SA, including repeated observations that the linguistic benefits of SA appear selective (e.g., fluency over accuracy). One explanation is that different contexts (e.g., instructed, abroad) and different activities/tasks (e.g., meaning-focused, form-focused) lead to linguistic “trade-offs” that prioritize specific linguistic abilities (Jackson and Suethanapornkul 2013; Michel 2017; Robinson 2005; Skehan 2009, but see Leonard & Shea 2017, who found evidence for neither trade-offs nor positive relations). In task-based language learning, for example, tasks based on concrete or familiar information can advantage accuracy and fluency, while interactive tasks appear to advantage accuracy but not fluency (Skehan 2001, 2009). In addition, trade-offs between fluency and accuracy have been found when attentional resources are taxed, such as in very complex, multi-stimuli activities (Robinson, 2005).
To date, this body of research has been used to understand CAFL relationships and the extent to which the conditions/contexts of language use potentially influence performance. Anchoring these understandings about CAFL relationships within Skill Acquisition Theory (SAT, DeKeyser 2017) has the potential to help better theorize linguistic development during SA given that SAT conceptualizes L2 learning as driven by language use. Research from task-based language learning has the potential to inform the ways in which CAFL elements are related, while SAT has the potential to inform how CAFL elements develop.

SAT proposes that three types of knowledge underpin linguistic development. First, establishing reliable and accurate declarative knowledge (i.e., knowledge about language) is argued to be an essential first step (usually established explicitly in L2 learning, but not necessarily, see DeKeyser 2017). Procedural knowledge is understood to underpin the conscious rule-governed behaviour that rehearses this declarative knowledge. Procedural knowledge is reflected in production that is mostly accurate, but slow and effortful. Over time, extended practice and opportunities for meaningful language use can lead to automatization, “a fast, parallel, fairly effortless process that is not limited by short-term memory capacity, is not under direct subject control, and is responsible for the performance of well-developed skilled behaviors” (Schneider, Dumais, & Shiffrin 1984, p.1).

Applied to SA, the informal interactions typically afforded during SA are largely, though not exclusively, communicatively-driven and meaning-focused (see Kinginger & Blattner 2008; Mitchell et al. 2017), indicating that this context likely prioritizes fluent over accurate and complex production (Robinson, 2005; Skehan, 2009). Repeated and extended opportunities to engage in communicatively-driven and meaning-focused activities will likely drive proceduralization and potential automatization of declarative knowledge (DeKeyser 2017). That
is, in CAFL terms, these types of activities will likely promote fluency over accuracy. In addition, learners with extensive declarative and/or procedural knowledge would be expected to demonstrate more rapid fluency gains (than learners with limited declarative and/or procedural knowledge) because these learners have a larger declarative knowledge based available for proceduralization/automatization (DeKeyser, 2017).

On the other hand, instructed contexts are hypothesized to prioritize the development of accuracy and complexity skills given that these contexts tend to be more form-focused than communicatively driven. In contrast to increasing fluency which exploits existing knowledge, accuracy improvement requires that new declarative knowledge needs to be established, or existing knowledge restructured, which is a slow and gradual process that requires extended exposure to language, and is made more complex by competition of various kinds with existing L1 morphosyntactic knowledge (see also Schmidt, 1992). Again, repeated and extended opportunities for slow and effortful language use that rehearses declarative knowledge newly established from instruction can lead to proceduralization of these skills (DeKeyser 2017).

In sum, fluency improvement is hypothesised to be more detectable during SA, especially for learners with extensive linguistic knowledge prior to SA. This is because fluency improvement operates on existing knowledge by optimizing the access and assembly procedures of stored linguistic knowledge for use in real time. Accuracy and syntactic complexity are likely to develop more slowly because this knowledge must be established and/or restructured. Lexical knowledge is also hypothesised to improve more quickly than grammatical knowledge, in meaning-focussed settings such as SA. In instructed contexts, however, accuracy and complexity might develop more consistently than fluency given that an important focus of classroom learning is to promote attention to morphosyntax. These accounts suggest potential between-
context differences for linguistic development and for the relationships among the CAFL components. These proposals inform the current study.

CURRENT STUDY
The current study begins to address the issues identified above by (i) investigating advanced level L2 French and L2 Spanish learners’ CAFL development over 21 months including a nine-month stay abroad; and (ii) investigating the extent to which relationships among CAFL elements change over time in L2 speech. Based on previous research, we expected that oral fluency and lexis would improve the most during SA, and also that gains in these domains would be evident early on in the sojourn. As SA research on accuracy is mixed, and our understanding of syntactic complexity changes during SA is limited, we could not adopt strong expectations for development in these domains. However, given the relatively advanced starting proficiency of our participants, we expect that our investigation of a longer stay abroad (i.e., one academic year) might provide evidence for accuracy and syntactic complexity development not detectable using shorter research designs. In terms of relationships among CAFL elements, since informal interactions during SA can be expected to be mostly meaning-focused, we expected that SA contexts would promote positive relationships between fluency and lexis. In contrast, we expect positive relationships between accuracy and complexity when participants return home given that instructional contexts tend to emphasize accuracy and complexity. These are, however, tentative predictions because of the limited nature and somewhat conflicting conclusions of previous research.
RESEARCH QUESTIONS

1. What is the overall pattern of development of CAFL among advanced instructed learners of L2 French and Spanish, before, during and following a period of SA?

2. What are the evolving relationships between different CAFL strands over time, before, during and following SA?

METHOD

Participants

Participants were university students majoring in either French (n=29) or Spanish (n=27) at a large university in England and were recruited at the end of year two (of a four-year degree) before studying abroad for nine months during their third year. Participants were experienced classroom learners who could be described as language-oriented students because (i) they were majoring in languages, against current trends at that time (British Academy 2011), and (ii) they indicated in interviews that their main reason for studying abroad was language improvement (see Mitchell et al. 2017). We recruited learners of both French and Spanish in order to inform understanding about the contribution of SA to linguistic development in the UK’s most studied foreign languages (and development in these two languages is treated comparatively in the study). Furthermore, we planned for participant attrition given this study’s longitudinal design over 21 months. Recruiting learners from well represented languages was therefore critical.

The L2 French participants included 27 L1 English speakers, a L1 Spanish speaker, and a L1 Finnish speaker. The mean age was 21 (range = 20-24) and mean length of previous French study was 11 years (range = 5-17). Following pathways typical of the British ‘third year abroad’, six participants spent the academic year as workplace interns, 15 as teaching assistants, and eight
as university exchange students, all dispersed in urban and rural France (see Mitchell et al. 2017).

The L2 Spanish participants were 25 L1 English speakers and two L1 Polish speakers. The mean age was 20.5 (range = 20-25) and mean length of previous Spanish study was 5.9 years (range =2-14), i.e., shorter than the mean length of previous study for the French group. Two participants were workplace interns, sixteen were teaching assistants, and nine were exchange students at Spanish universities. Eighteen participants sojourned to different parts of Spain and nine (all teaching assistants) in Mexico. A main SA contextual difference between the French and Spanish groups was that the former were much more likely to live in international student housing, while the participants in Spain and Mexico were much more likely to live in shared flats or homes with local Spanish speakers, and consequently reported somewhat greater use of the target language in daily life throughout their stay (see Mitchell et al. 2017).

**Design and outcome measures**

The study included six testing points over 21-months: Presojourn in the UK (May 2011), Insojourn1 abroad (October 2011), Insojourn2 abroad (February 2012), Insojourn3 abroad (May 2012), Postsojourn1 in the UK (October 2012), and Postsojourn2 in the UK (February 2013). The participants completed a range of tasks, including an oral interview, an oral picture-based narrative, and a written argumentative task, as well as receptive and productive proficiency measures. All tasks were piloted prior to data collection with five comparable learners per language. Tasks were also completed by comparison groups of L1 French and L1 Spanish speakers (n=10 per language).
Spoken picture-based narrative

This study draws on the picture-based stories in either French or Spanish which were narrated by participants at each data collection point. This task was selected as it was considered to provide the most robust comparable evidence on the various CAFL strands. We used three stories in rotation to minimize potential task familiarity effects (see Dominguez et al. 2013; McManus 2015; Mitchell et al. 2017). Each story was repeated once, 12 months apart: the “Cat” story at Presojourn (May 2011) and Insojourn3 (May 2012), the “Sisters” story at Insojourn1 (October 2011) and Postsojourn1 (October 2012), and the “Brothers” story at Insojourn2 (February 2012) and Postsojourn2 (February 2013). In order to maintain between-story equivalency, the stories were designed to be structurally similar: all stories were set in the past, contained similar numbers of colour images, contrasted perfective and habitual events, and contained two characters. To test for potential between-story differences, the stories narrated by L1 speakers and then compared on two oral fluency measures (see ‘Linguistic analyses’ for operationalization). These comparisons indicated no reliable differences (all confidence intervals (CI) for Cohen’s $d$ passed through zero): Mean length of run (Cat vs. Sisters, $d$ [CI] = -.20 [-1.07, .69], Cat vs. Brothers, $d$ [CI] = .20 [-.69, 1.07], Sisters vs. Brothers, $d$ = .42 [-.48, 1.29]) and speech rate (Cat vs. Sisters, $d$ [CI] = -.32 [-1.19, .57], Cat vs. Brothers, $d$ [CI] = -.03 [-.91, .85], Sisters vs. Brothers, $d$ = .31 [-.58, 1.18]). As previously noted, each story was administered approximately twelve months apart (e.g., cat story in May 2011 and May 2012) in order to reduce any familiarity effects.

Participants first previewed the story (< 2 mins) and could ask clarification questions before retelling the story with continuing access to the pictures. The mean narration time was 3:47 minutes.
Data preparation

Oral narratives were digitally recorded and transcribed following CHAT conventions (MacWhinney, 2000). Transcribers were L1 speakers or advanced-level L2 speakers of either French or Spanish. Important for further analysis, line breaks were introduced at the start of each new Analysis of Speech unit (ASU). ASUs were defined following Foster et al. (2000, p.365): “a single speaker’s utterance consisting of an independent clause or subclausal unit, together with any subordinate clause(s) associated with either”. Transcription accuracy was checked by at least two members of the research team before analysis. All files were automatically tagged using CHAT’s part-of-speech morphosyntactic tagger (MOR and POST commands), and manually disambiguated.

Linguistic analyses

Fluency

Oral fluency was operationalized as 1) speech rate, the number of syllables/total speaking time (in seconds), including pauses, and 2) mean length of run, the mean number of syllables between silent pauses of .25 seconds or greater (Freed, Segalowitz, and Dewey, 2004; Towell, Hawkins, and Bazergui, 1996). PRAAT was used to identify and measure pause length. The research team created protocols for the manual counting of syllables in ELAN. For each file, six trained coders (three per language) listened to the audio and read the transcript. Any disagreements were discussed until coders agreed. All acoustic syllables were counted, including filled pauses, repetitions, and repairs in order to capture speakers’ global fluency abilities (Gilabert & Munoz, 2010).
Lexis

Lexis was operationalized as lexical diversity, computed using the VocD command in CLAN (MacWhinney, 2000), resulting in a score $D$ (Malvern & Richards, 2002). $D$ is computed by taking random samples of a text, calculating type-token ratios of these samples, and then using these type-token ratios to create a random-sampling type-token ratio for the text (see Malvern & Richard, 2002). $D$ scores tend to range from 10-100; the higher the score, the greater diversity of words produced.

Accuracy

Following recommended practice in the field (Housen et al. 2012), we calculated two broad measures of accuracy: percentage of error-free ASUs (%EF-ASU), and percentage of error-free clauses (%EF-C). Error types were of the following types: missing or incorrect inflections (nominal and verbal), auxiliaries, prepositions, gender attribution, and agreement, selection of inappropriate lexical item, incorrect word order. Mispronunciations not affecting morphosyntactic accuracy judgements were not classified as errors.

Syntactic complexity

Two commonly used measures of syntactic complexity were calculated: the ratio of clauses to ASUs (Clauses/ASU) and mean length of ASUs (ML-ASU)$^1$. A clause was defined following Bulté and Housen (2012, p. 39): “a unit consisting of a subject (visible or implied) plus a predicate, i.e., a construction with a finite or nonfinite predicator or verb as its head”. We treated modal verbs (e.g., pouvoir, vouloir in French; deber, tener que, poder, haber que in Spanish) as
auxiliaries and therefore bare nonfinite VPs following these verbs were treated as complement phrases but not as separate clauses.

Transcripts were divided into ASUs and clauses following detailed analysis protocols. Adequate reliability was assured through double rating of transcripts and discussion. Every new ASU started on its own line with the participant number (e.g., *129). The symbol \[^ c\] was manually inserted within the ASU line where appropriate to mark clause boundaries. Following Jensen and Howard (2014), repetitions, corrections and hesitations were ignored for these syntactic complexity analyses. CLAN software was then used to calculate ML-ASU and Clauses/ASU.

**Statistical analyses**

Because all datasets were normally distributed (according to Shapiro-Wilks tests, all datasets p > .05), we present the results of parametric tests.

First, a two-way 2 x 6 repeated measures ANOVA with L2 Group (French, Spanish) as the between-subjects factor and Time (Presojourn, Insojourn1, Insojourn2, Insojourn3, Postsojourn1, and Postsojourn2) as the within-subjects factor was used to compare scores on all eight CAFL measures (as described above) from each of the six data points. The alpha level was set at .05.

Second, we calculated Pearson product-moment correlations with bootstrapped confidence intervals (CIs) for each of the CAFL measures at every data collection point in order to examine the nature of CAFL relationships and how these changed over time.

To interpret magnitudes of change, we present Cohen’s d effect sizes (ES) with CIs for d for all paired comparisons (within-subject over time and between-subject at each data point), and
values with bootstrapped CIs for all correlations. Within-subject ES were calculated using means and standard deviations of two test phases. Between-subject ES were calculated using means and standard deviations of two groups at the same test point. ES ($d$, $r$) were interpreted using Plonsky and Oswald’s (2014) field-specific benchmarks: within-subject: $ds$ around .60 are considered small, 1.00 medium, and 1.40 large; between-subject: $ds$ around .40 are considered small, .70 medium, and 1.00 large; and $r$ field-specific benchmarks: $rs$ close to .25 are considered small, .40 medium, and .60 large. CIs that include zero are additionally used to interpret ES reliability.

RESULTS
Results are presented by CAFL element. For each measure, we first present change over time (see Tables 2 and 4) and then between-group comparisons, including Cohen’s $d$ ES (see Table 2) (for CIs for $d$, see Supplementary Data). Correlations are then presented between the different CAFL elements at each test phase (Presojourn, Insojourn-1, Insojourn-2, Insojourn-3, Postsojourn-1, and Postsojourn-2: see Table 4). Descriptives for all measures are presented in Supplementary Data.

Fluency
Figure 1 shows the distribution of learners’ speech rate (SR) and mean length of run (MLR) scores over time. For SR, the statistically significant Time x Group interaction ($F(5, 44) = 5.28$, $p < .001$) indicated French-Spanish differences over time. Although no between-group differences were found Presojourn and at Insojourn1 (negligible ES), SR was faster among Spanish learners at Insojourn2 and Insojourn3 (large ES), and these differences were maintained
after SA at Postsojourn1 and Postsojourn2 (medium ES). We also found a statistically significant main effect for Time ($F(5, 44) = 76.23, p < .001$) because Presojourn scores were significantly lower than all other data points (large ES, see Table 3).

A similar pattern of results was found for MLR. We found a statistically significant Time x Group interaction ($F(5, 41) = 39.82, p < .001$) due to higher MLR scores in the Spanish group than the French group towards the end of SA (medium ES at Insojourn2, small ES at Insojourn3). In terms of within-group performance over time, a statistically significant main effect for Time ($F(5, 41) = 39.82, p < .001$) indicated that fluency scores were lower at Presojourn than at all other test points. The magnitudes of these differences over time were medium-to-large for both French and Spanish learners. Following the return to instructed contexts, a small decline in fluency was found (between Postsojourn1 and Postsojourn2) on both measures, but these did not substantially negate the gains made in-sojourn.

[Figure 1]

**Lexis**

Figure 2 shows the distribution of lexical diversity (D) scores over time. The (borderline) statistically significant Time x Group interaction ($F(5, 47) = 2.40, p = .051$) indicated that lexical diversity scores over time were, to some extent, influenced by group because of higher scores in the Spanish group. The magnitudes of these differences were medium before and after SA (Postsojourn1), and large in-sojourn. We also found a statistically significant main effect for Time ($F(5, 47) = 76.04, p < .001$) because lexical diversity scores at all other time points were higher than at Presojourn (medium-to-large ES for both groups). Table 1 indicates significant
improvement in lexical diversity scores following the move abroad (at Insojourn1), followed by further movement on immediate return to the instructed context (Postsojourn1).

[Figure 2]
[Table 1]

Accuracy

Figure 3 shows accuracy scores over time for percentage of error-free clauses (%EF-C) and percentage of error-free AS-units (%EF-ASU).

For %EF-C, results indicate a Time x Group interaction ($F(5, 48) = 4.31, p = .003$) but this time French learners performed more accurately than Spanish learners at Presojourn (medium ES), Insojourn1 (small ES) and Insojourn3 (medium ES). There was also a statistically significant main effect for Time ($F(5, 48) = 19.79, p < .001$) because %EF-C scores at all other test phases were more accurate than at Presojourn ($p < .001$). The magnitudes of these improvements were greater for Spanish learners (medium-to-large ES) than for French learners (small-to-medium ES) (see Table 3). Spanish learners additionally continued to improve postsojourn (small ES), while the French group did not, perhaps due to a ceiling effect for the French group.

Consistent with %EF-C, main effects for Time were statistically significant for %EF-ASU ($F(5, 48) = 17.93, p < .001$) because of less accurate performance at Presojourn for both groups of learners (small ES). Both groups made little further progress.

[Figure 3]
[Table 2]
[Table 3]
**Syntactic complexity**

Figure 4 visualizes summary findings for mean length of AS-unit (ML-ASU) and ratio of clauses to AS-units (clauses/ASU), showing varied performance over time for both groups.

For ML-ASU, we found between-group differences for performance over time ($F(5, 45) = 4.29, p = .003$) because of higher syntactic complexity scores in the French group at the beginning of the stay abroad (medium ES at Insojourn-1) and also after SA (large and medium ES at Postsojourn1 and Postsojourn2, respectively). Although we found a statistically significant main effect for Time ($F(5, 45) = 18.93, p < .001$), this is because Postsojourn scores in the instructed context were higher than those at Presojourn (French only, medium ES) and Insojourn3 (French and Spanish, medium-to-large and medium ES, respectively).

For Clauses/ASU, between-group differences were only found at Postsojourn1 and Postsojourn2 because of higher syntactic complexity scores in the French group (large and medium ES, respectively), even though the Time x Group interaction was not statistically significant ($F(5, 48) = 2.02, p = .092$). Within-group comparisons for performance over time was statistically significant ($F(5, 48) = 12.19, p < .00$), because Insojourn1 scores were higher than at Presojourn for both groups (French = small ES, Spanish = medium ES), indicating increased syntactic complexity early on in learners’ stay abroad. And, though for the Spanish learners only, we also found that Clauses/ASU scores improved on return to the instructed context after SA (at Postsojourn1, medium ES).

[Figure 4]

**Interim summary of longitudinal development in CAFL elements**
Figure 5 visualizes French and Spanish learners’ linguistic development for each measure, as they move from the instructed context to SA and back again. Using the within-group ES (Tables 2 and 4), these plots show performance over time with Presojourn as the baseline. Data points occurring within the red boxes are interpreted as reliable change compared to Presojourn (see Supplementary Data for ES). Both groups’ lexis and fluency scores are located within the red box, thus indicating meaningful and reliable improvement over time, whereas performance over time for accuracy and syntactic complexity is more variable, especially syntactic complexity.

[Figure 5]

**CAFL relationships over time**

The nature of the CAFL relationships at each data point were statistically examined using correlations, which indicated dynamic CAFL relationships over time. We report here all (borderline) significant correlations (see Supplementary Data for full reporting of all correlations).

At Presojourn, like Leonard and Shea (2017), we found significant relationships involving all CAFL elements (see Table 4). The strongest relationship was a positive correlation between lexis and fluency (medium). Lexis also positively correlated with syntactic complexity (small) and accuracy (small), and accuracy and fluency correlated positively (small).

[Table 4 here]

At Insojourn1, we found positive correlations between lexis and fluency only: a medium correlation with SR ($r = .498, p < .001$), and a borderline small correlation with MLR ($r = .270, p = .069$). Lexis and SR continued to correlate positively throughout the study, whereas correlations between lexis and MLR were more varied and often weaker (see Table 5). We
additionally found that Clauses/ASU (syntactic complexity) correlated negatively with both lexis ($r = -.436, p = .002$) and fluency (SR, $r = -.361, p = .011$).

[Table 5]

No other meaningful accuracy correlations were found until Postsojourn1, when small-to-medium positive correlations re-emerged between accuracy and fluency (%EF-C + SR: $r = .379$, $p = .005$; %EF-C + MLR: $r = .397$, $p = .004$; %EF-ASU +SR: $r = .352$, $p = .010$; %EF-ASU + MLR: $r = .324$, $p = .022$), and, at Postsojourn-2, borderline and small positive correlations emerged with lexis (%EF-C + D: $r = .253$, $p = .076$; %EF-ASU + D: $r = .268$, $p = .060$).

Towards the end of SA and on learners’ return to instructed contexts, correlations re-emerged between lexis and syntactic complexity: large-to-medium positive correlations with Clauses/ASU at Insojourn3 ($r = .595, p < .001$) and Postsojourn1 ($r = .343, p = .015$); and small-to-medium negative correlations with ML-ASU at Postsojourn1 ($r = -.291, p = .040$) and Postsojourn2 ($r = -.368, p = .009$). Syntactic complexity also correlated positively at Postsojourn1 with fluency (Clauses/ASU + SR, $r = .304$, $p = .027$) and at Postsojourn2 with accuracy (Clauses/ASU + % EF-C, $r = .313$, $p = .019$).

DISCUSSION

The present study investigated L2 French and L2 Spanish learners’ CAFL development over 21 months, measuring oral performance on a picture based narrative task once before, three times during, and twice after a nine-month stay abroad. We developed predictions regarding likely development both from previous empirical SA research and from our interpretation of Skill Acquisition Theory as relevant to the development of CAFL. In general, these predictions were upheld for both languages. Some minor differences did emerge between L2 development in
French and in Spanish, which can be attributed primarily to French participants’ somewhat longer prior experience of instruction (which led to some apparent ceiling effects, e.g., in accuracy), and to Spanish participants’ somewhat greater engagement in local social networks and higher levels of L2 use Insojour (which may partly account for their greater fluency gains). But, overall, the developmental pattern was similar across both groups.

Our expectation that fluency and lexis would benefit the most from SA was supported and corroborates previous research (Briggs 2015; Juan-Garau (2018); Llanes and Serrano 2011; Valls-Ferrer and Mora (2014); Pérez-Vidal 2014; Segalowitz and Freed 2004). ES magnitudes of improvement between Presojourn—Insojour1 and Presojourn—Insojourn2 were small-to-medium for fluency (SR, MLR) and medium-to-large for lexis, indicating considerable improvement over the course of SA. Our results offer more fine-grained understandings of fluency and lexis development than previous studies due to the inclusion of multiple data points during SA, indicating that contextual effects connected to the move abroad appear to quickly benefit fluency and lexis, a finding broadly consistent with Serrano et al. (2012). Furthermore, the gains made during SA for fluency and lexis appear to have been maintained on return to the instructed context.

It seems likely that the first, lengthy pre-SA period of instruction at home allowed participants to develop a substantial store of L2 knowledge, which was not yet fully proceduralized/automatized because opportunities for meaningful practice can be relatively absent in instructed settings. The period of SA with its richer opportunities for interaction made it possible for these relatively advanced instructed learners to quickly proceduralize/automatize this store of knowledge through immediate opportunities for meaningful practice, as suggested by DeKeyser (2007). Moreover, the period of SA allowed for this automatization to become
consolidated during the rest of the sojourn. While small declines in fluency were noted following the return home, development during SA appears to have been sufficiently advanced for levels of fluency not to have become seriously impaired.

Regarding lexis, a large increase in lexical diversity occurred immediately following the move abroad (Insojourn1). Unlike the sustained gains in fluency, however, lexical diversity then declined slightly in the course of SA (from Insojourn1 to Insojourn3), before rising again immediately following the return to the instructed context (Postsojourn1). This difference between these two CAFL components, which is not immediately predictable from SAT, may perhaps be accounted for by the context-specific nature of lexical learning. The initial move to the SA context exposed learners to a wide range of new daily experiences and types of meaning-focused interaction, which may account for the initial lexical gain. However, those contexts may then have become routinized as the sojourn progressed (see Mitchell et al. 2017 and McManus 2019 for evidence that social networking with locals declined somewhat over time). Thus, once an everyday vocabulary had been mastered of the kind required by the narrative task, there may have been relatively limited instrumental demand for further lexical development Insojourn. That the second change of context (the return to the classroom) was associated with a further rise in lexical diversity is intriguing, but our data does not allow us to account for this finding as we do not have classroom observation data.

Although we could not adopt strong expectations for L2 learners’ development of accuracy and syntactic complexity based on our synthesis of the literature, we thought that a nine-month stay abroad might also promote development for these CAFL subcomponents among our group of more advanced learners (compared to previous SA research). This tentative expectation was partially supported by the data which demonstrated evidence of improvement
following the move from the instructed context to SA, and continuing gradual improvement during the 9-month SA, thus supporting the accuracy findings of Serrano et al. (2012) and Leonard and Shea (2017). In terms of accuracy development after SA, we found improvement for accuracy in Spanish (%EF-C, %EF-ASU). Regarding complexity, we again see the most substantial improvement immediately following the move from the instructed context to SA, but this element then essentially plateaus (at least in French) during SA (though there is some ongoing improvement in Spanish, at least for Clauses/ASU). On return to the instructed context, however, complexity improves somewhat once more in both French (ML-ASU) and Spanish (Clauses/ASU, ML-ASU).

As previously suggested, these results join a growing body of research involving advanced learners showing that in addition to fluency and lexis, accuracy improvements during SA seem possible as well (e.g., Pérez-Vidal et al. 2015). These findings are in line with suggestions that SA allows advanced learners, in particular, to proceduralize/automatize their existing linguistic knowledge through opportunities for meaningful practice, and also to construct new morphosyntactic knowledge and/or restructure existing morphosyntactic knowledge, though this less explicit experience is a slow and gradual process (DeKeyser 2007). On return to the home university, accuracy continued to improve, perhaps reflecting more conscious effort once again on the part of participants to meet curricular expectations of the classroom (upon returning they still had one year of university to complete).

Our results comparing the relationships among CAFL elements over time showed significant and long-lasting relationships between fluency and lexis, thus contrasting with Mora and Valls-Ferrer (2012), who found no relationships among fluency and lexis after SA. It is possible that our use of a monologic task, in contrast to Mora and Valls-Ferrer’s oral interviews,
could explain this difference. Interviews typically require inter-speaker coordination, accommodation of interruptions, and other non-linguistic artefacts that are absent in narrative monologues, all of which could complicate the measurement of fluency and the subsequent correlational analyses investigating CAFL relationships. It is possible that our reliance on a picture-based narrative may have led to greater consistency of fluency measurement and relationships between fluency and lexis at most data collection waves.

Accuracy relationships among other CAFL elements were not stable (unlike the lexis—fluency pairing): Accuracy correlated with other measures before and after SA, but not during it. Our finding that accuracy correlated with fluency only before and after SA, but not during SA, could be related to participants’ developing sensitivity about the expectations of language use in the classroom and abroad (see also Kinginger & Blattner 2011). That is to say, classroom contexts potentially favour (morpho)syntactic development over fluency, while the SA context likely favours meaning-focused communication over (morpho)syntactic development, an idea also evidenced to some extent in participants’ comments about their language practices insojourn. Thus, it is possible that these learners were rebalancing attention to grammar and effective communication depending on the context, with grammatical accuracy prioritized in the classroom (before and after SA) and communicative adequacy prioritized during SA.

**Limitations and future research**

As previously noted, SA research has overwhelmingly examined linguistic development using before and after designs with the limitation that our understanding of development over the course of SA is not well understood. The present study contributes to this gap in our understanding, but more work is needed. For example, the present analyses are based on
learners’ oral productions of oral narratives. Picture-based oral narratives are popular elicitation
tasks in L2 research and provide robust evidence in key CAFL domains (notably fluency and accuracy). However, in other domains (complexity, lexis), they may be insufficiently demanding to provide a fully valid picture of advanced learners’ capabilities; future research would clearly benefit from widening the range of tasks to include more challenging and open-ended production tasks. A different set of tasks could lead to a different patterning of results, as indicated for example by work in task-based language teaching suggesting that task complexity can influence CAFL relationships (e.g., Skehan 2009).

Future research should also examine development of specific target features that are typically included in global accuracy measures, such as morphosyntactic agreement in gender and number (e.g., Edmonds & Gudmestad 2018, for evidence of improvement in grammatical gender during SA). As the research community continues to evaluate CAFL elements, understanding the development of features that are involved in CAFL operationalizations is important, as is understanding how CAFL elements develop together. One potentially rich area for future research would be to examine in depth the development of particular CAFL relationships (e.g., fluency-lexis) using multiple CAFL operationalizations in a variety of tasks.

CONCLUSION

The current study’s findings indicate that advanced instructed learners’ language skills appear sensitive to changes in learning context. Fluency and lexis appeared quick to develop, early in SA, and then changed little, apart from lexical improvement on immediate return to the classroom. Gradual development was found for accuracy and complexity, with important contextual influences at play. While offering some consistencies with previous research, these
findings additionally begin to challenge traditional narratives that SA is most beneficial for oral fluency and lexis over accuracy. Our focus on more advanced learners, with greater amounts of prior classroom study before SA, and over a longer period of time including the return to study post-SA, provides some explanation for these findings. A further lesson is the importance of supporting SA participants upon their return, so that the linguistic gains made while time abroad can be retained or further developed with instruction and opportunities for informal language practice.

REFERENCES


### Tables

**Table 1.**
Within-group ES (Cohen’s $d$) for fluency and lexis

<table>
<thead>
<tr>
<th></th>
<th>Fluency</th>
<th></th>
<th>Lexis</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Speech rate</td>
<td>Mean length of run</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>French</td>
<td>Spanish</td>
<td>French</td>
<td>Spanish</td>
</tr>
<tr>
<td>Presojourn–Insojourn1</td>
<td>1.20</td>
<td>1.35</td>
<td>1.31</td>
<td>.89</td>
</tr>
<tr>
<td>Presojourn–Insojourn2</td>
<td>1.01</td>
<td>1.59</td>
<td>1.20</td>
<td>1.25</td>
</tr>
<tr>
<td>Presojourn–Insojourn3</td>
<td>1.31</td>
<td>1.57</td>
<td>1.37</td>
<td>1.32</td>
</tr>
<tr>
<td>Presojourn–Postsojourn1</td>
<td>1.52</td>
<td>1.81</td>
<td>1.64</td>
<td>1.40</td>
</tr>
<tr>
<td>Presojourn–Postsojourn2</td>
<td>1.06</td>
<td>1.43</td>
<td>1.16</td>
<td>1.15</td>
</tr>
<tr>
<td>Insojourn1–Insojourn2</td>
<td>-.38*</td>
<td>.30*</td>
<td>-.33*</td>
<td>.55*</td>
</tr>
<tr>
<td>Insojourn1–Insojourn3</td>
<td>-.05*</td>
<td>.23*</td>
<td>-.16*</td>
<td>.54*</td>
</tr>
<tr>
<td>Insojourn1–Postsojourn1</td>
<td>.10*</td>
<td>.47*</td>
<td>.14*</td>
<td>.68</td>
</tr>
<tr>
<td>Insojourn1–Postsojourn2</td>
<td>-.30*</td>
<td>.15*</td>
<td>-.35*</td>
<td>.39*</td>
</tr>
<tr>
<td>Insojourn2–Insojourn3</td>
<td>.37*</td>
<td>-.07*</td>
<td>.20*</td>
<td>-.05*</td>
</tr>
<tr>
<td>Insojourn2–Postsojourn1</td>
<td>.56</td>
<td>.16*</td>
<td>.52*</td>
<td>.10*</td>
</tr>
<tr>
<td>Insojourn2–Postsojourn2</td>
<td>.09*</td>
<td>-.14*</td>
<td>-.02</td>
<td>-.17*</td>
</tr>
<tr>
<td>Insojourn3–Postsojourn1</td>
<td>.17*</td>
<td>.24*</td>
<td>.33</td>
<td>.16*</td>
</tr>
<tr>
<td>Insojourn3–Postsojourn2</td>
<td>-.28*</td>
<td>-.07*</td>
<td>-.22*</td>
<td>-.14*</td>
</tr>
<tr>
<td>Postsojourn1–Postsojourn2</td>
<td>-.46*</td>
<td>-.30*</td>
<td>-.54</td>
<td>-.29*</td>
</tr>
</tbody>
</table>

*Note:* $* =$ CIs for $d$ include zero (see supplementary data for full reporting)
Table 2.  
Between-group ES (Cohen’s $d$) for accuracy, syntactic complexity, lexis, and fluency

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Syntactic complexity</th>
<th>Lexis</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%EF-C $d$</td>
<td>%EF-ASU $d$</td>
<td>ML-ASU $d$</td>
<td>Clauses/ASU $d$</td>
</tr>
<tr>
<td>Presojourn</td>
<td>-.77</td>
<td>-.39*</td>
<td>-.48*</td>
<td>-.34*</td>
</tr>
<tr>
<td>Insojourn1</td>
<td>-.56</td>
<td>-.34*</td>
<td>-.81</td>
<td>.15*</td>
</tr>
<tr>
<td>Insojourn2</td>
<td>-.22*</td>
<td>-.33*</td>
<td>-.26*</td>
<td>-.26*</td>
</tr>
<tr>
<td>Insojourn3</td>
<td>-.69</td>
<td>-.52*</td>
<td>-.36*</td>
<td>-.36*</td>
</tr>
<tr>
<td>Postsojourn1</td>
<td>-.02*</td>
<td>-.10*</td>
<td>-1.17</td>
<td>-1.17</td>
</tr>
<tr>
<td>Postsojourn2</td>
<td>13*</td>
<td>.11*</td>
<td>-.71</td>
<td>-.71</td>
</tr>
</tbody>
</table>

Note. * = CIs for $d$ include zero (see supplementary data, Table S7, for full reporting).
Table 3.
Within-group ES (Cohen’s $d$) for accuracy and syntactic complexity

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th></th>
<th>Syntactic complexity</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% EF-C</td>
<td>% EF-ASU</td>
<td>Clauses/ASU</td>
<td>Mean length of ASU</td>
</tr>
<tr>
<td></td>
<td>French</td>
<td>Spanish</td>
<td>French</td>
<td>Spanish</td>
</tr>
<tr>
<td>Presojourn–Insojourn1</td>
<td>.83</td>
<td>1.37</td>
<td>.68</td>
<td>1.35</td>
</tr>
<tr>
<td>Presojourn–Insojourn2</td>
<td>.75</td>
<td>1.57</td>
<td>.35*</td>
<td>1.12</td>
</tr>
<tr>
<td>Presojourn–Insojourn3</td>
<td>.90</td>
<td>1.20</td>
<td>.05*</td>
<td>.31*</td>
</tr>
<tr>
<td>Presojourn–Postsojourn1</td>
<td>.95</td>
<td>2.04</td>
<td>.52*</td>
<td>1.44</td>
</tr>
<tr>
<td>Presojourn–Postsojourn2</td>
<td>.90</td>
<td>2.14</td>
<td>.05*</td>
<td>.47*</td>
</tr>
<tr>
<td>Insojourn1–Insojourn2</td>
<td>-.03*</td>
<td>.29*</td>
<td>-.31*</td>
<td>.05*</td>
</tr>
<tr>
<td>Insojourn1–Insojourn3</td>
<td>.18*</td>
<td>-.08*</td>
<td>-.59</td>
<td>-.89</td>
</tr>
<tr>
<td>Insojourn1–Postsojourn1</td>
<td>.17*</td>
<td>.77</td>
<td>-.08*</td>
<td>.17*</td>
</tr>
<tr>
<td>Insojourn1–Postsojourn2</td>
<td>.14*</td>
<td>.90</td>
<td>-.59</td>
<td>-.89</td>
</tr>
<tr>
<td>Insojourn2–Insojourn3</td>
<td>-.19*</td>
<td>-.34*</td>
<td>-.28*</td>
<td>-.78</td>
</tr>
<tr>
<td>Insojourn2–Postsojourn1</td>
<td>.19*</td>
<td>.46*</td>
<td>-.20*</td>
<td>.09*</td>
</tr>
<tr>
<td>Insojourn2–Postsojourn2</td>
<td>.15*</td>
<td>.58*</td>
<td>-.25*</td>
<td>-.75</td>
</tr>
<tr>
<td>Insojourn3–Postsojourn1</td>
<td>-.02*</td>
<td>.79</td>
<td>.45*</td>
<td>1.00</td>
</tr>
<tr>
<td>Insojourn3–Postsojourn2</td>
<td>-.05*</td>
<td>-.10*</td>
<td>1.00</td>
<td>1.64</td>
</tr>
<tr>
<td>Postsojourn1–Postsojourn2</td>
<td>-.03*</td>
<td>.13*</td>
<td>-.42*</td>
<td>-1.00</td>
</tr>
</tbody>
</table>

*Note.* $* =$ CIs for $d$ include zero (see supplementary data for full reporting).
Table 4.

Significant correlations for CAFL at Presojourn (see supplementary data for full reporting)

<table>
<thead>
<tr>
<th>Measures</th>
<th>$r$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fluency and lexis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SR + D</td>
<td>.483</td>
<td>.000</td>
</tr>
<tr>
<td>MLR + D</td>
<td>.457</td>
<td>.001</td>
</tr>
<tr>
<td><strong>Fluency and accuracy</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SR + %EF-C</td>
<td>.320</td>
<td>.019</td>
</tr>
<tr>
<td>SR + %EF-ASU</td>
<td>.311</td>
<td>.024</td>
</tr>
<tr>
<td>MLR + %EF-C</td>
<td>.298</td>
<td>.036</td>
</tr>
<tr>
<td>MLR + %EF-ASU</td>
<td>.273</td>
<td>.055</td>
</tr>
<tr>
<td><strong>Syntactic complexity and lexis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clauses/ASU + D</td>
<td>.304</td>
<td>.032</td>
</tr>
</tbody>
</table>
Table 5.

Correlations between fluency and lexis over time (see supplementary data for full reporting)

<table>
<thead>
<tr>
<th>Measures</th>
<th>Data point</th>
<th>$r$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR and D</td>
<td>Presojourn</td>
<td>.483</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>Insojourn1</td>
<td>.498</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>Insojourn2</td>
<td>.406</td>
<td>.003</td>
</tr>
<tr>
<td></td>
<td>Insojourn3</td>
<td>.379</td>
<td>.007</td>
</tr>
<tr>
<td></td>
<td>Postsojourn1</td>
<td>.403</td>
<td>.004</td>
</tr>
<tr>
<td></td>
<td>Postsojourn2</td>
<td>.513</td>
<td>.000</td>
</tr>
<tr>
<td>MLR and D</td>
<td>Presojourn</td>
<td>.457</td>
<td>.001</td>
</tr>
<tr>
<td></td>
<td>Insojourn1</td>
<td>.270</td>
<td>.069</td>
</tr>
<tr>
<td></td>
<td>Insojourn2</td>
<td>.265</td>
<td>.063</td>
</tr>
<tr>
<td></td>
<td>Insojourn3</td>
<td>.407</td>
<td>.004</td>
</tr>
<tr>
<td></td>
<td>Postsojourn1</td>
<td>.233</td>
<td>.104</td>
</tr>
<tr>
<td></td>
<td>Postsojourn2</td>
<td>.447</td>
<td>.001</td>
</tr>
</tbody>
</table>
Figures

Figure 1. Fluency development over time

![Figure 1: Fluency development over time](image-url)
Figure 2. Lexical development over time
Figure 3. Accuracy development over time
Figure 4. Syntactic complexity development over time
Figure 5. Within-group effect size (Cohen’s $d$) changes using Presojourn scores as baseline

1 Because ML-ASU could potentially include a confound due to subordination, a reviewer recommended that we additionally calculate words per clause. Since these new calculations did not change the patterning of results, we report ML-ASU in the main text and have provided descriptives for Words/Clause in Supplementary Data.