

# A Longitudinal Study of Advanced Learners' Linguistic Development Before, During, and After Study Abroad

---

<sup>1,\*</sup>KEVIN MCMANUS, <sup>2</sup>ROSAMOND MITCHELL, AND

<sup>3</sup>NICOLE TRACY-VENTURA

<sup>1</sup>Pennsylvania State University, <sup>2</sup>University of Southampton and <sup>3</sup>West Virginia University

\*E-mail: [kmcmanus@psu.edu](mailto:kmcmanus@psu.edu)

## INTRODUCTION

Study abroad (SA) research that documents learners' linguistic abilities before, during, and after SA makes an important contribution to understanding how context can shape L2 development (Pérez-Vidal 2014, 2017; Mitchell *et al.* 2017). For example, while SA appears to benefit oral fluency and vocabulary skills to a greater extent than at-home instructed contexts (e.g. Segalowitz and Freed 2004; Pizziconi 2017), the reverse has often been found for accuracy and syntactic complexity skills (e.g. Isabelli 2010; DeKeyser 2017). 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-García *et al.* 2018 and Taguchi and Collentine, 2018). In

contrast to some theories of L2 learning that prioritize meaningful language use and interaction (e.g. Duff 2019) or linguistic knowledge (e.g. Rothman and Slabakova 2018), Skill Acquisition Theory (SAT; 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 programmes with little prior language instruction (Dietrich 2018; Sanz and Morales-Front 2018). These learners tend to have a 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 learning. 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 and Pérez-Vidal 2014, 2017). Secondly, 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 (complexity, accuracy, fluency, lexis) framework to conceptualize L2 development (Housen *et al.*, 2012). While CAFL is primarily

a descriptive framework, its findings can usefully be interpreted with reference to relevant theoretical accounts of L2 development, including SAT (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 SA

Research broadly confirms the widespread belief that SA benefits oral fluency (Segalowitz and Freed 2004; Mora and Valls-Ferrer 2012; Juan-Garau 2018). Pre- and post-test SA comparisons of speed fluency (e.g. speech rate (SR)) 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- and three-month SA programmes: an extra month abroad appeared to afford no additional speed fluency benefits (see also Lara *et al.* 2015; Huensch and Tracy-Ventura 2017). 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 (Mora and Valls-Ferrer 2012; Leonard and Shea 2017; Juan-Garau 2018).

SA has also been shown to improve learners' vocabulary knowledge (Serrano *et al.* 2012; Briggs 2015; Tracy-Ventura *et al.* 2016; Zaytseva *et al.* 2018). Some research suggests that length of SA appears important in explaining the size and diversity of learners' vocabularies. For example, Ife *et al.* (2000) showed better performance on word association and translation tests for learners who studied abroad for two semesters rather than one (see also Serrano *et al.* 2011, 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 2014). 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 and Schwieter 2018 for review). Recent studies with more proficient learners, however, indicate accuracy improvement (Rees and Klapper 2007; Serrano

*et al.* 2012; Polat and Kim 2014; Leonard and Shea, 2017), 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 (Mora and Valls-Ferrer 2012; Pérez-Vidal and Barquin 2014; Lara *et al.* 2015), Leonard and Shea's (2017) experienced participants (with six 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 and Shea 2017) and (ii) after a three-month stay abroad (Mora and 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 of 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 and Schwieter 2018), our evidence base about linguistic development resulting from SA is primarily based on low-proficiency North Americans taking part in short-term SA programmes (although longstanding exceptions exist, especially involving European learners, e.g. Pérez-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 and 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 (Robinson 2005; Skehan 2009; Jackson and Suethanapornkul 2013; Michel 2017, but see Leonard and 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/context of language use potentially influence performance. Anchoring these understandings about CAFL relationships within 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 et al. 1984).

Applied to SA, the informal interactions typically afforded during SA are largely, though not exclusively, communicatively driven and meaning-focused (see Kinginger and 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 base 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 that 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 hypothesized 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 hypothesized 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, an L1 Spanish speaker, and an L1 Finnish speaker. The mean age was 21 years (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', 6 participants spent the academic year as workplace interns, 15 as teaching assistants, and 8 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 2 L1 Polish speakers. The mean age was 20.5 years (range = 20–25) and mean length of previous Spanish study was 5.9 years (range = 2–14), that is, shorter than the mean length of previous study for the French group. Two participants were workplace interns, 16 were teaching assistants, and 9 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 that 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 were narrated by L1 speakers and then compared on two oral fluency measures (see ‘Linguistic analyses’ for operationalization). These comparisons indicated no reliable differences (all 95% confidence intervals (CIs) for Cohen’s  $d$  passed through zero): Mean length of run (Cat vs. Sisters,  $d$  [CI] =  $-0.20$  [ $-1.07, 0.69$ ], Cat vs. Brothers,  $d$  [CI] =  $0.20$  [ $-0.69, 1.07$ ], Sisters vs. Brothers,  $d = 0.42$  [ $-0.48, 1.29$ ]) and speech rate (Cat vs. Sisters,  $d$  [CI] =  $-0.32$  [ $-1.19, 0.57$ ], Cat vs. Brothers,  $d$  [CI] =  $-0.03$  [ $-0.91, 0.85$ ], Sisters vs. Brothers,  $d = 0.31$  [ $-0.58, 1.18$ ]). As previously noted, each story was administered approximately 12 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 min) and could ask clarification questions before retelling the story with continuing access to the pictures. The mean narration time was 3:47 min.

## 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: 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 (i) speech rate (SR), the number of syllables/total speaking time (in seconds), including pauses and (ii) mean length of run (MLR), the mean number of syllables between silent pauses of 0.25 s or greater (Towell *et al.* 1996; Freed *et al.* 2004). PRAAT (Boersma & Weenink 2015) was used to identify and measure pause length. The research team created protocols for the manual counting of syllables in ELAN (Sloetjes and Wittenburg 2008). 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 and Muñoz 2010).

### Lexis

Lexis was operationalized as lexical diversity, computed using the VocD command in CLAN (MacWhinney 2000), resulting in a score D (Malvern and Richards 2002). D is computed by taking random samples of a text, calculating type-token ratios of those samples, and then using those type-token ratios to create a random-sampling type-token ratio for the text (see Malvern and Richards 2002). D scores tend to range from 10 to 100; the higher the score, the greater the 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, and 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).<sup>1</sup> A clause was defined following Bulté and Housen (2012: 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 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 data sets were normally distributed (according to Shapiro–Wilk tests, all data sets  $p > 0.05$ ), we present the results of parametric tests.

First, a two-way  $2 \times 6$  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 0.05.

Secondly, we calculated Pearson product-moment correlations with bootstrapped 95% 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 95% CIs for  $d$  for all paired comparisons (within-subject over time and between-subject at each data point) and  $r$  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 0.60 are considered small, 1.00 medium, and 1.40 large; between-subject: *ds* around 0.40 are considered small, 0.70 medium, and 1.00 large; and *r* field-specific benchmarks: *rs* close to 0.25 are considered small, 0.40 medium, and 0.60 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 1 and 2) and then between-group comparisons, using Cohen’s *d* ES (see Table 3) (for CIs for *d*, see Supplementary Data). Correlations are then presented between the different CAFL elements (See Tables 4 and 5 and Supplementary Data). Descriptives for all measures are presented in Supplementary Data.

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

	Fluency				Lexis	
	SR		MLR		D	
	French	Spanish	French	Spanish	French	Spanish
Presojourn–Insojourn1	1.20	1.35	1.31	.89	1.57	2.77
Presojourn–Insojourn2	1.01	1.59	1.20	1.25	0.98	1.76
Presojourn–Insojourn3	1.31	1.57	1.37	1.32	0.75	1.26
Presojourn–Postsojourn1	1.52	1.81	1.64	1.40	1.77	3.34
Presojourn–Postsojourn2	1.06	1.43	1.16	1.15	0.81	1.71
Insojourn1–Insojourn2	–0.38 <sup>a</sup>	0.30 <sup>a</sup>	–0.33 <sup>a</sup>	0.55 <sup>a</sup>	–0.70	–0.70
Insojourn1–Insojourn3	–0.05 <sup>a</sup>	0.23 <sup>a</sup>	–0.16 <sup>a</sup>	0.54 <sup>a</sup>	–0.75	–1.35
Insojourn1–Postsojourn1	0.10 <sup>a</sup>	0.47 <sup>a</sup>	0.14 <sup>a</sup>	0.68	0.54 <sup>a</sup>	0.16 <sup>a</sup>
Insojourn1–Postsojourn2	–0.30 <sup>a</sup>	0.15 <sup>a</sup>	–0.35 <sup>a</sup>	0.39 <sup>a</sup>	–0.72	–0.74
Insojourn2–Insojourn3	0.37 <sup>a</sup>	–0.07 <sup>a</sup>	0.20 <sup>a</sup>	–0.05 <sup>a</sup>	–0.13 <sup>a</sup>	–0.54 <sup>a</sup>
Insojourn2–Postsojourn1	0.56	0.16 <sup>a</sup>	0.52 <sup>a</sup>	0.10 <sup>a</sup>	1.10	0.92
Insojourn2–Postsojourn2	0.09 <sup>a</sup>	–0.14 <sup>a</sup>	–0.02	–0.17 <sup>a</sup>	–0.08 <sup>a</sup>	–0.04 <sup>a</sup>
Insojourn3–Postsojourn1	0.17 <sup>a</sup>	0.24 <sup>a</sup>	0.33	0.16 <sup>a</sup>	1.12	1.67
Insojourn3–Postsojourn2	–0.28 <sup>a</sup>	–0.07 <sup>a</sup>	–0.22 <sup>a</sup>	–0.14 <sup>a</sup>	0.04 <sup>a</sup>	0.51 <sup>a</sup>
Postsojourn1–Postsojourn2	–0.46 <sup>a</sup>	–0.30 <sup>a</sup>	–0.54	–0.29 <sup>a</sup>	–1.10	–0.96

<sup>a</sup>CIs for *d* include zero (see Supplementary Data for full reporting).

Table 2: Within-group ES (Cohen’s *d*) for accuracy and syntactic complexity

	Accuracy				Syntactic complexity			
	% EF-C		% EF-ASU		Clauses/ASU		Mean length of ASU	
	French	Spanish	French	Spanish	French	Spanish	French	Spanish
Presojourn–Insojourn1	0.83	1.37	0.84	1.14	0.68	1.35	0.74	0.36 <sup>a</sup>
Presojourn–Insojourn2	0.75	1.57	0.87	1.21	0.35 <sup>a</sup>	1.12	0.40 <sup>a</sup>	0.57
Presojourn–Insojourn3	0.90	1.20	1.03	1.07	0.05 <sup>a</sup>	0.31 <sup>a</sup>	−0.67	−0.41 <sup>a</sup>
Presojourn–Postsojourn1	0.95	2.04	0.96	1.58	0.52 <sup>a</sup>	1.44	1.06	0.48 <sup>a</sup>
Presojourn–Postsojourn2	0.90	2.14	1.06	1.94	0.05 <sup>a</sup>	0.47 <sup>a</sup>	0.41 <sup>a</sup>	0.23 <sup>a</sup>
Insojourn1–Insojourn2	−0.03 <sup>a</sup>	0.29 <sup>a</sup>	0.16 <sup>a</sup>	0.11 <sup>a</sup>	−0.31 <sup>a</sup>	0.05 <sup>a</sup>	−0.55	0.28 <sup>a</sup>
Insojourn1–Insojourn3	0.18 <sup>a</sup>	−0.08 <sup>a</sup>	0.35 <sup>a</sup>	0.05 <sup>a</sup>	−0.59	−0.89	−1.05	−0.92
Insojourn1–Postsojourn1	0.17 <sup>a</sup>	0.77	0.25 <sup>a</sup>	0.48 <sup>a</sup>	−0.08 <sup>a</sup>	0.17 <sup>a</sup>	−0.14 <sup>a</sup>	0.13 <sup>a</sup>
Insojourn1–Postsojourn2	0.14 <sup>a</sup>	0.90	0.34 <sup>a</sup>	0.84	−0.54	−0.88	−0.55	−0.15 <sup>a</sup>
Insojourn2–Insojourn3	0.19 <sup>a</sup>	−0.34 <sup>a</sup>	0.16 <sup>a</sup>	−0.04 <sup>a</sup>	−0.28 <sup>a</sup>	−0.78	−1.07	−1.09
Insojourn2–Postsojourn1	0.19 <sup>a</sup>	0.46 <sup>a</sup>	0.07 <sup>a</sup>	0.36 <sup>a</sup>	0.20 <sup>a</sup>	0.09 <sup>a</sup>	0.72	−0.18 <sup>a</sup>
Insojourn2–Postsojourn2	0.15 <sup>a</sup>	0.58 <sup>a</sup>	0.14 <sup>a</sup>	0.70	−0.25 <sup>a</sup>	−0.75	0.01 <sup>a</sup>	−0.41
Insojourn3–Postsojourn1	−0.02 <sup>a</sup>	0.79	−0.10 <sup>a</sup>	0.38 <sup>a</sup>	0.45 <sup>a</sup>	1.00	1.64	1.11
Insojourn3–Postsojourn2	−0.05 <sup>a</sup>	0.90	−0.03 <sup>a</sup>	0.68	0.00 <sup>a</sup>	0.10 <sup>a</sup>	1.07	0.75
Postsojourn1–Postsojourn2	−0.03 <sup>a</sup>	0.13 <sup>a</sup>	0.08 <sup>a</sup>	0.33 <sup>a</sup>	−0.42 <sup>a</sup>	−1.00	−0.70	−0.29 <sup>a</sup>

<sup>a</sup>CIs for *d* include zero (see [Supplementary Data](#) for full reporting).

Table 3: Between-group ES (Cohen’s *d*) for accuracy, syntactic complexity, lexis, and fluency

	Accuracy		Syntactic complexity		Lexis	Fluency	
	%EF-C <i>d</i>	%EF-ASU <i>d</i>	ML-ASU <i>d</i>	Clauses/ASU <i>d</i>	D <i>d</i>	SR <i>d</i>	MLR <i>d</i>
Presojourn	−0.77	−0.39 <sup>a</sup>	−0.48 <sup>a</sup>	−0.34 <sup>a</sup>	0.84	0.16 <sup>a</sup>	0.34 <sup>a</sup>
Insojourn1	−0.56	−0.34 <sup>a</sup>	−0.81	0.15 <sup>a</sup>	1.60	0.27 <sup>a</sup>	−0.12 <sup>a</sup>
Insojourn2	−0.22 <sup>a</sup>	−0.33 <sup>a</sup>	−0.26 <sup>a</sup>	−0.26 <sup>a</sup>	1.56	0.99	0.75
Insojourn3	−0.69	−0.52 <sup>a</sup>	−0.36 <sup>a</sup>	−0.36 <sup>a</sup>	1.07	0.59	0.60
Postsojourn1	−0.02 <sup>a</sup>	−0.10 <sup>a</sup>	−1.17	−1.17	0.81	0.70	0.46 <sup>a</sup>
Postsojourn2	13 <sup>a</sup>	0.11 <sup>a</sup>	−0.71	−0.71	1.47	0.74	0.61

<sup>a</sup>CIs for *d* include zero (see [Supplementary Data, Table S7](#), for full reporting).

Table 4: Significant correlations for CAFL at Presojourn

Measures	<i>r</i>	<i>p</i>
<i>Fluency and lexis</i>		
SR + D	0.483	0.000
MLR + D	0.457	0.001
<i>Fluency and accuracy</i>		
SR + %EF-C	0.320	0.019
SR + %EF-ASU	0.311	0.024
MLR + %EF-C	0.298	0.036
MLR + %EF-ASU	0.273	0.055
<i>Syntactic complexity and lexis</i>		
Clauses/ASU + D	0.304	0.032

Note: See [Supplementary Data](#) for full reporting.

Fluency

[Figure 1](#) shows the distribution of learners’ SR and MLR scores over time. For SR, the statistically significant Time × Group interaction ( $F(5, 44) = 5.28, p < 0.001$ ) 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

Table 5: Correlations between fluency and lexis over time

Measures	Data point	<i>r</i>	<i>p</i>
SR and D	Presojourn	0.483	0.000
	Insojourn1	0.498	0.000
	Insojourn2	0.406	0.003
	Insojourn3	0.379	0.007
	Postsojourn1	0.403	0.004
	Postsojourn2	0.513	0.000
MLR and D	Presojourn	0.457	0.001
	Insojourn1	0.270	0.069
	Insojourn2	0.265	0.063
	Insojourn3	0.407	0.004
	Postsojourn1	0.233	0.104
	Postsojourn2	0.447	0.001

Note: See [Supplementary Data](#) for full reporting.

for Time ( $F(5, 44) = 76.23, p < 0.001$ ) 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  $\times$  Group interaction ( $F(5, 41) = 39.82, p < 0.001$ ) 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 ( $F(5, 41) = 39.82, p < 0.001$ ) 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.

Lexis

[Figure 2](#) shows the distribution of lexical diversity (D) scores over time. The (borderline) statistically significant Time  $\times$  Group interaction ( $F(5, 47) = 2.40, p = 0.051$ ) 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 ( $F(5, 47) = 76.04, p < 0.001$ ) because lexical diversity scores at all other time points were higher than at Presojourn (medium-to-large ES for both groups). [Table 4](#) indicates significant improvement in lexical



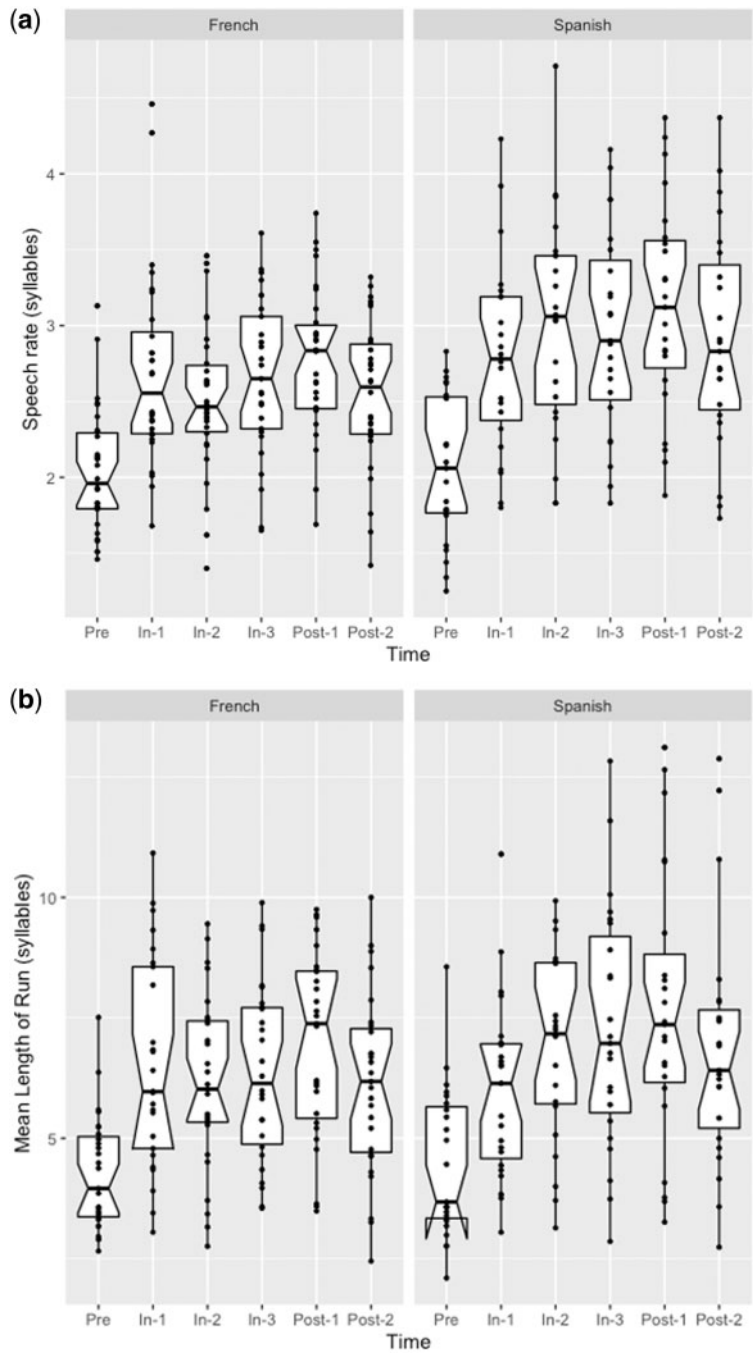


Figure 1: Fluency development over time. (a) Changes in speech rate over time; (b) changes in mean length of run over time.

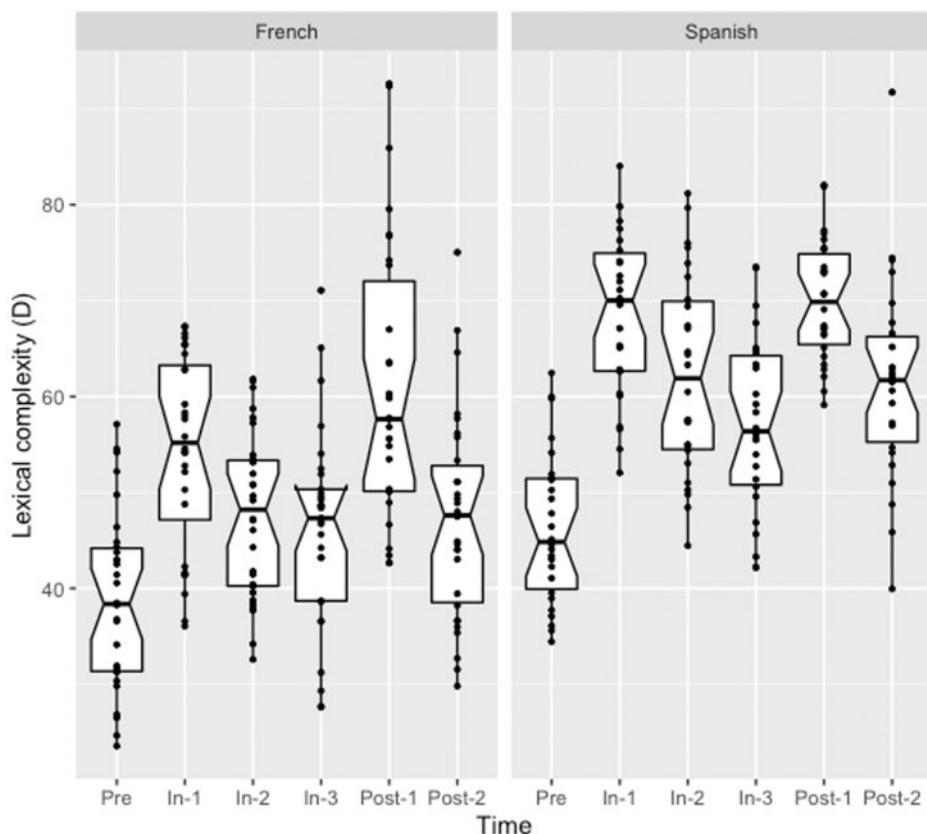


Figure 2: Lexical development over time.

diversity scores following the move abroad (at Insojourn1), followed by further movement on immediate return to the instructed context (Postsojourn1).

## 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  $\times$  Group interaction ( $F(5, 48) = 4.31$ ,  $p = 0.003$ ) 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 ( $F(5, 48) = 19.79$ ,  $p < 0.001$ ) because %EF-C scores at all other test phases were more accurate than at Presojourn ( $p < 0.001$ ). 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

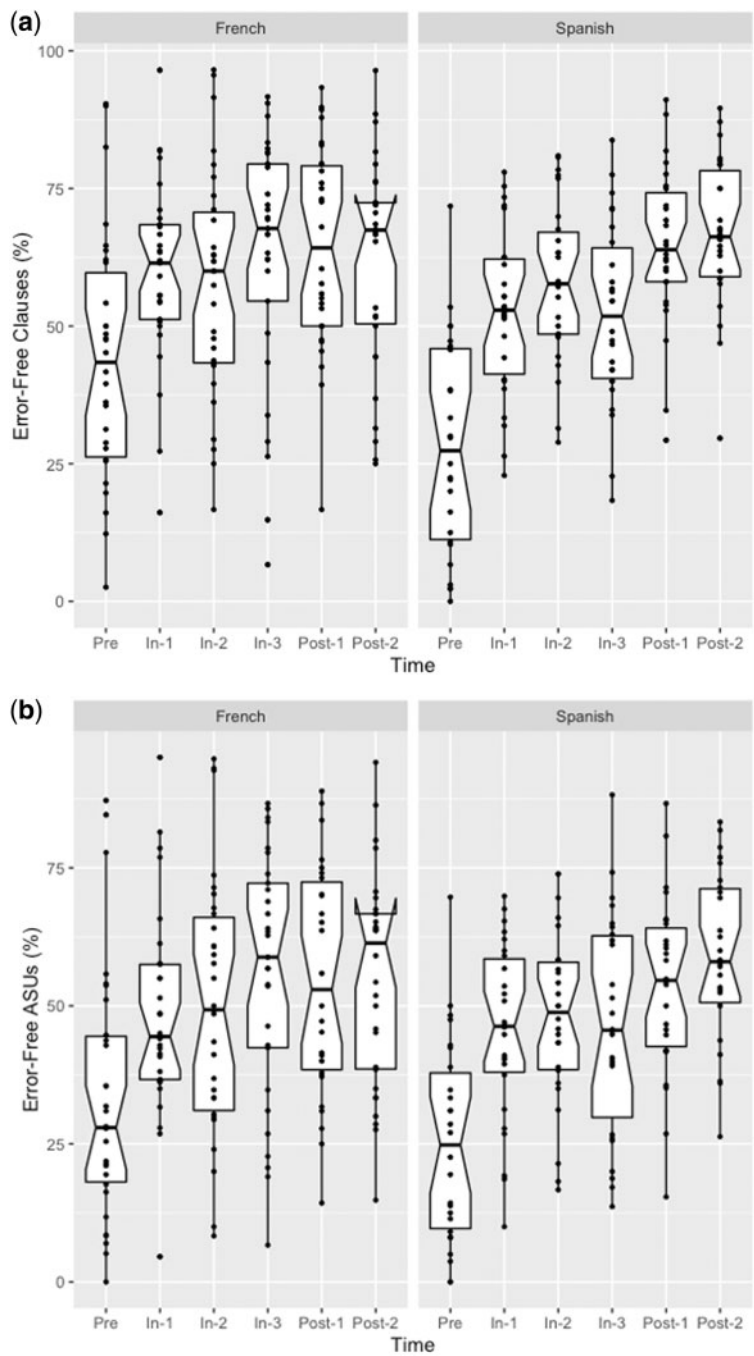


Figure 3: Accuracy development over time. (a) Changes in error-free clauses over time; (b) changes in error-free ASUs over time.

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 ( $F(5, 48) = 17.93, p < 0.001$ ) because of less accurate performance at Presojourn for both groups of learners (small ES). Both groups made little further progress.

## 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 ( $F(5, 45) = 4.29, p = 0.003$ ) 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 ( $F(5, 45) = 18.93, p < 0.001$ ), 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  $\times$  Group interaction was not statistically significant ( $F(5, 48) = 2.02, p = 0.092$ ). Within-group comparisons for performance over time was statistically significant ( $F(5, 48) = 12.19, p < 0.00$ ), 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).

## 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 1 and 2), these plots show performance over time with Presojourn as the baseline. Data points occurring within the red boxes are interpreted as a 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.

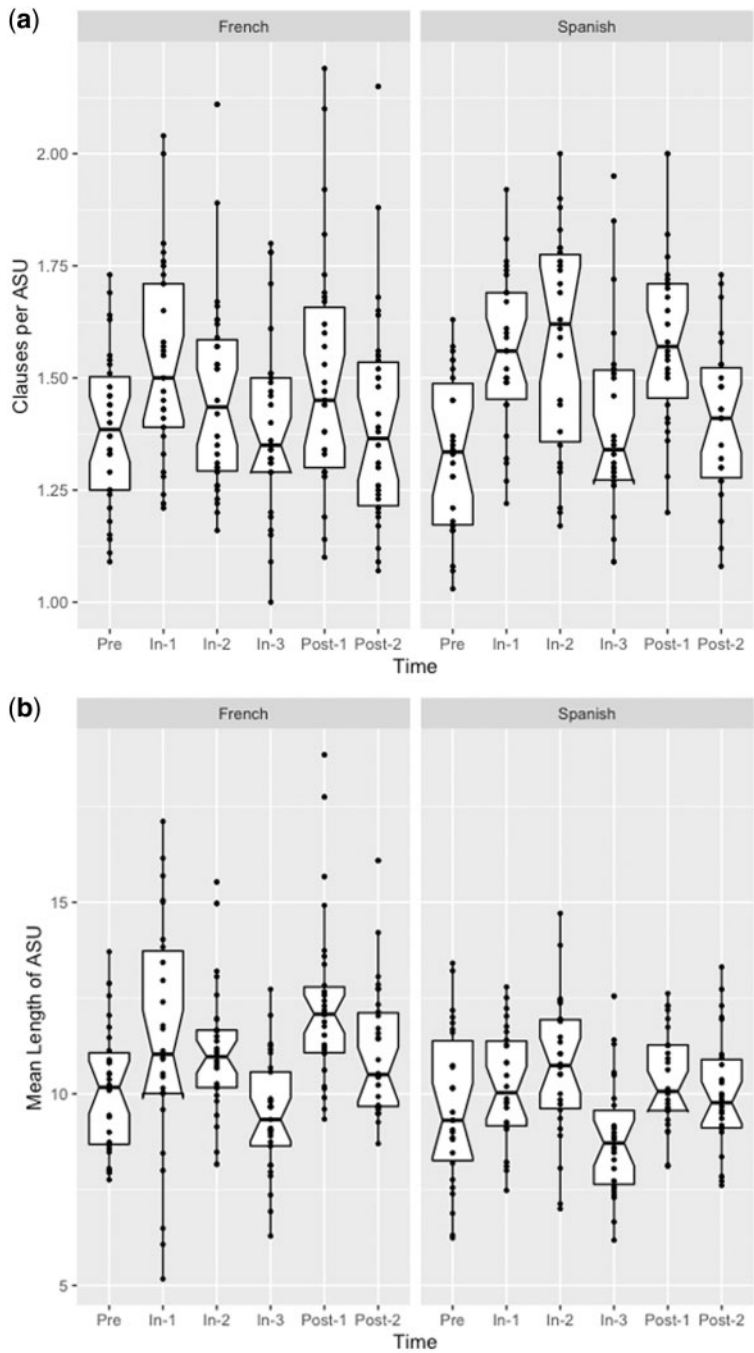


Figure 4: Syntactic complexity development over time. (a) Changes in clauses per ASU over time; (b) changes in mean length of ASU over time.

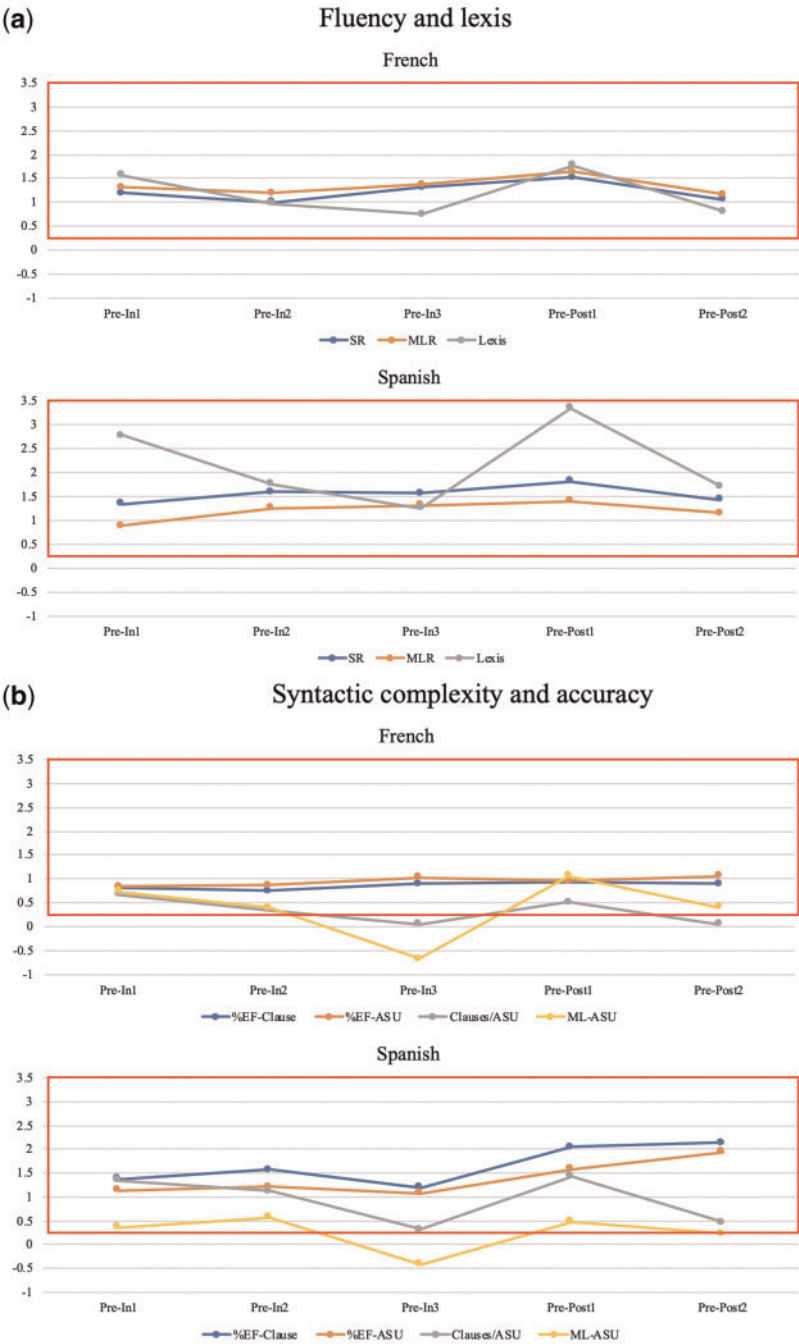


Figure 5: Within-group effect size (Cohen's  $d$ ) changes using Presojourn scores as baseline.



## CAFL relationships over time

The nature of the CAFL relationships at each data point was 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).

At Insojourn1, we found positive correlations between lexis and fluency only: a medium correlation with SR ( $r = 0.498$ ,  $p < 0.001$ ) and a borderline small correlation with MLR ( $r = 0.270$ ,  $p = 0.069$ ). 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 ( $r = -0.436$ ,  $p = 0.002$ ) and fluency (SR,  $r = -0.361$ ,  $p = 0.011$ ).

No other meaningful accuracy correlations were found until Postsojourn1, when small-to-medium positive correlations re-emerged between accuracy and fluency (%EF-C + SR:  $r = 0.379$ ,  $p = 0.005$ ; %EF-C + MLR:  $r = 0.397$ ,  $p = 0.004$ ; %EF-ASU + SR:  $r = 0.352$ ,  $p = 0.010$ ; %EF-ASU + MLR:  $r = 0.324$ ,  $p = 0.022$ ), and, at Postsojourn-2, borderline and small positive correlations emerged with lexis (%EF-C + D:  $r = 0.253$ ,  $p = 0.076$ ; %EF-ASU + D:  $r = 0.268$ ,  $p = 0.060$ ).

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 ( $r = 0.595$ ,  $p < 0.001$ ) and Postsojourn1 ( $r = 0.343$ ,  $p = 0.015$ ); and small-to-medium negative correlations with ML-ASU at Postsojourn1 ( $r = -0.291$ ,  $p = 0.040$ ) and Postsojourn2 ( $r = -0.368$ ,  $p = 0.009$ ). Syntactic complexity also correlated positively at Postsojourn1 with fluency (Clauses/ASU + SR,  $r = 0.304$ ,  $p = 0.027$ ) and at Postsojourn2 with accuracy (Clauses/ASU + %EF-C,  $r = 0.313$ ,  $p = 0.019$ ).

## 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 SAT 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. accuracy), and to Spanish participants' somewhat greater engagement in local social networks and higher levels of L2 use Insojourn (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 (Segalowitz and Freed 2004; Llanes and Serrano 2011; Mora and Valls-Ferrer 2012; Pérez-Vidal 2014; Briggs 2015; Juan-Garau 2018). ES magnitudes of improvement between Presojourn–Insojourn1 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). The 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 given evidence of improvement following the move from the instructed context to SA and continuing gradual improvement during the nine-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 2014*). 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 [Kinging and Blattner 2008](#)). 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 and 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.

## SUPPLEMENTARY DATA

[Supplementary material](#) is available at *Applied Linguistics* online.

## NOTE

- 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](#).

## ACKNOWLEDGEMENTS

We are very grateful to the LANGSNAP participants for their time and contributions to this study, and to Laurence Richard, Patricia Romero de Mills, and Amanda Huensch. We would also like to thank *Applied Linguistics'* editors Christina Higgins and John Hellermann and the valuable feedback from three anonymous reviewers.

## FUNDING

UK Economic and Social Research Council (Grant number: RES-062-23-2996).

## REFERENCES

- Boersma, P. and Weenink, D.** 2015. Praat: Doing phonetics by computer (Version 5.4.18) [Computer program].
- Briggs, J.** 2015. 'Out-of-class language contact and vocabulary gain in a study abroad context,' *System* 53: 129–40.

- British Academy.** 2011. 'Position statement: Language matters more and more,' available at <https://www.thebritishacademy.ac.uk/sites/default/files/LMmm%20-%20A%20Position%20Statement.pdf>. Accessed 31 January 2020.
- Bulté, B.** and **A. Housen.** 2012. 'Defining and operationalising L2 complexity' in A. Housen, F. Kuiken and I. Vedder (eds): *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. John Benjamins.
- Collentine, J.** 2009. 'Study abroad research: Findings, implications, and future directions' in M. H. Long and C. J. Doughty (eds): *The Handbook of Language Teaching*. Wiley.
- DeKeyser, R.** 1997. 'Beyond explicit rule learning: Automatizing second language morphosyntax,' *Studies in Second Language Acquisition* 19: 195–221.
- DeKeyser, R.** 2007. 'Study abroad as foreign language practice' in R. DeKeyser (ed.): *Practice in a Second Language: Perspectives from Applied Linguistics and Cognitive Psychology*. Cambridge University Press.
- DeKeyser, R.** 2010. 'Monitoring processes in Spanish as a second language during a study abroad program,' *Foreign Language Annals* 43: 80–92.
- DeKeyser, R.** 2017. 'Knowledge and skill in ISLA' in S. Loewen and M. Sato (eds): *Routledge Handbook of Instructed Second Language Acquisition*. London: Routledge.
- Dietrich, A.** 2018. 'History and current trends in US study abroad' in C. Sanz and A. Morales-Front (eds): *The Routledge Handbook of Study Abroad and Practice*. Routledge.
- Dominguez, L., N. Tracy-Ventura, M. J. Arche, R. Mitchell, and F. Myles.** 2013. 'The role of dynamic contrasts in the L2 acquisition of Spanish past tense morphology,' *Bilingualism: Language and Cognition* 16: 558–77.
- Du, H.** 2013. 'The development of Chinese fluency during study abroad in China,' *The Modern Language Journal* 97: 131–43.
- Duff, P.** 2019. 'Social dimensions and processes in second language acquisition: Multilingual socialization in transnational contexts,' *The Modern Language Journal* 103: 6–22.
- Edmonds, A.** and **A. Gudmestad.** 2018. 'Gender marking in written L2 French: Before, during, and after residence abroad,' *Study Abroad Research in Second Language Acquisition and International Education* 3: 58–83.
- Foster, P., A. Tonkyn, and G. Wigglesworth.** 2000. 'Measuring spoken language: A unit for all reasons,' *Applied Linguistics* 21: 354–75.
- Freed, B., N. Segalowitz, and D. P. Dewey.** 2004. 'Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs,' *Studies in Second Language Acquisition* 26: 275–301.
- Gilbert, R.** and **C. Muñoz.** 2010. 'Differences in attainment and performance in a foreign language: The role of working memory capacity,' *International Journal of English Studies* 10: 19–42.
- Housen, A., F. Kuiken, and I. Vedder.** 2012 (eds). *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. John Benjamins.
- Howard, M.** and **J. Schwieter.** 2018. 'The development of second language grammar in a study abroad context,' in C. Sanz and A. Morales-Front (eds): *The Routledge Handbook of Study Abroad and Practice*. Routledge.
- Huensch, A.** and **N. Tracy-Ventura.** 2017. 'Understanding L2 fluency behavior: The effects of individual differences in L1 fluency, cross-linguistic differences, and proficiency over time,' *Applied Psycholinguistics* 38: 755–85.
- Ife, A., G. Vives Boix, and P. Meara.** 2000. 'The impact of study abroad on the vocabulary development of different proficiency groups,' *Spanish Applied Linguistics* 4: 55–84.
- Isabelli-García, C.** 2010. Acquisition of Spanish gender agreement in two learning contexts: Study abroad and at home. *Foreign Language Annals* 43: 289–303.
- Isabelli-García, C., J. Bown, J. L. Plews, and D. P. Dewey.** 2018. 'Language learning and study abroad,' *Language Teaching* 514: 439–84.
- Jackson, D.** and **S. Suethanapornkul.** 2013. 'The cognition hypothesis: A synthesis and meta-analysis of research on second language task complexity,' *Language Learning* 63: 330.
- Jensen, J.** and **M. Howard.** 2014. 'The effects of time in the development of complexity and accuracy during study abroad: A study of French and Chinese learners of English,' *EUROSLA Yearbook* 14: 31–64.
- Juan Garau, M.** 2014. 'Oral accuracy growth after formal instruction and study abroad: Onset level, contact factors and long-term effects' in C. Pérez-Vidal (ed.): *Language Acquisition in Study Abroad and Formal Instruction Contexts*. John Benjamins.



- Juan-Garau, M.** 2018. 'Exploring oral L2 fluency development during a three-month stay abroad through a dialogic task' in C. Sanz and A. Morales-Front (eds): *The Routledge Handbook of Study Abroad Research and Practice*. Routledge.
- Kinginger, C.** and **G. Blattner.** 2008. 'Histories of engagement and sociolinguistic awareness in study abroad: Colloquial French' in L. Ortega and H. Byrnes (eds): *The Longitudinal Study of Advanced L2 Capacities*. Lawrence Erlbaum.
- Lara, R., J. C. Mora,** and **C. Pérez-Vidal.** 2015. 'How long is long enough? L2 English development through study abroad programmes varying in duration,' *Innovation in Language Learning and Teaching* 9: 46–57.
- Leonard, K. R.** and **C. E. Shea.** 2017. 'L2 speaking development during study abroad: Fluency, accuracy, complexity, and underlying cognitive factors,' *The Modern Language Journal* 101: 179–93.
- Llanes, À.** 2011. 'The many faces of study abroad: An update on the research on L2 gains emerged during a study abroad experience,' *International Journal of Multilingualism* 8: 189–215.
- Llanes, À.,** and **C. Muñoz.** 2009. 'A short stay abroad: Does it make a difference?,' *System* 37: 353–65.
- Llanes, À.** and **R. Serrano.** 2011. 'Length of stay and study abroad: Language gains in two versus three months abroad,' *Revista Española de Lingüística Aplicada* 24: 95–110.
- MacWhinney, B.** 2000. *The CHILDES Project: The Database* (vol. 2). Psychology Press.
- Malvern, D.** and **B. Richards.** 2002. 'Investigating accommodation in language proficiency interviews using a new measure of lexical diversity,' *Language Testing* 19: 85–104.
- McManus, K.** 2015. 'L1-L2 differences in the acquisition of form-meaning pairings: A comparison of English and German learners of French,' *Canadian Modern Language Review* 71: 51–77.
- McManus, K.** 2019. 'Relationships between social networks and language development during study abroad,' *Language, Culture, and Curriculum* 32: 270–84.
- Michel, M.** 2017. 'Complexity, accuracy and fluency (CAF),' in S. Loewen and S. Masatoshi (eds): *The Routledge Handbook of Instructed Second Language Acquisition*. Routledge.
- Mitchell, R. F., N. Tracy-Ventura,** and **K. McManus.** 2017. *Anglophone Students Abroad: Identity, Social Relationships and Language Learning*. Routledge.
- Mora, J. C.** and **M. Valls-Ferrer.** 2012. 'Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts,' *TESOL Quarterly* 46: 610–41.
- Pérez-Vidal, C.** 2014 (ed.). *Language Acquisition in Study Abroad and Formal Instruction Contexts*. John Benjamins.
- Pérez-Vidal, C.** 2017. 'Study abroad and ISLA,' in S. Loewen and S. Masatoshi (eds): *The Routledge Handbook of Instructed Second Language Acquisition*. Routledge.
- Pérez-Vidal, C.,** and **E. Barquín.** 2014. 'Comparing progress in academic writing after formal instruction and study abroad' in C. Pérez-Vidal (ed.): *Language Acquisition in Study Abroad and Formal Acquisition Contexts* (pp. 217–34). John Benjamins.
- Pizziconi, B.** 2017. 'Japanese vocabulary development in and beyond study abroad: The timing of the year abroad in a language degree curriculum,' *The Language Learning Journal* 45: 133–52.
- Plonsky, L.** and **F. L. Oswald.** 2014. 'How big is "big"? Interpreting effect sizes in L2 research,' *Language Learning* 64: 878–912.
- Polat, B.** and **Y. Kim.** 2014. 'Dynamics of complexity and accuracy: A longitudinal case study of advanced untutored development,' *Applied Linguistics* 35: 184–207.
- Rees, J.** and **J. Klapper.** 2007. 'Analysing and evaluating the linguistic benefit of residence abroad for UK foreign language students,' *Assessment & Evaluation in Higher Education* 32: 331–53.
- Ren, W.** 2018. 'Developing L2 pragmatic competence in study abroad contexts' in C. Sanz and A. Morales-Front (eds): *The Routledge Handbook of Study Abroad and Practice*. Routledge.
- Robinson, P.** 2005. 'Cognitive complexity and task sequencing: Studies in a componential framework for second language task design,' *IRAL-International Review of Applied Linguistics in Language Teaching* 43: 1–32.
- Rothman, J.** and **R. Slabakova.** 2018. 'The generative approach to SLA and its place in modern second language studies,' *Studies in Second Language Acquisition* 40: 417–42.
- Sanz, C.** and **A. Morales-Front.** 2018. 'Introduction: Issues in study abroad research and practice' in C. Sanz and A. Morales-Front (eds): *The Routledge Handbook of Study Abroad and Practice*. Routledge.
- Schmidt, R.** 1992. 'Awareness and second language acquisition,' *Annual Review of Applied Linguistics* 13: 206–26.

- Schneider, W., Dumais, D., and Schiffrin, R.** 1984. 'Automatic and control processing and attention,' in R. Parasuraman (ed.): *Varieties of Attention*. Academic Press.
- Segalowitz, N. and B. F. Freed.** 2004. 'Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts,' *Studies in Second Language Acquisition* 26: 173–99.
- Serrano, R., À. Llanes, and E. Tragant.** 2011. 'Analyzing the effect of context of second language learning: Domestic intensive and semi-intensive courses vs. study abroad in Europe,' *System* 39: 133–43.
- Serrano, R., E. Tragant, and A. Llanes.** 2012. 'A longitudinal analysis of the effects of one year abroad,' *Canadian Modern Language Review* 68: 138–63.
- Skehan, P.** 2001. 'Tasks and language performance' in M. Bygate, P. Skehan, and M. Swain (eds): *Research Pedagogic Tasks: Second Language Learning, Teaching, and Testing*. Longman.
- Skehan, P.** 2009. 'Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis,' *Applied Linguistics* 30: 510–32.
- Sloetjes, H. and P., Wittenburg** 2008. 'Annotation by category – ELAN and ISO DCR'. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).
- Taguchi, N. and Collentine, J.** 2018. 'Language learning in a study-abroad context: Research agenda,' *Language Teaching* 514: 553–66.
- Towell, R., R. Hawkins, and N. Bazergui.** 1996. 'The development of fluency in advanced learners of French,' *Applied Linguistics* 17: 84–119.
- Tracy-Ventura, N.** 2017. 'Combining corpora and experimental data to investigate language learning during residence abroad: A study of lexical sophistication,' *System* 71: 35–45.
- Tracy-Ventura, N., R. F. Mitchell, and K. McManus.** 2016. 'The LANGSNAP longitudinal corpus' in M. Alonso-Ramos (ed.): *Spanish Learner Corpus Research: Current Trends and Future Perspectives*. John Benjamins.
- Zaytseva, V., C. Pérez-Vidal, and I. Miralpeix.** 2018. 'Vocabulary acquisition during study abroad: A comprehensive review of the research' in C. Sanz and A. Morales-Front (eds): *The Routledge Handbook of Study Abroad Research and Practice*. Routledge.