Understanding the Long-Term Evolution of L2 Lexical Diversity: The Contribution of a Longitudinal Learner Corpus

Nicole Tracy-Ventura, Amanda Huensch,
and Rosamond Mitchell

1 Introduction

1.1 Goals of the Study

The central goal of Second Language Acquisition (SLA) research is understanding the learner's underlying L2 knowledge system, its development, and what impacts upon both. To address these questions, most SLA research adopts cross-sectional designs comparing groups from different proficiency levels, or relatively short-term experimental designs. From a practicality point of view, this is reasonable, but for some questions of central concern for SLA research, taking a longitudinal perspective is necessary.

Compared to SLA, learner corpus research (LCR) has more traditionally involved the collection and analysis of large datasets of second language production. Large amounts of data from numerous participants help to improve generalizability of findings, and computerized tools make analysis of such data more feasible (Granger et al. 2015). Work in LCR is well ahead in this regard, as is work in L1 acquisition, and many sophisticated computerized tools are used in these fields. In the spirit of "open science" (Marsden et al. 2015), electronic learner corpora also ease the process of data sharing. Where learner corpora are publicly available, other researchers can exploit them, thus broadening the impact of the data collected which is particularly important for scarce longitudinal data (MacWhinney 2017; Meunier 2015). As evidenced by this volume, and other recent publications (e.g., Granger 2009; Hasko 2013; Myles 2015), there is now a growing interest in ways that SLA can benefit from the work being done in LCR and vice versa.

In this chapter we describe our work on the long-term evolution of L2 lexical diversity in a group of French and Spanish L2 university

learners for whom we have been building a learner corpus since a first wave of data collection in May 2011. Our longitudinal learner corpus includes over 700,000 words (including data from L1 speakers), 88 percent of which is spoken data primarily from an oral interview. Metadata have been collected, the files are formatted using the Codes for the Human Analysis of Transcripts (CHAT: MacWhinney 2000) and have also been morpho-syntactically tagged. All of the files (audio, transcripts, and tagged files) are shared publicly on our website, and available on talkbank.org. Our newest data were collected in June 2016 and are allowing us to investigate the evolution of our participants' L2 skills four years after returning from a year abroad and three years after graduating from university.

Language proficiency is a dynamic phenomenon, subject to attrition as well as development. Understanding how second language proficiency evolves over time, and how both development and attrition relate to phases of increased and decreased input and interaction, should be of central concern to the field of SLA, increasing our understanding of the contribution of L2 use to the creation and stabilization of the underlying L2 knowledge system. However, there are few longitudinal studies which have collected information both on L2 proficiency and L2 use, and which have spanned more than one to two years. Research which has focused on the long-term evolution of bilingual speakers' language skills tends to center on first language attrition (L1: see Schmid 2011). In contrast, foreign language attrition, and variables that influence it, has received very little attention (Bardovi-Harlig & Stringer 2010; Schmid & Mehotcheva 2012; Weltens 1989). Longitudinal methods are needed to address questions of foreign language attrition as well as development and long-term retention.

In this article, we demonstrate how we are using our longitudinal corpus to investigate the possible outcomes of attrition, development, or retention in lexical diversity among advanced instructed FL learners three years after formal instruction has ended. Using data on L2 use collected systematically at the same time as the L2 production data, we explore the relations between evolving patterns of contact with L2 and these different potential proficiency outcomes.

Corpus-based methods are particularly suited for the analysis of L2 lexis. Compared to controlled and psycholinguistic tests of vocabulary knowledge, learner corpora provide more authentic samples of language, demonstrating what learners are capable of producing in real time. In previous research on L1 attrition of lexical diversity, interview data distinguished better between monolingual speakers and bilingual attriters than data elicited from more controlled tasks (Schmid & Jarvis 2014), suggesting that corpus data can contribute helpfully to

the study of lexical maintenance and attrition. And of course, corpus-based methods make it possible to conduct lexical analyses of large amounts of data with automatic or semi-automatic tools.

1.2 Long-Term Evolution of Foreign Language Proficiency

As discussed above, there are three potential outcomes in the long-term evolution of foreign language proficiency: attrition, retention, or development. Although much research has focused on instructed language learning, little is known about what happens to language abilities once learners are no longer engaged in formal instruction. Interest in L2 attrition is evident from a number of early SLA studies (e.g., Bahrick 1984a, 1984b; Weltens 1989; Weltens et al. 1989), but this line of research has not yet developed into a serious research agenda. One potential reason for this is the difficulty of establishing the highest proficiency level attained among FL learners (Schmid & Mehotcheva 2012), or what Bardovi-Harlig and Stringer (2010) refer to as 'peak attainment'. In L1 attrition research participants who did not move out of the L1 context until aged 13–15 or over are assumed to have a fully developed L1 (and previous research has shown that under these conditions the L1 largely resists attrition: Schmid 2011). In contrast, no such assumption can be made for instructed FL learners, who are known to vary in the proficiency level attained, even given similar amounts of FL instruction or exposure. Therefore, it is extremely important to empirically establish the actual proficiency level attained, in order to study meaningfully the nature of retention and/or attrition during subsequent changing conditions of L2 use. Clearly, longitudinal research designs are required to address these issues.

Studies of all types of attrition (L1, L2) have also shown considerable individual variation. To explain this individual variation, researchers have appealed to a range of non-linguistic factors, including age at onset of (potential) attrition, level of proficiency, length of exposure, attitudes and motivation, and current contact with the language. The influence of these interacting factors has proved difficult to disentangle however. Thus, research on L1 attrition suggests that (in)frequency of use alone does not explain attrition, but rather it depends on context of use. For example, Schmid and Dusseldorp (2010) found that when an L1 is used for professional purposes, those speakers experience less attrition than those who do not use their L1 in more formal contexts. Research on FL attrition (Murtagh 2003; Weltens 1989; Xu 2010) has provided some indication that exposure post-instruction is not a strong predictor of attrition/retention, and that this is true for both extent of exposure and length of exposure.

Instead, the limited research suggests that proficiency level attained might better predict the amount of attrition/retention (Mehotcheva 2010; Murtagh 2003; Xu 2010). It has also been shown that in FL settings attrition is non-linear; initial attrition can be rapid but subsequently stabilizes (Weltens 1989).

A study of special relevance here is that of Mehotcheva (2010), who researched a population of mixed L2/FL learners similar to that under consideration in this study. She examined the FL Spanish attrition of L1 Dutch and L1 German university students, between 12 and 72 months following a period of study abroad (SA) in Spain. She had two groups of participants, one longitudinal ($n = 5$), where data were collected twice over a one-year period post-SA. The other participants contributed to a cross-sectional study. Here, four different groups were examined, at varying lengths of time post-SA. Data came from a variety of measures, including a sociolinguistic interview (which included questions about linguistic history and use, including during SA), a C-test, and a picture-naming test (designed to investigate lexical access). Analysis of the longitudinal data demonstrated that the participants experienced attrition, particularly reduced access to lexical items. Specifically, they used more foreign words, pseudowords and disfluency markers, and a lexical diversity measure (*D* scores in the interview) declined over the one-year period. Analysis of the cross-sectional data demonstrated that higher initial proficiency (measured via self-assessment) was a significant predictor of language retention. Little influence was found for attitude and motivation, disuse, language contact, or length of exposure during the post-SA and instructional period.

Other longitudinal studies focusing on the benefits of SA have in general found that linguistic gains abroad are retained over the following year, during which participants continued to receive formal instruction (Howard 2009; Huensch & Tracy-Ventura 2017; Llanes 2012; Regan 2005). For example, Llanes (2012) included a measure of lexical complexity, Guiraud's Index of Lexical Richness (number of types divided by the square root of the total number of tokens) and found that her participants retained the gains they made during SA in both oral and written language.

Several studies focusing on L1 attrition (Schmid & Dusseldorp 2010; Schmid & Jarvis 2014; Yilmaz & Schmid 2012) have investigated the difference in lexical diversity between L1 attriters and monolingual controls on tasks such as narrative retells and interviews. Such studies have also examined extralinguistic factors which correlate with levels of lexical diversity in these tasks and other measures as well (e.g., lexical access, disfluency) but with inconsistent results. For

example, findings from Schmid and Dusseldorp (2010) suggested a relationship between length of emigration and lower lexical diversity scores. However, Schmid and Jarvis (2014) did not find any significant relationships between the extralinguistic variables they studied and lexical diversity scores. Results of their study did suggest that data elicited in interviews better predicted speakers as controls or attriters than data from more controlled tasks.

Schmid and Mehotcheva (2012) provide a useful overview of theoretical frameworks which have been employed in conceptualizations of FL attrition and are applicable to long-term retention as well. Some psycholinguistic theorists appeal to frequency/recency of exposure and use of FL. Thus for example, the Dynamic Model of Multilingualism proposed by Herdina and Jessner (2002) assumes that language proficiency is always in a state of dynamic change, with positive/negative growth taking place in different parts of the language system, depending on the extent and nature of active language use, and ongoing competition between different linguistic systems. The Neurolinguistic Theory of Bilingualism and associated Activation Threshold Hypothesis (ATH) advanced by Paradis (1993, 2004) argues somewhat similarly, that bilingualism comprises a "system of systems" where active use of part of the system (e.g., a particular language) lowers the neurolinguistic activation threshold for that part and inhibits other competing systems or subsystems. The practical effect is that the most commonly used system(s) are the most accessible, and attrition "is a result of lack of long term stimulation" (Paradis 2007, 125). However Paradis also allows that motivation as well as an advanced level of proficiency may have a protective effect. Given that different parts of the system may attrite at different rates, Paradis (2007) predicts that the lexicon is more susceptible to attrition than grammar, which is more strongly sustained by procedural memory.

To conclude this brief review, it is clear that while different theoretical frameworks have been proposed, there is limited empirical research on long-term FL retention/attrition which might help to choose among these. There is some evidence that lexical attrition occurs, but how it might differ in oral vs. written data from the same learners is not known. There is a suggestion that proficiency is a predictor of language retention, but other non-linguistic factors such as motivation or language contact have perhaps surprisingly not been shown to be predictors. Given the expense and effort of educational investment in FL learning, a fuller understanding of the long-term evolution of FL proficiency and the conditions which will promote retention is needed. This study aims to make a preliminary

contribution to addressing these questions, using a longitudinal learner corpus of French and Spanish which is complemented with extensive data on participants' overall proficiency and patterns of FL exposure and use. In our previous research (Huensch et al. 2019) we focused on the attrition/retention/development of oral fluency and proficiency, and found that both language contact/use and peak proficiency attained were important variables in the retention of fluency and proficiency three years after formal instruction ended. The current study expands on this work by utilizing a corpus-based approach to investigate the attrition/retention/development of FL lexis, and its relation with (a) lexical proficiency attained at the end of SA and (b) post-SA FL exposure. The current study investigates the following research questions:

(1) To what extent does lexical diversity, operationalized as *D* (Malvern & Richards 2002) and the Moving Average Type–Token Ratio (MATTR), change four years after residence abroad, in speech and writing?
(2) To what extent can post-SA language exposure and peak lexical diversity attainment explain changes in lexical diversity of learners four years after residence abroad?

2 Methods

2.1 Participants

Participants in the current study are part of the Languages and Social Networks Abroad Project (LANGSNAP: Mitchell et al. 2017), an ongoing longitudinal study that began in 2011. LANGSNAP began as a project examining the social and individual factors that influenced language learning during residence abroad/SA. Over nearly two years, data were collected six times from 56 British university French and Spanish degree students: once at the end of Year 2 of their four-year degree (May 2011), three times during a nine-month stay abroad program during Year 3 (October 2011, February 2012, May 2012), and twice after return to their home university during Year 4 (October 2012, February 2013). In May 2016 a follow-up study was initiated to continue the project and examine the long-term evolution of their FL proficiency and patterns of FL use.

All of the original LANGSNAP participants ($n = 56$) were invited to participate in the current study. Contact was made primarily through a private project Facebook group, and 33 participants contributed once again: 15 Spanish L2 participants and 18 French L2 participants. Table 30 provides background data about the participants included in

Table 30 Age and years studying L2 of the 33 LANGSNAP 3.0 participants

	Spanish L2	**French L2**
Participants	15 (12 females)	18 (17 females)
Mean age	25.5	24.7
Years studying L2 at pretest	6.1	10.5

the current study. At the time of data collection, 12/15 of the Spanish participants were living in the UK, with the others living in Canada and Spain. Of the French participants, 12/18 were living in the UK, with the others living in Australia, Belgium, Finland, France, and Thailand.

2.2 Procedure

Participants living in the UK in 2016 were met by a member of the research team at a time and location convenient for them. Those living abroad were visited when feasible ($n = 3$) and if not, all tasks were completed online, including the oral tasks via Skype. Each data collection session lasted approximately 1.5 hours, and task sequencing was randomized.

2.3 Learner Corpus

Three communicative tasks produced the learner corpus analyzed in this study. For continuity, we selected from among the tasks that were used in the original LANGSNAP study (see Table 31), and administered an oral interview, an oral picture-based narrative, and a written argumentative essay (for more details see Tracy-Ventura et al. 2016). For both the oral picture-based narrative and the written argumentative essay, we used prompts which the participants had last completed four years earlier: the Cat Story (oral narrative) and a prompt on 'Gay marriage and adoption' (argumentative essay). Where tasks are used repeatedly, there could conceivably be some risk of priming (e.g., later performances might be more fluent than earlier ones), or a ceiling effect. However, we believed that the long gaps between task administrations reduced the priming risk. For the investigation of lexical development more particularly, it is known that the specific task prompt can influence measures of lexical diversity (see Tracy-Ventura et al. 2016; Kyle in press). In designing this

Table 31 Project timeline

Data collection wave	Location	Oral tasks	Written task (argumentative essay)
Pre-sojourn: May 2011	Home university	Oral interview Cat Story	Gay Marriage & Adoption
In-sojourn 1: Oct 2011	Abroad	Oral interview Sisters Story	Legalization of Marijuana
In-sojourn 2: Feb 2012	Abroad	Oral interview Brothers Story	Taxes on Junk Food
In-sojourn 3: May 2012	Abroad	Oral interview Cat Story	Gay Marriage & Adoption
Post-sojourn 1: Oct 2012	Home university	Oral interview Sisters Story	Legalization of Marijuana
Post-sojourn 2: Feb 2013	Home university	Oral interview Brothers Story	Taxes on Junk Food
Post-sojourn 3: June 2016	Varied	Oral interview Cat Story	Gay Marriage & Adoption

study, this was seen as a positive reason to compare data from collection waves which used the same prompts. In the current study, those waves are pre-sojourn, in-sojourn 3, and post-sojourn 3 (see Table 31).[1]

The semi-structured oral interview was administered by a member of the research team; the pre-established questions were designed to elicit information about participants' experiences since their last interview in February 2013. In particular, they were asked about what they had been doing since graduation (e.g., jobs, traveling), their current leisure activities, who they currently live with and spend time with, any plans for the future, and if they could do their degree over again, whether they would study languages.

The Cat Story was a picture-based story retell task, borrowed from Domínguez et al. (2013). The story depicts the experiences of a little

[1] A reviewer pointed out that it is important to demonstrate that any changes between in-sojourn 3 and post-sojourn 3 were not the result of gains made during the final year of instruction. A comparison of *D* scores between in-sojourn 3 and post-sojourns 1 and 2 indicated that in-sojourn 3 was the peak score for the interview and writing tasks. Post-sojourn 1 scores were higher than in-sojourn 3 scores for the narrative task, but that was likely due to task effects (see Tracy-Ventura et al. 2016). The same pattern of results was true for MATTR. In-sojourn 3 is the most appropriate choice of data collection wave for the current study because it represents the peak *D* and MATTR scores for the interview and because it used the same narrative task as post-sojourn 3 (Cat Story).

Table 32 LANGSNAP corpus word counts by task

	Oral narrative	Oral interview	Argumentative essay	Total
French				
L2 Learners 2011–2013 ($n = 29$)	65,905	222,014	36,339	324,258
L2 Learners in 2016 ($n = 18$)	6,365	16,957	3,094	26, 416
Spanish				
L2 Learners 2011–2013 ($n = 27$)	53,497	214,364	36,059	303,920
L2 Learners in 2016 ($n = 15$)	4,480	15,994	3,140	23,614
Total	130,247	469,329	78,632	678, 208

girl and her cat one day when the cat got lost. Participants were given a few minutes to look over the pictures and ask the researchers any questions they had about the story before beginning their retell, with support from the pictures. For those completing the task online, the pictures were made available via a link; all interviews and story retells were audio-recorded, via Skype where necessary.

The written argumentative essay was based on the prompt 'Do you believe that gay people have the right to get married and have children?' This prompt was borrowed from Lozano and Mendikoetxea (2013) and administered through a specially written computer program. Having seen the prompt, participants had three minutes of planning time and 15 minutes of writing time, with a target of 200 words. While writing, they could still see the prompt and a word counter. Buttons allowed for easy insertion of accented letters. If they finished early, they could click on a submit button. After 15 minutes, the program automatically closed and they were not allowed to write any more.

The total word counts for the LANGSNAP corpus are provided in Table 32 and are separated by task. These counts include the data from the original project which ran from 2011 to 2013, plus the new data (LANGSNAP 3.0).[2] As shown in this table, most data is oral, with the oral narrative and the oral interview together comprising 88 percent of the corpus.

[2] Data were also collected from L1-speaking controls. Those word counts are not included here. The L1 data are also freely available.

2.4 Other Materials

Participants also filled out a background questionnaire modeled after Mehotcheva (2010: see https://languageattrition.org/). This questionnaire collected information about language use and activities (e.g., work, travel, lifestyle) since participants' graduation. For example, participants answered questions related to L2 use at work, or partners with whom they spoke their L2. Questions also focused on whether they had pursued further L2 instruction, how confident they were speaking the L2, whether they planned to live abroad again, etc. In addition, they completed a reflective interview in English (see the Appendix) which asked questions about their perceived L2 abilities, their current professional/academic activities, the languages used in their current social networks, etc. Finally, participants completed the Language Engagement Questionnaire (LEQ, available on IRIS, see McManus et al. 2014) first designed for in-sojourn data collection waves in LANGSNAP, which asked them to first select any languages they used on a regular basis and then indicate how frequently they used each language for a variety of activities (e.g., watch movies, write emails, engage in small talk).

2.5 Analysis

The oral and written data were transcribed following CHAT conventions for use with CLAN (MacWhinney 2000). All transcripts were checked at least three times for accuracy by different members of the research team. Two CLAN commands provided the measures of lexical diversity used in the current study: VOCD and MATTR. VOCD is the command that generates a *D* score, a measure that has been used in previous research on L1 (e.g., Schmid & Jarvis 2014) and FL attrition (e.g., Mehotcheva 2010). We also chose to run our analyses using MATTR, which has been found to be less sensitive to text length (Fergadiotis et al. 2015) and is therefore a potentially more valid measure of lexical diversity.

To run VOCD, a minimum of 50 tokens is needed. The way that *D* is calculated is through a procedure that

> takes the average TTR from 100 random samples of 35 words drawn from a text, then 100 random samples of 36 words, then 37 words, all the way to 50 words. It plots the average TTR values for each sample size on a curve, and then uses a formula with a single parameter to find the best fit between the formula-generated curve and the observed TTR curve.
>
> (Castañeda-Jiménez & Jarvis 2014, 504)

MATTR (Covington & McFall 2010) on the other hand,

> calculates the lexical diversity of a sample using the Moving Average Type–Token Ratio (MATTR). This index is based on a moving window that computes TTRs for each successive window of fixed length (N). Initially, a window length is selected (e.g., 10 words) and the TTR for words 1–10 is estimated. Then, the TTR is estimated for words 2–11, then 3–12, and so on to the end of the text. For the final score, the estimated TTRs are averaged.
> (MacWhinney 2000, 95)

For this study we used a window length of 75 words, the shortest text length (Covington 2007).

Language exposure was analyzed based on participants' responses on the background questionnaire, the LEQ, and what they reported in their English and L2 interviews. Based on data from these four sources, each participant was coded as either having 'intense' or 'limited' exposure depending on the consistency and intensity of the L2 contact since graduation. Initial coding was based on five categories in which intensity and frequency were coded separately (e.g., neither intense/consistent; intense & sporadic; intense & consistent; limited & sporadic; limited & consistent). However, this coding proved complex because of participants' diverging experiences. Thus, a decision was made to limit the coding to either intense or limited. For example, participants in the limited category ($n = 20$) were living in their home countries and on the LEQ either reported practically no engagement with the L2 or, in seven cases, did not even list the L2 as a language used. On the other hand, nine participants had lived abroad again in a French- or Spanish-speaking country for an extended period of time (i.e., 9–12 months), or were still living abroad, and were coded as intense. Other participants coded as intense ($n = 4$) used their L2 with significant others or extensively at work. It should be noted, however, that even participants coded as intense made little use of L2 in writing.

To investigate research question (1), two separate three-way mixed repeated measures ANOVAs were conducted to understand the effects of language, time, and task on lexical diversity. For both tests, language (French, Spanish) was the between-subjects variable, and time (pre-sojourn, in-sojourn 3, post-sojourn 3) and task (interview, narrative, writing) were the within-subjects variables. Language was included as a variable because the data come from a smaller number of participants ($n = 33$) compared to the original project ($n = 56$). Analysis of the 2011–2013 data suggested that the two groups were similar (e.g., evidence of development while abroad and potential attrition after return home – see Mitchell et al. 2017). We felt it was

important to replicate the analysis with this 2016 subset to examine whether the results would mirror those of the earlier data. For one test *D* was the dependent variable and for the other MATTR was the dependent variable. The data were within acceptable ranges concerning the assumptions of ANOVA tests (e.g., normally distributed, lacking extreme values) except that the assumption of sphericity was violated; thus, a Greenhouse–Geisser correction was applied. Alpha was set at 0.05. For posthoc comparisons, effect sizes are reported using *d* and interpreted using Plonsky and Oswald (2014)'s field-specific recommendations for within-groups: small effect ($d = 0.60$), medium effect ($d = 1.00$), and large effect ($d = 1.40$), and between-groups: small effect ($d = 0.40$), medium effect ($d = 0.70$), and large effect ($d = 1.00$).

3 Results

Research question (1) investigated to what extent lexical diversity, operationalized as *D* and MATTR, changed in speech and writing four years after residence abroad. First, descriptive statistics with the mean and standard deviations for tokens and types are provided in Table 33 for each task, data collection wave, and L2 group. As shown, the interview had the highest average number of words, while the written argumentative essay had the lowest average number of words.

Table 34 provides the means and standard deviations for all *D* scores separated by language group for each task and each time point. Results of the three-way mixed ANOVA with *D* scores demonstrated that there was no three-way interaction between language, time, and task, $F(2.94, 79.38) = 1.08$, $p = 0.361$, partial $\eta^2 = 0.04$. There was, however, a two-way interaction between time and task, $F(2.94, 79.38) = 7.40$, $p < 0.001$, partial $\eta^2 = 0.22$, but no two-way interaction between time and language, $F(1.89, 51.05) = 2.03$, $p = 0.144$, partial $\eta^2 = 0.07$, or task and language, $F(1.39, 37.43) = 0.65$, $p = 0.475$, partial $\eta^2 = 0.02$. A main effect of time, $F(1.89, 51.05) = 19.08$, $p < 0.001$, partial $\eta^2 = 0.41$, and a main effect of task, $F(1.39, 37.43) = 74.58$, $p < 0.001$, partial $\eta^2 = 0.73$, were also found.

Given the significant interaction between time and task, a one-way repeated measures ANOVA with time as the within-subjects variable was conducted for each task (with the language groups combined). Results for the interview task indicated a significant effect of time, $F(1.85, 55.44) = 42.56$, $p < 0.001$, partial $\eta^2 = 0.59$. Posthoc analyses demonstrated significant differences between all three time points: pre-sojourn–in-sojourn 3 ($p < 0.001$, $d = 1.12$),

Table 33 Descriptive statistics of types and tokens over time

	French			Spanish		
	Pre-sojourn	**In-sojourn 3**	**Post-sojourn 3**	**Pre-sojourn**	**In-sojourn 3**	**Post-sojourn 3**
Interview						
Tokens	1,353.82 (412.73)	1,277.06 (454.59)	942.06 (221.56)	1,104.47 (248.86)	1,653.20 (548.73)	1,066.27 (355.02)
Types	311.78 (75.30)	323.82 (79.28)	301.72 (49.72)	285.53 (53.78)	413.73 (98.53)	332.07 (81.19)
Narrative						
Tokens	438.33 (134.42)	393.59 (162.70)	374.41 (124.91)	292.27 (70.79)	302.36 (99.89)	320.00 (114.43)
Types	140.39 (32.13)	140.53 (38.97)	148.06 (37.74)	108.67 (17.50)	127.29 (28.47)	133.86 (38.98)
Writing						
Tokens	218.22 (27.24)	224.53 (30.59)	174.94 (43.04)	191.00 (28.76)	226.93 (30.80)	209.33 (39.88)
Types	110.89 (14.61)	117.59 (15.13)	102.50 (23.33)	105.47 (13.66)	124.00 (13.67)	116.73 (21.84)

Table 34 Descriptive statistics for D *scores over time*

	French			Spanish		
	Pre-sojourn	**In-sojourn 3**	**Post-sojourn 3**	**Pre-sojourn**	**In-sojourn 3**	**Post-sojourn 3**
Interview	60.16 (18.26)	77.46 (12.43)	88.70 (11.13)	67.84 (8.03)	79.64 (10.51)	84.59 (7.78)
Narrative	46.38 (8.97)	46.63 (11.18)	58.00 (9.04)	45.73 (7.24)	54.92 (8.75)	55.76 (9.23)
Writing	78.89 (20.34)	84.39 (17.00)	85.87 (25.10)	77.55 (18.11)	83.13 (15.94)	78.40 (14.81)

pre-sojourn–post-sojourn 3 ($p < 0.001$, $d = 1.92$), and in-sojourn 3–post-sojourn 3 ($p = 0.001$, $d = 0.86$).

Effect sizes of a small and medium effect between in-sojourn 3–post-sojourn 3 and pre-sojourn–in-sojourn 3, respectively, suggest that *D* scores increased throughout the testing period, with the largest effect occurring across the entire testing period between pre-sojourn and post-sojourn 3.

Similarly, results for the narrative task indicated a significant effect of time, $F(1.90, 53.08) = 18.51$, $p < 0.001$, partial $\eta^2 = 0.40$. Posthoc analyses demonstrated significant differences between pre-sojourn–post-sojourn 3 ($p = 0.001$, $d = 1.27$) and in-sojourn 3–post-sojourn 3 ($p = 0.008$, $d = 0.67$), but not pre-sojourn–in-sojourn 3 ($p = 0.060$, $d = 0.45$). These results suggest that when the scores of both groups are analyzed together, increases in *D* scores during SA were minimal (marginal significance and negligible effect), but that they improved between returning home from SA and three years after graduation (small effect) and throughout the testing period (medium effect). Unlike results from the oral tasks, results for the writing task did not indicate a significant effect of time, $F(1.87, 56.03) = 1.37$, $p = 0.263$, partial $\eta^2 = 0.04$.

Table 35 provides the means and standard deviations of the MATTR scores separated by language group for each task and each time point. Results of the three-way mixed ANOVA with MATTR scores demonstrated similar results. There was no statistically significant three-way interaction between language, time, and task, $F(3.48, 93.89) = 2.04$, $p = 0.104$, partial $\eta^2 = 0.07$. There were significant two-way interactions: one between time and task, $F(3.48, 93.89) = 6.97$, $p < 0.001$, partial $\eta^2 = 0.21$; and one between time and language, $F(1.90, 51.26) = 6.51$, $p = 0.003$, partial $\eta^2 = 0.19$. Thus, one-way repeated measures ANOVAs with time as the within-subjects variable were conducted for each task (with the language groups combined). Results for the interview task indicated a significant effect of time, $F(1.91, 57.40) = 44.79$, $p < 0.001$, partial $\eta^2 = 0.60$. Posthoc analyses demonstrated significant differences between all three time points: pre-sojourn–in-sojourn 3 ($p < 0.001$, $d = 1.15$), pre-sojourn–post-sojourn 3 ($p < 0.001$, $d = 1.94$), and in-sojourn 3–post-sojourn 3 ($p = 0.001$, $d = 0.99$). These results suggest that similar to the results for *D*, MATTR scores increased between each of the testing periods on the interview task, with the largest effect occurring across the entire testing period. Similarly, results for the narrative task indicated a significant effect of time, $F(1.63, 45.68) = 17.63$, $p < 0.001$, partial $\eta^2 = 0.39$. Posthoc analyses again demonstrated significant differences between all three time points: pre-sojourn–in-sojourn 3 ($p = 0.048$, $d = 0.58$),

pre-sojourn–post-sojourn 3 ($p < 0.000$, $d = 1.23$), and in-sojourn 3–post-sojourn 3 ($p = 0.016$, $d = 0.62$). Like the narrative results for *D*, increases in MATTR scores were minimal during SA (marginal significance and negligible effect) but improved throughout the testing period (medium effect). These results suggest that when both language groups are analyzed together, the group means continue to improve at post-sojourn 3 in both oral tasks. Similar to the results for *D*, results for the writing task did not indicate a significant effect of time, $F(1.97, 59.15) = 1.60$, $p = 0.212$, partial $\eta^2 = 0.05$. In sum, these results suggest that there are no major differences between language groups, but rather the main difference appears to be between the oral and written tasks. Therefore, other variables likely explain the continued improvement of the group as a whole at post-sojourn 3, and two variables in particular are examined in research question (2).

Research question (2) investigates to what extent variables such as language exposure (since the last data collection wave in February 2013) and lexical diversity attained at the end of SA could predict changes in lexical diversity from in-sojourn 3 to post-sojourn 3. This analysis is restricted to the tasks that show significant changes over time (narrative and interview), and only results from the MATTR score regression analyses are reported.[3] Therefore, a standard multiple regression was conducted with the dependent variable of change in MATTR between in-sojourn 3 and post-sojourn 3. The two independent variables were peak attainment at the end of residence abroad (MATTR value at in-sojourn 3) and 'language exposure' (a nominal, dichotomous variable coded as 0 = 'limited' and 1 = 'intense'). Due to the small sample size, only two independent or explanatory variables could be included. The data were within acceptable ranges concerning the assumptions of a regression test.

Results of the multiple regression are displayed in Table 36. For the MATTR scores on the narrative, results indicated a statistically significant model, with an R^2 value of 0.58. Both language exposure ($p = 0.008$) and the MATTR score at in-sojourn 3 ($p < 0.001$) contributed significantly to the model. These results suggest that the two explanatory variables predict 58 percent of the variance in the change scores, with the MATTR score at in-sojourn 3 explaining the most variance (it has the higher score in the β column and a lower p value).

[3] Results from the regression analyses conducted with *D* scores patterned the same as those for MATTR scores. We chose to only report MATTR score results in this chapter both for space considerations and because of the finding that MATTR is less sensitive to text length (Fergadiotis et al. 2015) and thus a potentially more reliable measure of lexical diversity than *D*.

Table 35 Descriptive statistics for MATTR scores over time

	French			Spanish		
	Pre-sojourn	**In-sojourn 3**	**Post-sojourn 3**	**Pre-sojourn**	**In-sojourn 3**	**Post-sojourn 3**
Interview	0.65 (0.04)	0.68 (0.02)	0.71 (0.02)	0.64 (0.03)	0.69 (0.03)	0.70 (0.02)
Narrative	0.62 (0.03)	0.61 (0.04)	0.66 (0.03)	0.59 (0.05)	0.64 (0.03)	0.64 (0.03)
Writing	0.72 (0.04)	0.73 (0.04)	0.74 (0.05)	0.72 (0.04)	0.74 (0.04)	0.73 (0.03)

Table 36 Regression results

		B	***SE*$_b$**	***β***	***F***	***p***	***R*^{2}**
Narrative change	Intercept	0.472	0.078		$F(2,28) = 19.96$	0.000	0.58
	Exposure	0.027	0.009	0.355		0.008	
	MATTR S3$^+$	–0.732	0.123	–0.736		0.000	
Interview change	Intercept	0.776	0.106		$F(2,30) = 25.28$	0.000	0.62
	Exposure	0.021	0.008	0.342		0.009	
	MATTR S3$^+$	–1.116	0.157	–0.870		0.000	

Note. The degrees of freedom change between the two tasks because we have missing data points in the narratives. +S3 = in-sojourn 3

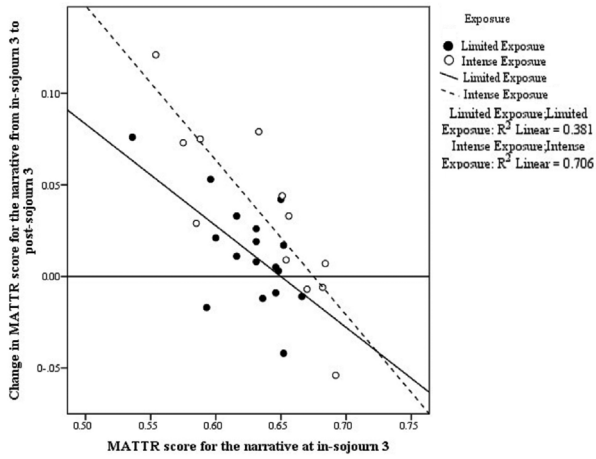

Figure 7 Scatterplot of Narrative regression results

According to Plonsky and Ghanbar (2018), this is considered a large effect. Figure 7 is a scatterplot which displays the results for each participant. Those participants who were categorized as limited exposure are represented by the black dots and those categorized as intense exposure are represented by the white dots. As shown in Figure 7, those participants who had lower MATTR scores at in-sojourn 3, tended to make the most gains.

Results of the regression testing MATTR scores on the interview also indicated a statistically significant model with an R^2 value of 0.62. Again, both variables contributed significantly to the model: language exposure ($p = 0.009$) and MATTR score at in-sojourn 3 ($p < 0.001$), with MATTR scores at in-sojourn 3 explaining the most variance. The scatterplot displaying these results is shown in Figure 8. There was less variation in MATTR scores at in-sojourn 3 in the interview than the narrative; however, the overall results are the same and suggest that changes in MATTR scores at post-sojourn 3 can be predicted by both MATTR scores at in-sojourn 3 and exposure post-instruction.

4 Discussion

The purpose of this study was to investigate the long-term evolution of lexical diversity post-SA and post-formal language instruction in two groups of participants, one L2 French and one L2 Spanish. Since graduating with their BA in French and Spanish, a majority of the participants were working in the UK in jobs that required little to no

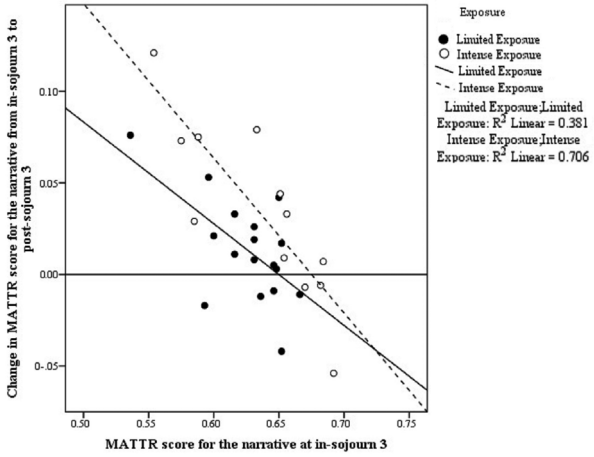

Figure 8 Scatterplot of Interview regression results

use of their L2s, and also reported limited use of the L2 in informal social settings. Another group was living abroad again and/or was maintaining social relations with one or more speakers of their L2, and used that language often in their everyday lives. These differences in L2 use allowed us to investigate the claims of previous researchers, regarding the relationship between L2 use and L2 vocabulary retention/development/attrition, for example the claim of Paradis (2007) that lexical knowledge is the most sensitive to attrition compared to grammar or phonetics, and that frequency of use is linearly related. Using a longitudinal learner corpus that was carefully designed to include oral and written data from the same participants, collected multiple times from 2011 to 2016, and formatted and annotated for semi-automatic analysis, we were able to compare participants' vocabulary use over several years and examine how factors such as ongoing language exposure/use and lexical diversity scores at the end of a year abroad predict retention four years later.

Research question one examined to what extent lexical diversity, operationalized as *D* and MATTR, changed four years after residence abroad. Two-way interactions between time and task were found for both the *D* and MATTR scores, and posthoc tests revealed that there were differences over time in the oral tasks but not the written tasks. In particular, the participants showed linear progress over time that continued at post-sojourn 3 during the 2016 data collection wave in both the oral narrative and interview tasks. The fact that lexical diversity did not change significantly over time in writing but did in

speaking could be due to higher pretest scores in the writing task compared to the oral tasks, i.e., the existence of a ceiling effect for that particular task. However, participants also did not report much formal L2 writing from 2013 to 2016 (e.g., only three participants reported writing emails several times a week or more in the L2). Increases in oral lexical diversity over time could be explained by increased automaticity in lexical access that is improved in online speech production with continued practice/use.

The result that the group as a whole showed continued improvement at post-sojourn 3 (three years after formal instruction had ended) in the oral tasks was surprising and helped to motivate exploration of extralinguistic factors that could help explain this development. This was the focus of research question (2), which we investigated using a multiple regression. Due to the low number of participants ($n = 33$), it was only possible to include two predictor variables in the multiple regression: post-SA language exposure/use and peak lexical diversity attainment, variables that have been the focus of previous research on FL attrition (Bardovi-Harlig & Stringer 2010; Schmid & Mehotcheva 2012). Recall that research on L1 attrition has not found L1 use to be a significant predictor, except when it is used for professional purposes (Schmid & Dusseldorp 2010). That is, those people who use it for professional purposes have less L1 attrition than those who do not. In the current study we ran two multiple regressions, one using the MATTR scores in the narrative and another with the scores in the interview. In contrast to the L1 attrition research and the limited research on FL attrition (Mehotcheva 2010), in the current study language exposure/use was a significant predictor of change over time, as were MATTR scores at in-sojourn 3. The latter explained more of the variance, which suggests that higher MATTR scores at the end of their year abroad may initially help protect participants from attrition even with limited exposure after formal instruction. Neisser (1984) claimed that there may be a general critical threshold at which, once learners reach that level, their linguistic knowledge becomes permanent and immune to decay. As we continue this project into the future, we will be able to further test this claim.

The fact that L2 exposure/use post-instruction emerged as a significant predictor in the regression provides support for Paradis' (2004) Activation Threshold Hypothesis. Previous research on FL attrition has not found exposure to be a significant predictor. However, several methodological issues could help explain conflicting results. For example, Mehotcheva (2010) had only five participants in her longitudinal group, and in her cross-sectional data proficiency was assessed based on self-report. Additional longitudinal research is needed to

corroborate these findings. Our research investigating oral fluency with these same participants (Huensch et al. 2019) also found exposure/use and peak attainment to be significant predictors of change three years after formal instruction.

We have only just begun to explore the variables that predict attrition/maintenance/development of the LANGSNAP participants' L2 French and Spanish abilities. Much future work remains. In particular, we plan to examine individual participants in more depth, providing case studies of those participants who demonstrated evidence of the most extreme development and attrition. Additionally, we plan to extend our analysis to include other aspects of lexical complexity, such as lexical sophistication, as well as accuracy and syntactic complexity. Lastly, in the future we will examine the role that individual differences, such as motivation and attitudes, play in FL retention and attrition.

Several limitations of the current study should be acknowledged. One issue is task familiarity. This was the third time that these participants orally retold the Cat Story and wrote an argumentative essay in response to the prompt about gay marriage and adoption. However, four years had passed since the last time they completed these tasks, which likely helped to address this concern. Additionally, performance on the writing task may have been limited by a ceiling effect; the more open-ended oral tasks were not affected in this way, however. Another potential limitation is the dichotomous operationalization of language exposure that was used. It is possible that if we were able to further differentiate the participants, then the results for language exposure/use may explain more of the variance in change scores.

These limitations notwithstanding, the current study takes us one small step closer to understanding variables that influence the long-term evolution of FL proficiency. By building a longitudinal learner corpus that includes the collection of metadata, is formatted using agreed-upon conventions, and is shared publicly, we aim to provide a resource that will support further investigations drawing on SLA theory, and encourage other researchers to conduct additional studies using these data.

Acknowledgments

LANGSNAP was funded by the ESRC (award number RES-062-23-2996). LANGSNAP 3.0 was funded by a *Language Learning* Small Research Grant, USF World Faculty Mobility Grant, and a USF New Researcher Grant. We are grateful to our LANGSNAP colleagues, the participants, and various research assistants for their contribution to this study.

References

Bahrick, H. (1984a). Fifty years of second language attrition: Implications for programmatic research. *The Modern Language Journal* 68(2), 105–118.

(1984b). Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. *Journal of Experimental Psychology: General* 113(3), 1–29.

Bardovi-Harlig, K. & Stringer, D. (2010). Variables in second language attrition: Advancing the state of the art. *Studies in Second Language Acquisition* 32 (1), 1–45.

Castañeda-Jiménez, G. & Jarvis, S. (2014). Exploring lexical diversity in second language Spanish. In K. L. Geeslin (ed.), *The Handbook of Spanish Second Language Acquisition*, 498–513. Oxford: Wiley-Blackwell.

Covington, M. (2007). MATTR User Manual, retrieved from https://athenaeum.libs.uga.edu/handle/10724/19840 (accessed June 13, 2020).

Covington, M. & McFall, J. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics* 17(2), 94–100.

Domínguez, L., Tracy-Ventura, N., Arche, M., Mitchell, R., & Myles, F. (2013). The role of dynamic contrasts in the L2 acquisition of Spanish past tense morphology. *Bilingualism: Language and Cognition* 16(3), 558–577.

Fergadiotis, G., Wright, H., & Green, S. (2015). Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research* 58(3), 840–852.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer (ed.), *Corpora and Language Teaching*, 13–32. Amsterdam: John Benjamins.

Granger, S., Gilquin, G., & Meunier, F. (2015). Introduction: Learner corpus research – past, present, and future. In S. Granger, G. Gilquin, & F. Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*, 1–5. Cambridge: Cambridge University Press.

Hasko, V. (2013). Capturing the dynamics of second language development via learner corpus research: A very long engagement. *The Modern Language Journal* 97(S1), 1–10.

Herdina, P. & Jessner, U. (2002). *A Dynamic Model of Multilingualism: Changing the Psycholinguistic Perspective*. Clevedon: Multilingual Matters.

Howard, M. (2009). Short- versus long-term effects of naturalistic exposure on the advanced learner's L2 development: A case-study. In E. Labeau & F. Myles (eds.), *The Advanced Learner Variety: The Case of French*, 93–123. Oxford: Peter Lang.

Huensch, A. & Tracy-Ventura, N. (2017). L2 utterance fluency development before, during, and after residence abroad: A multidimensional investigation. *The Modern Language Journal* 101(2), 275–293.

Huensch, A., Tracy-Ventura, N., Bridges, J., & Cuesta-Medina, J. (2019). Variables affecting the maintenance of L2 fluency post-study abroad in

the short and long term. *Study Abroad Research in Second Language Acquisition and International Education* 4(1), 96–125.

Kyle, K. (in press). Lexis. In N. Tracy-Ventura & M. Paquot (eds.), *Handbook of Second Language Acquisition and Corpora*. New York, NY: Routledge.

Llanes, À. (2012). The short- and long-term effects of a short study abroad experience: The case of children. *System* 40(2), 179–190.

Lozano, C. & Mendikoetxea, A. (2013). Learner corpora and second language acquisition: The design and collection of CEDEL2. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (eds.), *Automatic Treatment and Analysis of Learner Corpus Data*, 65–100. Amsterdam: John Benjamins.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. (3rd edn.) Mahwah, NJ: Lawrence Erlbaum.

(2017). A shared platform for studying second language acquisition. *Language Learning* 67(S1), 254–275.

Malvern, D. & Richards, B. (2002). Investigation accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing* 19(1), 85–104.

Marsden, E., Mackey, A., & Plonsky, L. (2015). The IRIS repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (eds.), *Advancing Methodology and Practice: The IRIS Repository of Instruments for Research into Second Languages*, 1–21. New York, NY: Routledge.

McManus, K., Mitchell, R., & Tracy-Ventura, N. (2014). Understanding insertion and integration in a study abroad context: The case of English-speaking sojourners in France. *Revue Française de Linguistique Appliquée* 19(2), 97–116.

Mehotcheva, T. (2010). After the Fiesta Is Over: Foreign Language Attrition of Spanish in Dutch and German Erasmus Students. Unpublished Ph.D. dissertation, University of Groningen.

Meunier, F. (2015). Developmental patterns in learner corpora. In S. Granger, G. Gilquin, & F, Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*, 379–400. Cambridge: Cambridge University Press.

Mitchell, R., Tracy-Ventura, N., & McManus, K. (2017). *Anglophone Students Abroad: Identity, Social Relationships, and Language Learning*. New York, NY: Routledge.

Murtagh, L. (2003). Retention and Attrition of Irish as a Second Language. A Longitudinal Study of General and Communicative Proficiency in Irish among Second Level School Leavers and the Influence of Instructional Background, Language Use and Attitude/Motivation Variables. Unpublished Ph.D. dissertation, Rijksuniversiteit Groningen.

Myles, F. (2015). Second language acquisition theory and learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*, 309–332. Cambridge: Cambridge University Press.

Neisser, U. (1984). Interpreting Harry Bahrick's discovery: What confers immunity against forgetting? *Journal of Experimental Psychology: General* 113(1), 32–35.

Paradis, M. (1993). Linguistic, psycholinguistic, and neurolinguistic aspects of 'interference' in bilingual speakers: The activation threshold hypothesis. *International Journal of Psycholinguistics* 9(2), 133–145.

(2004). *A Neurolinguistic Theory of Bilingualism*. Amsterdam: John Benjamins.

(2007). L1 attrition features predicted by a neurolinguistic theory of bilingualism. In B. Köpke, M. S. Schmid, M. Keijzer, & S. Dostert (eds.), *Language Attrition: Theoretical Perspectives*, 121–133. Amsterdam: John Benjamins.

Plonsky, L. & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R^2 values. *The Modern Language Journal* 102(4), 713–731.

Plonsky, L. & Oswald, F. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning* 64(4), 878–912.

Regan, V. (2005). From community back to classroom: What variation analysis can tell us about context of acquisition. In J.-M. Dewaele, (ed.), *Focus on French as a Foreign Language: Multidisciplinary Approaches*, 191–209. Clevedon: Multilingual Matters.

Schmid, M. (2011). *Language Attrition*. Cambridge: Cambridge University Press.

Schmid, M. & Dusseldorp, E. (2010). Quantitative analyses in a multivariate study of language attrition: The impact of extralinguistic factors. *Second Language Research* 26(1), 125–160.

Schmid, M. & Jarvis, S. (2014). Lexical access and lexical diversity in first language attrition. *Bilingualism: Language and Cognition* 17(4), 729–748.

Schmid, M. & Mehotcheva, T. (2012). Foreign language attrition. *Dutch Journal of Applied Linguistics* 1(1), 102–124.

Tracy-Ventura, N., Mitchell, R., & McManus, K. (2016). The LANGSNAP longitudinal learner corpus: Design and use. In M. A. Ramos (ed.), *Spanish Learner Corpus Research: State of the Art*, 117–142. Amsterdam: John Benjamins.

Weltens, B. (1989). *The Attrition of French as a Foreign Language*. Dordrecht/Providence: Foris Publications.

Weltens, B., Van Els, T., & Schils, E. (1989). The long-term retention of French by Dutch students. *Studies in Second Language Acquisition* 11(2), 205–216.

Xu, X. (2010). English Language Attrition and Retention in Chinese and Dutch University Students. Unpublished Ph.D. dissertation, University of Groningen.

Yilmaz, G. & Schmid, M. (2012). L1 accessibility among Turkish-Dutch bilinguals. *The Mental Lexicon* 7(3), 249–274.

Appendix

Reflective Interview Questions

(1) After graduating from X University, have you had any significant travel experiences abroad, i.e., lived in a country other than the UK for more than a month? Have you returned to the place where you did your year abroad? Have you spent time in other countries where you used your foreign language(s)?
(2) How do you think your language abilities in French/Spanish compare now to when you returned from your year abroad and when you finished university? Better, worse, the same? Any other languages to compare?
(3) In what ways do you use your various languages now? Tell me about your YA language (French/ Spanish) first. Are there any other languages that you use regularly? What motivates you to continue to use that/these language(s)?
(4) Have you kept in touch with any people you met abroad? Who? (Other Erasmus or locals) How have you done so? What language (s) do you use with each other?
(5) Are there important people in your life with whom you use French/Spanish? Remember when you were abroad we asked about your top 5? Are there any people in your current top 5 who are French/Spanish users?
(6) Do you think your study abroad experience has benefitted you professionally? What about socially? Personally? Culturally? If so, how? If not, why not?