Studies in Second Language Acquisition (2024), **46**, 169–199
doi:10.1017/S0272263123000384

RESEARCH ARTICLE

Measuring the development of lexical richness of L2 Spanish: A longitudinal learner corpus study

María Díez-Ortega[1] and Kristopher Kyle[2]

[1]Stanford Language Center, Stanford University, Stanford, CA, USA; [2]Department of Linguistics, University of Oregon, Eugene, OR, USA
Corresponding author: María Díez-Ortega; Email: merydior@gmail.com

(Received 12 May 2022; Revised 10 April 2023; Accepted 16 June 2023)

Abstract

Research has indicated that lexical richness is an important indicator of second language (L2) proficiency. However, most research has examined written, cross-sectional English L2 corpora and does not necessarily indicate how spoken lexical use develops over time or whether observed trends are stable across L2s. This study adds to previous research on the development of spoken vocabulary by investigating lexical features of L2 Spanish learners over a 21-month period, using the LANGSNAP corpus. Multiple lexical richness indices used in previous studies were examined including lexical diversity, word frequency, word concreteness, and bigram strength of association. Linear mixed-effects models were run to examine changes over time. The results suggest that although some features of lexical richness (e.g., word frequency) see meaningful change over time, others (e.g., bigram T score) may not be indicative of L2 oral development.

The development of oral communication in a second language (L2) is of critical importance to second language acquisition (SLA) research and L2 pedagogy. An essential component of oral communication skills is productive vocabulary use, which can affect learners' everyday interactions and their academic achievement (Daller et al., 2003). The study of productive vocabulary development and use has been an area of research interest in SLA more broadly and in learner corpus research more specifically over the past 25 years (Laufer & Nation, 1995; Meara & Bell, 2001; Wolfe-Quintero et al., 1998). Characteristics of L2 vocabulary use have been found to correlate with how comprehensible and accented an L2 learner's speech is perceived to be (Appel et al., 2019; Crowther et al., 2018; Saito, 2020; Saito et al., 2016) and have been associated with achievement on spoken assessment tasks (Berger et al., 2019; Crossley et al., 2014; Eguchi & Kyle, 2020; Kyle et al., 2016).

Despite the clear importance of productive L2 vocabulary, there is less agreement on precisely how vocabulary use should be measured. Generally, indices of productive vocabulary use fall under the umbrella of lexical richness (Read, 2000; Yule, 1944). Lexical richness refers to the breadth and depth of productive vocabulary knowledge

(Read, 2000), which is most often measured using the constructs of lexical diversity and lexical sophistication (see Kyle, 2020). Lexical diversity is most commonly measured using an index of lexical variety such as D, MATTR, or MTLD (see Jarvis, 2013a; Zenker & Kyle, 2021). Lexical sophistication refers to the proportion of advanced words used in a language production task and is most commonly measured using an index of word frequency (Kyle & Crossley, 2015; Laufer & Nation, 1995; Read, 2000). Several recent studies have emphasized the importance of taking a multivariate approach to the measurement of lexical sophistication (Eguchi & Kyle, 2020; Kim et al., 2018; Kyle et al., 2018). These studies have suggested, for example, that lexical sophistication is most accurately indexed when features beyond single-word frequency are considered, including psycholinguistic properties of words (such as concreteness; Crossley & Skalicky, 2019; Guo et al., 2013) and the strength of association between word combinations (Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Garner et al., 2019; Granger & Bestgen, 2014).

The majority of productive vocabulary use research has investigated data from cross-sectional corpora, with a relatively small number of studies investigating oral development longitudinally (Berger et al., 2019; Crossley et al., 2010; Crossley et al., 2011a, Crossley et al., 2019; Tavakoli, 2018). Longitudinal research on the same individuals is needed to inform developmental research and support SLA theories of vocabulary acquisition (Hasko, 2013; Meunier, 2015; Ortega & Byrnes, 2009). In addition, most productive vocabulary research has focused on L2 English, likely because of the availability of both English learner corpora and tools for the automatic analysis of English texts. There has not been enough research to determine the degree to which these findings extend to other L2s or even if different L1–L2 pairs may yield different results. For instance, Spanish has a rich verb morphology system that might affect lexical development in different ways than the English verb system does, especially when considering verb inflections (Montrul, 2004; Schnur & Rubio, 2021). In more practical terms, Spanish is the second most common native language in the world, with more than 496 million speakers according to the Instituto Cervantes annual report (Fernández Vítores, 2022). Spanish is also the most common L2 studied in the United States across all educational levels and is also commonly studied worldwide. It is therefore critical to investigate and better understand how multiple features of L2 Spanish may develop in both university and study abroad (SA) contexts, as this can have not only theoretical implications but also practical implications for pedagogy and for the design of SA programs.

A number of studies have investigated the characteristics of productive vocabulary use in L2 Spanish by looking at a single feature of lexical richness (Asencion-Delaney & Collentine, 2011; Berton, 2020; Castañeda-Jimenez & Jarvis, 2014; McManus et al., 2021; Schnur & Rubio, 2021; Tracy-Ventura, 2017). These studies contribute to the new field of L2 Spanish vocabulary development and learner corpus research, but more research is needed to make wider generalizations (Lozano, 2015; Mendikoetxea, 2013). The current study builds on previous research by investigating the productive lexical development of L2 Spanish use using spontaneous oral data from the longitudinal Languages and Social Networks Abroad Project corpus (LANGSNAP: Mitchell et al., 2017), a 21-month corpus of 27 university learners with regard to various features of lexical variation and sophistication.

Defining lexical richness

The term lexical richness was initially used in literary stylometric studies to refer to the size of a particular author's vocabulary (Yule, 1944). Although Yule's use of lexical

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

richness referred specifically to a particular calculation of lexical diversity (Yule's *K*), researchers in applied linguistics have used the term more broadly. Read (2000), for example, explains that lexical richness refers to the breadth and depth of lexical knowledge that is demonstrated in productive language use. Read further outlined three subconstructs of lexical richness—namely lexical density, lexical diversity, and lexical sophistication.

Lexical density refers to the proportion of content words in a text. Although it was hypothesized that more proficient users of a language (who have wider and deeper productive vocabulary knowledge) will produce more informationally dense texts, empirical evidence has suggested that density is more closely related to register (less interactive texts tend to be more lexically dense) than proficiency (Engber, 1995; Lu, 2012; O'Loughlin, 1995). Accordingly, lexical density indices are rarely used as measures of lexical richness.

Lexical diversity typically refers to the variety of words used in a text (Engber, 1995; Jarvis, 2013b; Kyle et al., 2021) and is a productive measure of lexical breadth. As language learners become more proficient, we presume that their productive vocabulary will grow. We also presume that individuals with a larger productive vocabulary will use a wider variety of lexical items to complete a particular language task. Accordingly, we presume that more proficient language users will produce texts that are more lexically diverse than less proficient users (when the language task is kept consistent). An important confound with many indices of lexical diversity is that they conflate text length and lexical variety (Koizumi & In'nami, 2012; McCarthy & Jarvis, 2010; Zenker & Kyle, 2021). In order to estimate lexical breadth, it is therefore important to use indices of lexical diversity that are stable across different text lengths such as moving average TTR (MATTR; Covington & McFall, 2010).

The third subconstruct of lexical richness is lexical sophistication. Lexical sophistication has been conceptualized from two related perspectives. The first perspective focuses on the learnability of a particular word and highlights lexical breadth. For example, words that are more frequent in an individual's language experience are (with some caveats) easier to learn (and use) than words that are less frequent (Ellis, 2002). We therefore presume that more proficient language learners will know (and use) a higher proportion of less frequent words (see Laufer & Nation, 1995; Read, 2000). As lexical sophistication research has matured, other features of word learnability such as concreteness (Brysbaert et al., 2014; Paivio, 1971) have been used to complement frequency indices. The second perspective focuses on reader and/or listener perceptions of lexical proficiency. Although perceptions of lexical proficiency are affected by word learnability features (and therefore vocabulary depth), they are also affected by features of vocabulary breadth such as the use of vocabulary items in the appropriate lexicogrammatical contexts and registers (Garner et al., 2019; Kim et al., 2018; Kyle et al., 2018; Nation, 2001). From this second perspective, indices related to both individual word use (e.g., frequency and concreteness) and indices related to collocation use (e.g., n-gram association strength) are used to measure lexical sophistication. In this paper we adopt the latter perspective, which is in line with a large body of lexical sophistication research published over the past decade (Crossley, Salsbury, et al., 2013; Eguchi & Kyle, 2020; Kyle et al., 2018; Kyle & Crossley, 2015).

Lexical richness in learner corpus research

Indices of lexical richness have been found to correlate with measures of L2 lexical proficiency (Berger et al., 2019; Crossley, Salsbury, et al., 2013; Kyle et al., 2018) and

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

holistic scores of speaking proficiency (Crossley et al., 2014; Eguchi & Kyle, 2020; Kyle & Crossley, 2015) as well as with judgements of communicative competence, such as fluency, comprehensibility or accentedness (Appel et al., 2019; Saito, 2020; Saito & Akiyama, 2017; Tavakoli & Uchihara, 2020; Uchihara & Saito, 2019). Empirical studies have demonstrated that both lexical diversity and lexical sophistication are important predictors of speaking and writing proficiency (Crossley & McNamara, 2013; Guo et al., 2013; Kyle & Crossley, 2015), but the majority of research to date has used written corpora. Findings of L2 writing studies do not always transfer directly to studies involving L2 speaking. For example, the assumption that as learners' language develop, they will use less frequent words seems to be accepted in L2 writing research (Crossley et al., 2011b; Kyle et al., 2018; Laufer & Nation, 1995). In contrast, studies examining L2 spoken corpora present mixed results (Bardel et al., 2012; Crossley et al., 2010; Crossley et al., 2011a; Crossley et al., 2015; Kyle & Crossley, 2015; Lindqvist et al., 2013). Spoken language differs from written language in that it generally involves less planning and lack of editing, especially in interpersonal communication. Linguistic features of spoken and written language are influenced not only by mode but also by register (Biber, 1988; Biber & Conrad, 2019; Kyle et al., 2022). Interpersonal registers vary in their situational context (e.g., everyday conversation versus office hours), which is also characterized by different linguistic features (Biber & Conrad, 2019). Less formal registers tend to be characterized by less sophisticated lexical items (though these items may still be diverse; Biber et al., 2004; Kyle et al., 2022). Consequently, research investigating spontaneous spoken data has found that more proficient learners tend to produce more frequent words (Crossley et al., 2011a; Crossley et al., 2019; Eguchi & Kyle, 2020; Kyle & Crossley, 2015). However, relatively few studies have examined this phenomenon longitudinally, and even fewer studies have done so in L2 Spanish.

In the SA context, a few studies have investigated development of lexical diversity (McManus et al., 2021; Tavakoli, 2018) and/or lexical sophistication (Tavakoli, 2018; Tracy-Ventura, 2017; Zaytseva et al., 2021). The mixed results suggest a complex interplay between time spent abroad, task, and mode. For example, after a 1-month stay abroad, Tavakoli (2018) found that English learners improved their oral lexical diversity (as measured by *D* and MTLD) in a dialogue task, but no significant changes were observed in the monologic task. In another study using monologic tasks, Leonard and Shea (2017) found no significant increase in the development of Spanish lexical diversity (as measured by *D*) after 3 months abroad. However, McManus et al (2021) found lexical diversity scores (as measured by *D*) to increase significantly after 9 months abroad using an oral narrative task, A few studies have also used spontaneous spoken tasks. Mora and Valls-Ferrer (2012) used oral interviews in their 15-month longitudinal study, finding that a 3-month SA period resulted in a significant increase in lexical diversity scores (as measured by Guiraud's index). Similarly, Serrano et al. (2012) also used oral interviews and found a significant increase in spoken lexical diversity (Guiraud's) after the first 3 months abroad, but improvement in written lexical diversity scores were not significant until after 8 months abroad. In contrast, two longitudinal studies found SA to be more beneficial for the development of written than oral lexical diversity as measured by Guiraud's index (Pérez-Vidal et al., 2012; Zaytseva et al., 2021), with formal instruction having a greater impact on oral lexical diversity than SA. However, the studies that used Guiraud's index should be interpreted with caution, given the well-documented intrinsic relationship between Guiraud's index and text length (e.g., McCarthy & Jarvis, 2010; Koizumi & In'nami, 2012; Zenker & Kyle, 2021). Fewer studies have examined lexical sophistication during SA using frequency band-based indices (Tracy-Ventura, 2017; Zaytseva et al., 2021). Despite the important

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

contributions these studies make to our understanding of longitudinal vocabulary development and the field of SA, more research is still needed in this area.

Multivariate approach to lexical richness

Lexical diversity has been widely studied in L2 research (Engber, 1995; Jarvis, 2013a, 2013b) and is calculated by considering the number of types (different words) and the number of tokens (total number of words) in a text. Because of an intrinsic link between text length and simple measures of diversity—such as the type-token ratio or Guiraud's (1960) index—measures such as moving average TTR (MATTR; Covington & McFall, 2010) and the measure of textual lexical diversity (MTLD; McCarthy & Jarvis, 2010) are increasingly used (see Koizumi & In'nami, 2012; Vidal & Jarvis, 2020; Zenker & Kyle, 2021). Although a consistent relationship between lexical diversity and lexical development has been found, lexical diversity indices only account for the use of different words, not how sophisticated the words themselves are. For example, the following Spanish sentences would get a similar diversity score: *el gato come la comida* (the cat eats the food) and *el mamífero devora un manjar* (the mammal devours a delicacy), yet one could argue the later one uses more advanced and sophisticated vocabulary. A combination of lexical diversity and lexical sophistication indices to examine lexical use in a multivariate manner is needed, providing a broader understanding of vocabulary development (Jarvis, 2017; Kyle, 2020).

Measures of lexical sophistication often make use of a reference corpus and include a variety of indices related to word frequency, range, and collocation. As learners become more proficient, they tend to use less frequent words (at least in written productions; Crossley et al., 2011b). An early approach to calculating frequency was the Lexical Frequency Profile (LFP; Laufer & Nation, 1995). An LFP is calculated by grouping word families of a reference corpus and dividing them into frequency bands. The percentage of words in a learner text that occur in each band is then calculated. Laufer and Nation found that more proficient writers tend to use more low-frequency words and more words from the university word list than novice writers, who tend to use high-frequency words found in the 1,000 and 2,000 frequency bands. More recent investigations of Spanish as an L2 have also indicated that more advanced writers tend to produce more low-frequency words (Berton, 2020; Schnur & Rubio, 2021). A recent longitudinal study found that using more low-frequency words can be a predictive feature of development when written and spoken texts are combined (Tracy-Ventura, 2017). An alternative method for calculating frequency scores is to use the mean frequency of words in a text. Mean frequency is calculated by identifying the precise frequency of each word in a learner text (based on a reference corpus) and then calculating the average frequency score. Mean frequency indices have been found to be reasonably strong predictors of L2 proficiency (Crossley, Cobb, & McNamara, 2013, p. 967). Tools such as Coh-Metrix (Graesser et al., 2004) and TAALES 2.0 (Kyle et al., 2018) calculate mean-frequency scores for English texts but not for other languages. Research in L2 English has found that the use of lower frequency words positively correlates with proficiency levels and holistic scores, especially in written registers (Crossley et al., 2011b; Kim et al., 2018; Laufer & Nation, 1995).

However, although we presume that as learners have access to a wider range of less frequent words as they become more proficient users of a language, this does not necessarily mean that they use (or should use) less frequent words in all contexts. For example, a number of studies have found a positive relationship between speaking

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

proficiency and word frequency across a range of relatively informal speaking task types (Berger et al., 2019; Crossley et al., 2011a; Crossley et al., 2014; Eguchi & Kyle, 2020). For example, Kyle and Crossley (2015) found a small, positive relationship between frequency and proficiency scores on an independent TOEFL speaking task, which asks test takers to provide their opinion on an everyday topic. Berger et al. (2019) found similar results using a corpus of L2 conversations rated for lexical proficiency. In a recent study that investigated the relationship between lexical sophistication and oral proficiency interview scores (Eguchi & Kyle, 2020), a strong, positive relationship was found between the "common word" factor (which included several frequency indices) and holistic oral proficiency interview scores. These findings suggest that for some spoken registers, advanced oral proficiency may be characterized by the use of comprehensible higher frequency words. Kyle et al. (2016), for example, found that opinion-based TOEFL iBT independent speaking-task responses included more frequent words than integrated speaking-task responses and that less formal integrated tasks (i.e., campus situation) included more frequent words than more formal integrated tasks that required the synthesis of technical academic information. The register of the task (i.e., campus situation versus academic) affected word frequency. Task type has also been shown to affect lexical features that correlate with holistic judgements of comprehensibility and accentedness (Appel et al., 2019; Crowther et al., 2018) and written lexical sophistication in L2 Spanish texts (Schnur & Rubio, 2021). Thus, register effects should be accounted for in the investigation of lexical sophistication.

It should be noted, however, that not all studies that involve informal speaking tasks have found positive relationships between frequency and proficiency (Bardel et al., 2012; Lindqvist et al., 2011). Using the same small sample of interview data from L2 learners of French ($n = 14$) and Italian ($n = 20$) but different methods of differentiating between basic and advanced vocabulary, Lindqvist et al. (2011) and Bardel et al. (2012) found that "advanced high" learners used a lower proportion of frequent words than "advanced low" learners in each language. Clearly, more research is needed to determine the factors that affect the production of high and low frequency words such as mode, register, L2, and the proficiency levels that are under investigation.

N-grams and strength of association

Collocation use is also an important indicator of proficient word use that taps into one aspect of vocabulary breadth (Gries, 2013; Nation, 2001; Paquot, 2019; Sinclair, 1991). Research analyzing lexical sophistication from a multivariate approach has found corpus-based measures of n-gram frequency and strength-of-association (SOA) to be strong predictors of language development and proficiency (Crossley et al., 2015; Eguchi & Kyle, 2020; Gablasova et al., 2017; Garner et al., 2019; Kyle et al., 2018). N-grams refer to multiword sequences of *n* words (e.g., *en el, soy un)*, and SOA measures the conditional probability that two words in an n-gram will occur together, based on a reference corpus. Common SOA measures include mutual information (MI), which tends to highlight highly exclusive collocations, and T score, which tends to highlight collocations between frequent words.

Strength-of-association and n-gram frequency are generally indicators of both L2 spoken and written proficiency. N-grams indices positively correlate with holistic scores of writing (Gablasova et al., 2017; Garner et al., 2019, 2020; Granger & Bestgen, 2014) and predict longitudinal development trajectories related to writing proficiency (Bestgen & Granger, 2014; Paquot, 2019). These measures also contribute to a large

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

percentage of the variance of holistic scores of lexical proficiency in writing (Garner et al., 2020; Granger & Bestgen, 2014; Kim et al., 2018; Kyle et al., 2018), oral lexical proficiency (Eguchi & Kyle, 2020; Kyle & Crossley, 2015), and rater judgements of comprehensibility (Saito, 2020). Eguchi and Kyle (2020), for example, found that advanced oral proficiency interview score samples were characterized by more strongly associated n-grams (measured using both MI and T scores) and n-grams used in wider contexts. These studies suggest that appropriate collocation use is an important predictor of spoken and written L2 (at least in English). However, related research of collocational use in Spanish is scarce (Vincze et al., 2016) and more research is needed to determine the degree to which these relationships are stable across L2s.

Psycholinguistic word information

In addition to frequency and n-gram measures, psycholinguistic word information indices are an important factor when modeling L2 development via lexical sophistication. These word norms are based on behavioral studies (Brysbaert et al., 2014; Stadthagen-Gonzalez et al., 2017) and are related to a word's saliency (Crossley et al., 2016; Crossley & Skalicky, 2019; Salsbury et al., 2011), which in turn affects the difficulty of learning and using a word (see Ellis, 2002). Therefore, psycholinguistic word information indices are a measure of vocabulary depth. Psycholinguistic word information includes indices such as concreteness (how concrete or abstract a word is), familiarity (how often that word is encountered), and imageability (how easy it is to create a mental image of a word), among others.

Psycholinguistic properties of word knowledge have contributed to the variance explaining lexical proficiency and holistic scores in both spoken and written assessment contexts (Crossley et al., 2016; Crossley et al., 2011a; Eguchi & Kyle, 2020; Kyle et al., 2018). Longitudinal studies of L2 speech samples have indicated that learners use words that are less concrete, less meaningful, and less imageable as a function of time (Crossley & Skalicky, 2019; Salsbury et al., 2011). Cross-sectional studies have found a similar relationship between learner proficiency and the use of less salient words (e.g., words that are less concrete) in L2 English (Crossley et al., 2011a; Eguchi & Kyle, 2020; Kyle & Crossley, 2015). To our knowledge, however, there has been no empirical research on how these norms can be used to index lexical development of L2 Spanish.

Studies of lexical richness of L2 English have found several measures to be relatively stable across written and spoken corpora (e.g., lexical diversity, concreteness) but not all (e.g., word frequency). Most studies of productive lexical use have been cross-sectional in nature. These studies provide an account of the lexical characteristics of learner produced texts at various benchmark levels, but they do not necessarily indicate how lexical use develops over time. Although the number of published longitudinal studies has been increasing (Berger et al., 2019; Crossley et al., 2019; Crossley & Skalicky, 2019), more research is needed (particularly in languages other than English) to understand the ways in which lexical use develops. In particular, more studies that investigate the development of productive lexical use from a multivariate perspective are needed (Eguchi & Kyle, 2020; Kim et al., 2018; Kyle et al., 2018). To date, only a small number of studies have examined lexical richness in Spanish (Asencion-Delaney & Collentine, 2011; Berton, 2020; Castañeda-Jimenez & Jarvis, 2014; Vincze et al., 2016), and only two have used longitudinal designs (McManus et al., 2021; Tracy-Ventura, 2017). As part of a larger study investigating complexity, accuracy, and fluency (CAF) measures, McManus et al. (2021) found that spoken lexical diversity scores as measured by *D*

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

(Malvern & Richards, 2002) increased over three 1-year collection points. Using a frequency-band approach, Tracy-Ventura (2017) found that participants used significantly more low-frequent words in the 3k–5k bands after studying abroad for 9 months. These studies have provided an excellent starting point for research into the development of L2 Spanish productive lexical use. However, more research is needed to understand how L2 Spanish develops with respect to lexical diversity, frequency, saliency, and collocation use.

Current study

The present study adds to previous findings of lexical richness in Spanish learner corpus research and in longitudinal development of spoken language by investigating several lexical and collocational features of language use in L2 Spanish learners over a 21-month period.

Multiple indices of lexical sophistication commonly used in previous learner corpus studies were calculated to allow for comparisons between longitudinal research and oral data in languages other than English. This study is guided by the following research questions:

(1) How do features of lexical richness develop over time in L2 Spanish?
(2) To what extent are indices of lexical richness in L2 Spanish collinear?

Method

Learner corpus

The learner corpus used for this study was a subset of the Spanish oral data from the longitudinal learner corpus LANGSNAP[1] (Mitchell et al., 2017; Tracy-Ventura et al., 2016). The LANGSNAP corpus includes written and oral data from 27 L2 Spanish learners who spent 9 months abroad. The data were collected at six points over a 21-month period: before departure, three visits during their stay, and two post-SA collection points. At each collection point, each learner completed a written argumentative task, a picture-based oral narrative task, and a semistructured interview, each of which was designed to elicit rich interactive language. In total, the corpus includes 486 texts (303,920 words). There were three prompts for the written argumentative essay and the picture-narrative task, each administered approximately a year apart. Preliminary analyses indicated that there were strong task and prompt effects in the written and oral narrative data, reflecting previous research (Biber & Gray, 2013; Kyle et al., 2016). Therefore, we decided to analyze the oral interviews in this study.

The semistructured oral interviews consisted of preestablished questions related to students' opinions and experiences about their lives abroad, their host family, or language learning. As described in Mitchell et al. (2017), the interviews were designed to elicit a variety of forms and vocabulary from interactive and spontaneous L2 speech samples that could be relevant for the analysis of complexity, accuracy, lexicon, and fluency measures. For example, even though at Visits 2 and 3 the preestablished topic is about students' immediate experiences, participants are also asked to reflect on future plans and what they would miss when returning home. Each of the 27 participants produced one spoken

[1]Available at http://langsnap.soton.ac.uk.

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

Table 1. Collection points, topics, and average words per learner

Collection point	Time	Location	Preestablished topic	Average number of words per interview
Previsit	May 2011	Home	Reasons to study languages	793
Visit 1	Oct 2011	Abroad	Describe the place where you live	2,021
Visit 2	Feb 2012	Abroad	What has happened since the last visit	2,036
Visit 3	May 2012	Abroad	What has happened since the last visit	1,862
Postvisit 1	Oct 2012	Home	How do you feel back home	948
Postvisit 2	Feb 2013	Home	How did your Spanish change while SA	927

text at each collection point, with a total of 162 texts and 254,828 words. The interviews were conducted by a member of the research team and lasted an average of 15 min. The LANGSNAP website provides files with the responses transcribed by the research team. The overview of the corpus and starting topics are shown in Table 1.

The 27 learners were language majors at a university in the United Kingdom where students are required to study abroad during their third year of their undergraduate degree. While abroad, students were exchange students (n = 9), teaching assistants (n = 16), or work interns (n = 2). There were more females (n = 20) than males (n = 7), and most students had English as their L1 (n = 25), except two L1 Polish speakers. Two thirds of the learners (n = 18) spent their year abroad in Spain and the rest (n = 9) in Mexico. Participants' ages at the time of collection varied from 20 to 25, and their mean length for studying Spanish before beginning data collection was 5.5 years (for more information on the participants, design, and collection of LANGSNAP, see Mitchell et al., 2017; Tracy-Ventura et al., 2016). Participants' overall proficiency was measured three times (before departure, after 5 months abroad, and after returning) using the Spanish Elicited Imitation Test (EIT; Ortega, 2000). Participants' proficiency was at an intermediate level at the beginning of the study. The results of a repeated measures analysis of variance showed a significant effect for time and large effect sizes between times (Mitchell et al., 2017), indicating that participants' overall language proficiency improved while abroad and it continued after their return.

Indices of lexical richness

Indices of lexical richness were calculated using a freely available, newly developed tool, TAALES_ES. The tool processes texts using the es-core-news-sm (version 2.1.0) model and Spacy (version 2.1.8). For the analyses in this paper, all words were lemmatized and homographs were distinguished by parts of speech (e.g., noun, verb, etc.). The scripts are freely available at https://github.com/LCR-ADS-Lab/TAALES_ES. Research in Spanish L2 acquisition has pointed out that appropriate verb inflection and the use of the subjunctive mood may signal development for L1 English speakers (Asencion-Delaney & Collentine, 2011; Collentine, 2010; Montrul, 2004; Schnur & Rubio, 2021). Given the rich verb morphology system in Spanish and previous research findings, verb lemmas were distinguished by tense and mood (but not person). This allowed (for example) differentiating between present indicative

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

conjugations of a verb (e.g., *comes* [you eat] and *comemos* [we eat] are represented as *comer_VERB_Ind_Pres*) and subjunctive conjugations (e.g., *comieras* [were you to eat] and *comiera* [were he/I/she to eat] are represented as *comer_VERB_Sub_Imp*). Spacy reports high accuracy for the features used in these analyses (Explosion Al, 2022), including large-grained parts of speech (Noun, Verb, Adjective, etc.; F1 = .982), tense (F1 = .973), and mood (F1 = .965). All corpus-based indices were calculated using a 450-million word subset of the Spanish version of the Corpus of the Web (ESCOW14; Schäfer, 2015; Schäfer & Bildhauer, 2012). The following indices were used to measure lexical richness:

Moving average type token ratio (MATTR)

MATTR is an index of lexical diversity that has been shown to be independent of text length (Covington & McFall, 2010; Zenker & Kyle, 2021). In this study, MATTR is calculated using a moving 50-word window. First, a type token ratio (TTR) is calculated for words 1–50 in an essay, followed by words 2–51, 3–52, and so on until the end of the essay is reached. Final MATTR scores are calculated by averaging the TTR scores for all 50-word windows. It is expected that higher proficiency language users will produce more lexically diverse texts given a particular language production task, which is indexed by higher MATTR scores.

Content word frequency

Content word frequency scores are calculated using all adjectives, adverbs, nouns, and verbs in a text. Mean content word frequency scores are calculated based on the average ESCOW frequency score for content words in a learner text. Traditionally, lower frequency content words (e.g., *chirrido* [squeak], *egregio* [egregious], *innumerables* [countless]) have been considered more difficult and/or less likely to be known by a language learner than more frequent content words (e.g., *hermano* [brother], *bosque* [forest], *verano* [summer]), and they are therefore considered more sophisticated, particularly in written registers.

Verb frequency

Verb frequency scores, which accounted for tense and mood, indicated the mean frequency for verbs based on the ESCOW corpus. Less frequent verbs, such as for instance, indicative past tense conjugations of *desestimar* (e.g., *desestimar_VERB_Ind_Past* [e.g., *desestimé*; I dismissed]), present subjunctive forms of *comprometer* (*comprometer_VERB_Sub_Pres* [e.g., *comprometa*; were I to compromise]), or future indicative forms of *describir* (*desribir_VERB_Ind_Fut* [e.g., *describiré*; I will describe; *describirán*; they will describe]) are considered more sophisticated than verbs that are more frequent, such as the present indicative of the verb *querer* (e.g., *querer_VERB_Ind_Pres* [e.g., *quiero*, I want; *queremos*, we want]), the present indicative of *decir (decir_VERB_Ind_Pres* [e.g., *dice*; she/he says]), or past tense indicative forms of *dar* (*dar_VERB_Ind_Past* [e.g., *di*; I gave]). It is expected that more proficient learners will (on average) use less frequent verb forms. By measuring verb frequency in this way, we do not assume that learners who produce one verb form can produce all tenses. We approach verb inflection as a feature of L2 Spanish development, as less common verb forms may take longer to learn (Asención-Delaney & Collentine, 2011; Montrul, 2004) and may be predictive of proficiency (Schnur & Rubio, 2021).

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

Bigram MI score

Bigram MI scores comprise the mean MI score for bigrams in a learner text. Word combinations that are more exclusive earn higher MI scores (e.g., caer_AUX_Inf-derrotar_ADJ [e.g., *cayendo derrotado*; falling defeated], *platillo volante* [flying saucer], *inversamente proporcional* [inversely proportional), and less exclusive word combinations earn lower MI scores (e.g., solo_ADJ-saber_VERB_Ind_Pres [e.g., *solo sé*; I only know], *cuando ellas* [when they (female)], trabajar_VERB_Inf-a_ADP [e.g., *trabajar a*; to work to]). Previous L2 English research has indicated that more advanced L2 learners tend to use more strongly associated bigrams.

Bigram T score

Bigram T scores comprise the mean T score for bigrams in a learner text. Frequently occurring word combinations tend to earn higher T scores (e.g., *muy bueno* [very good], *una vez* [one time], *en mi* [in my]), whereas less frequently occurring word combinations tend to earn lower T scores (e.g., *similar porque* [similar because], *familiar en* [relative in], leer_VERB_Ind_Pres-a_ADP [e.g., *leo a*; read to]). Previous research in English (e.g., Eguchi & Kyle, 2020; Garner et al., 2019; Granger & Bestgen, 2014) has indicated that more advanced L2 learners tend to use bigrams that earn higher T scores.

Word concreteness

Concreteness scores represent the average concreteness value for words in a text. Concreteness refers to the degree to which a word refers to a perceptible entity. In this study, concreteness scores collected by Guasch et al. (2016) were used. Word such as *abeja* (bee), *manzana* (apple), and *silla* (chair) earn higher concreteness scores, whereas words such as *amargura* (bitterness), *salud* (health), and *suerte* (luck) earn lower concreteness scores. Previous research has suggested that more proficient learners will (on average) use words with lower concreteness scores. Given the relatively small number of words included in the Guasch et al. (2016) study, concreteness scores were available for approximately 20% of the content words in the learner corpus.

Statistical analyses

To analyze the longitudinal data, a series of linear mixed-effects (LME) models were developed using R (R Core Team, 2021) to determine whether indices of lexical richness were predictive of language development as a function of time. This advanced statistical analysis allows us to examine development over time while also considering participants and their individual trajectories (Gries, 2015). The analyses were calculated using R package *lme4* (Bates et al., 2015). In each model, the lexical index (e.g., frequency_CW, MATTR_lemmas) was set as the dependent variable, time as the fixed effect, and participant as a random effect, which uses random intercepts. This model presumes that although the characteristics of each participant's productive lexical use may be at a different starting point before SA, their development would follow similar trajectories and increase or decrease at approximately the same rate. This is the equation used for all models: lmer (lexical_index ~ Time + (1 | Participant), data). The R package *lmerTest* (Kuznetsova et al., 2015) was used to estimate p values. To calculate the effect size of each model, we used the R package *MuMln* (Bartoń, 2020). Both R^2 marginal (R^2_m) and conditional (R^2_c) values are reported. The R^2_m values indicate the amount of the variance explained by the fixed effects alone within the group, whereas R^2_c explains

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

the amount of the variance by both fixed effects and random effects. Finally, the R package *emmeans* (Lenth et al., 2018) was used to obtain estimated marginal means and run post hoc pairwise comparison of the marginal means to identify significant differences among collection points and how time can predict growth in these models. Correlation analyses were conducted to determine the strength of the relationship between lexical indices used in this study. A repository with all files for all the analyses can be found at https://osf.io/atbws/?view_only=1b01cfd8aa3c41e8b7bac87288095757.

Results

Several LME models were conducted to examine the change in the characteristics of productive lexical use during the 21-month of the study period. To measure lexical richness, an index of lexical diversity and two indices of word frequency were calculated. Two indices of n-gram association strength were also measured. Additionally, scores for concreteness were calculated at each collection point. All indices meet the assumption of normality. The score of each lexical index was set as the dependent variable in each of the models, collection points were set as the fixed effects and participants as random effects. The results of each LME model are reported below. Descriptive statistics, visualization of group means and individual trajectories, and post hoc pairwise comparisons between collection points are included. Furthermore, a correlation matrix showing the relationship of all indices is included.

Lexical diversity

An LME model was conducted to investigate the change of lexical diversity as measured by MATTR during the 21 months. Descriptive statistics can be found in Table 2. The group means and individual trajectories are visualized in Figure 1.

The results of the LME model indicated a meaningful and significant relationship (p = .003) between the fixed effects (collection point) and lexical diversity. The model indicated that the fixed effects (collection points) explained 4.14% of the variance in lexical diversity scores (R^2m = .0414). The model also indicated that the combination of fixed and random effects accounted for 64.5% of the variance (R^2c = .6449) in lexical diversity scores, suggesting a high degree of variation across participants. Post hoc pairwise analyses showed a large, significant increase (p = .016, d = -0.897) in lexical diversity scores between Time 1 (predeparture) and Time 6 (9 months after returning to the home country). There is not a significant change in lexical diversity scores during the first 2 months abroad (Time 2). However, a medium, significant increase (p = .047, d = -0.794) was observed between Time 2 (after 2 months abroad) and Time 6. See Table 3 and 4 for a summary of the results and Figure 2 for a visualization of the results.

Table 2. Descriptive statistics for lexical diversity scores across six points

Time	Description	*n*	Mean	*SD*
1	Previsit (May 2011)	27	.708	.025
2	During visit (Oct 2011)	27	.709	.023
3	During visit (Feb 2012)	27	.716	.028
4	During visit (May 2012)	27	.719	.029
5	After return (Oct 2012)	27	.720	.027
6	After return (Feb 2013)	27	.722	.022

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

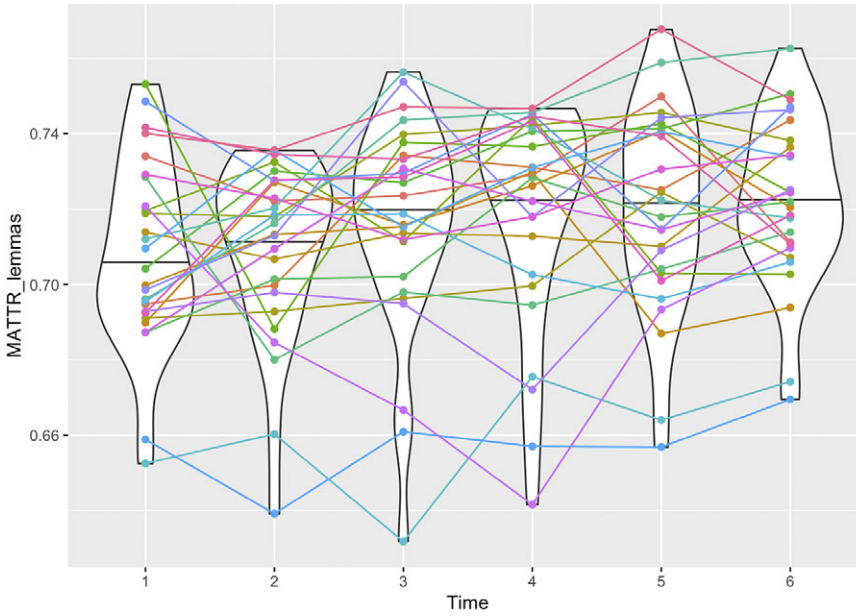

Figure 1. Visualization of group means and individual trajectories.

Table 3. Selected pairwise results for lexical diversity

Contrast 1	Contrast 2	estimate	*SE*	df	*t*	*p*	*d*
Time 1	Time 4	−0.011	0.004	130	−2.662	.090	−0.725
Time 1	Time 5	−0.012	0.004	130	−2.880	.052	−0.784
Time 1	Time 6	−0.014	0.004	130	−3.296	.016	−0.897

Table 4. Adjacent pairwise results for lexical diversity

Contrast 1	Contrast 2	estimate	*SE*	df	*t*	*p*	*d*
Time 1	Time 2	−0.002	0.004	130	−0.380	1.00	−0.103
Time 2	Time 3	−0.007	0.004	130	−1.577	.615	−0.439
Time 3	Time 4	−0.003	0.004	130	−0.705	.981	0.251
Time 4	Time 5	−0.001	0.004	130	−0.218	1.00	−0.059
Time 5	Time 6	−0.002	0.004	130	−0.416	1.00	0.113

Lexical sophistication

Content word frequency

Descriptive statistics for content word frequency can be found in Table 5 and are visualized in Figure 3.

The results of the LME model indicated a meaningful and significant relationship ($p < .001$) between collection point and content word frequency scores. The model indicated that the fixed effects (collection points) explained 17.4% of the variance in content word frequency scores (R^2m = .174). The model also indicated that the combination of fixed and random effects explained 59.9% of the variance (R^2c = .599),

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

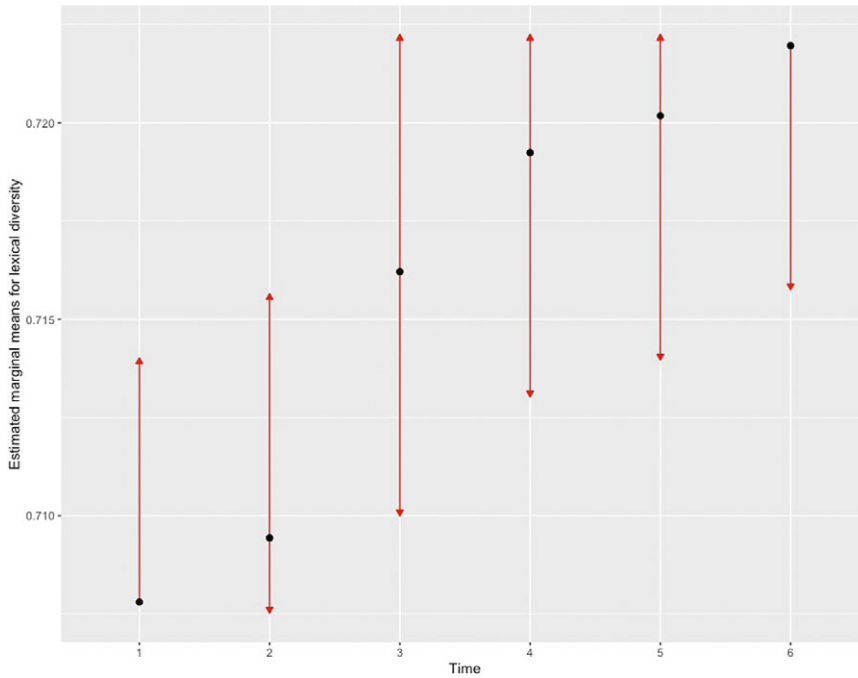

Figure 2. Visualization of pairwise comparisons for lexical diversity. Overlapping red lines indicate that comparisons are not significant ($p > .05$).

Table 5. Descriptive statistics for content word frequency scores across six points

Time	Description	*n*	Mean	*SD*
1	Previsit (May 2011)	27	10.929	0.351
2	During visit (Oct 2011)	27	11.192	0.252
3	During visit (Feb 2012)	27	11.288	0.213
4	During visit (May 2012)	27	11.247	0.195
5	After return (Oct 2012)	27	11.280	0.287
6	After Return (Feb 2013)	27	11.241	0.288

suggesting a high degree of variation across participants. Post hoc pairwise analyses showed a large, significant increase ($p < .001$, $d = -1.399$) in content word frequency scores between Time 1 (predeparture) and Time 2 (2 months in country). This increase remained during the remainder of the study period, but no further significant increases or decreases were observed after Time 2. See Table 6 and 7 for a summary of the results and Figure 4 for a visualization.

Verb frequency

An LME model was conducted to investigate the degree to which the average frequency of verbs changed during the study period. Descriptive statistics can be found in Table 8 and are visualized in Figure 5.

The results indicated a meaningful and significant relationship ($p < .001$) between collection point and verb frequency scores. The model indicated that the fixed effects

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

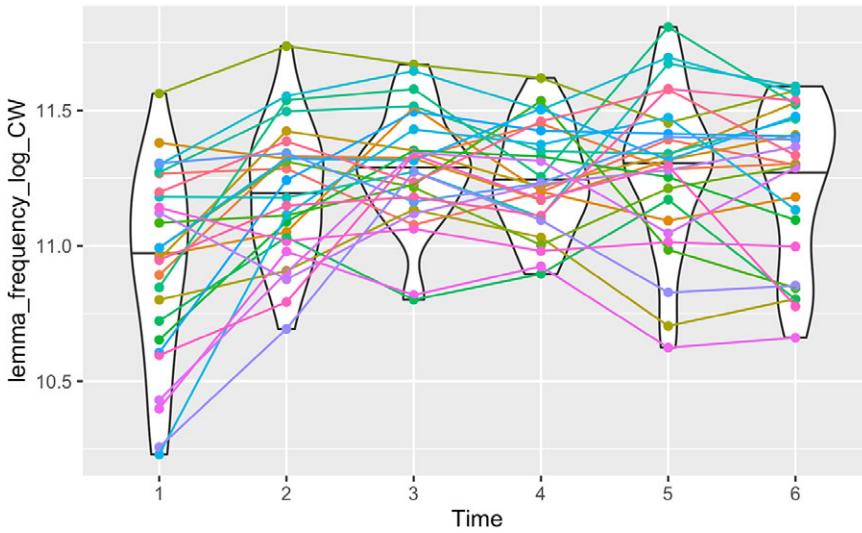

Figure 3. Visualization of group means and individual trajectories.

Table 6. Selected pairwise results for content word frequency

Contrast 1	Contrast 2	estimate	*SE*	df	*t*	*p*	*d*
Time 1	Time 4	−0.318	0.051	130	−6.233	<.0001	−1.696
Time 1	Time 5	−0.351	0.051	130	−6.874	<.0001	−1.871
Time 1	Time 6	−0.311	0.051	130	−6.100	<.0001	−1.660

Table 7. Adjacent pairwise results for content word frequency

Contrast 1	Contrast 2	estimate	*SE*	df	*t*	*p*	*d*
Time 1	Time 2	−0.262	0.051	130	−5.140	<.0001	-1.399
Time 2	Time 3	−0.096	0.051	130	−1.879	.420	-0.511
Time 3	Time 4	0.040	0.051	130	0.786	.969	0.214
Time 4	Time 5	−0.033	0.051	130	−0.641	.988	-0.174
Time 5	Time 6	0.040	0.051	130	0.774	.971	0.211

(collection points) explained 10.6% of the variance in verb frequency scores (R^2m = .106). The model also indicated that the combination of fixed and random effects accounted for 49.4% of the variance (R^2c = .494) in verb frequency scores, suggesting a high degree of individual variation. Post hoc pairwise analyses indicated a large, significant decrease ($p < .001$, $d = 1.321$) in verb frequency scores between Time 1 (predeparture) and Time 4 (after 9 months abroad). This decrease remained significant ($p < .05$, $d = 0.923$) at Time 5 (after 5 months back home). The analysis also indicated a large, significant decrease ($p < .001$, $d = 1.228$) between Time 2 (after 2 months abroad) and Time 4 (after 9 months abroad), which remained significant ($p = .032$, $d = .830$) at Time 5. See Table 9 and 10 for a summary of the results and Figure 6 for a visualization.

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

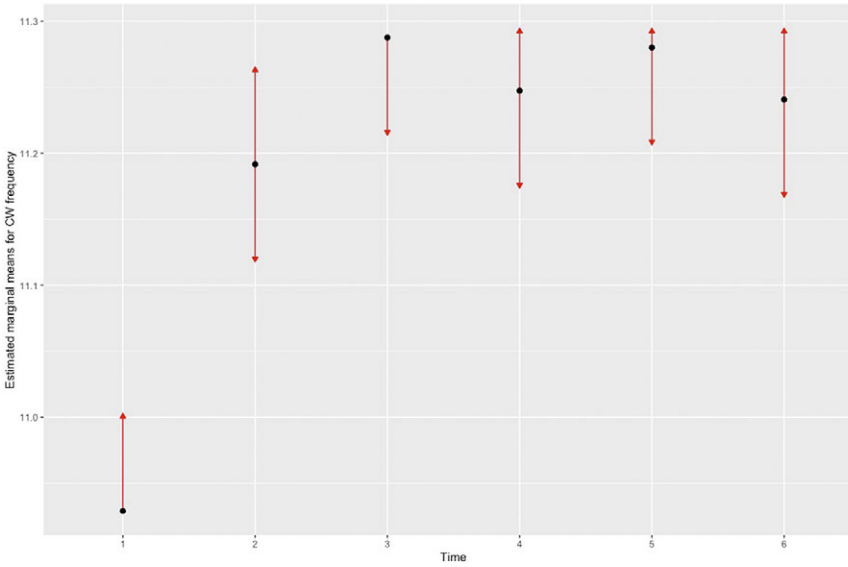

Figure 4. Visualization of pairwise comparisons for content word frequency.

Table 8. Descriptive statistics for verb frequency scores across six points

Time	Description	*n*	Mean	*SD*
1	Previsit (May 2011)	27	10.8	.245
2	During visit (Oct 2011)	27	10.8	.216
3	During visit (Feb 2012)	27	10.7	.291
4	During visit (May 2012)	27	10.6	.320
5	After return (Oct 2012)	27	10.6	.303
6	After return (Feb 2013)	27	10.7	.263

N-gram association strength

Mutual information

Descriptive statistics for MI scores can be found in Table 11 and are visualized in Figure 7.

The results indicate a meaningful and significant relationship (p = .009) between collection point and MI scores. The model indicated that the fixed effects (collection points) explained 4.7% of the variance in MI scores (R^2m = .047), and that the combination of fixed and random effects explained 52.8% of the variance in MI scores, (R^2c = .528), suggesting a high degree of variation across participants. Post hoc pairwise analyses showed a large, significant decrease (p = .035; d = .824), between Time 1 (previsit) and Time 2 (the first 2 months abroad), which remained significant until Time 3 (5 months after being abroad; p = .036; d = .821). No further significant increase or decrease was observed for the rest of the study. See Table 12 and 13 for a summary of the results and Figure 8 for a visualization.

T scores

Descriptive statistics for T scores can be found in Table 14 and visualized in Figure 9.

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

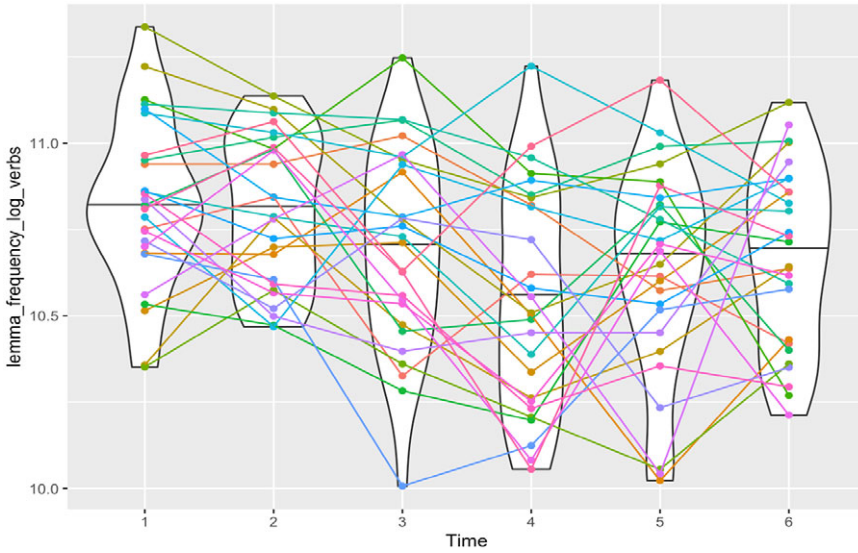

Figure 5. Visualization of group means and individual trajectories.

Table 9. Selected pairwise results for verb frequency

Contrast 1	Contrast 2	estimate	*SE*	df	*t*	*p*	*d*
Time 1	Time 4	0.274	0.056	130	4.854	<.0001	1.321
Time 1	Time 5	0.191	0.056	130	3.390	.012	0.923
Time 1	Time 6	0.148	0.056	130	2.629	.097	0.716

Table 10. Adjacent pairwise results for verb frequency

Contrast 1	Contrast 2	estimate	*SE*	df	*t*	*p*	*d*
Time 1	Time 2	0.019	0.056	130	0.341	1.000	0.093
Time 2	Time 3	0.106	0.056	130	1.876	.421	0.511
Time 3	Time 4	0.149	0.056	130	2.637	.096	0.718
Time 4	Time 5	−0.083	0.056	130	−1.464	.687	−0.398
Time 5	Time 6	−0.043	0.056	130	−0.761	.973	−0.207

The results indicate a nonsignificant relationship ($p = .107$) between collection point and T scores. The model indicated that the fixed effects (collection points) explained 2.7% of the variance in T scores (R^2m = .027), and the combination of fixed and random effects explained 53.2% of the variance (R^2c = .532).

Psycholinguistic Norms

Concreteness

Descriptive statistics for concreteness scores can be found in Table 15 and are visualized in Figure 10.

The results of the model indicate a meaningful and significant relationship ($p < .001$) between collection point and concreteness scores. The model indicated that the fixed

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

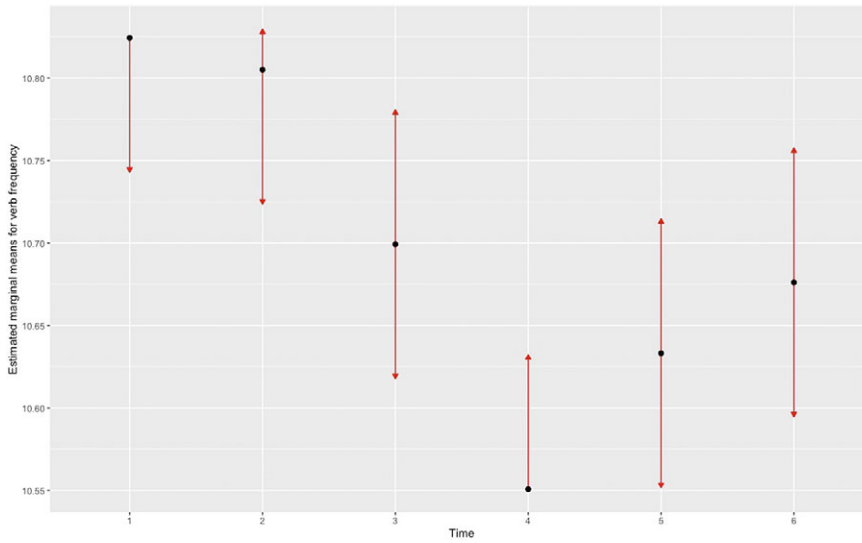

Figure 6. Visualization of pairwise comparisons for verb frequency.

Table 11. Descriptive statistics for bigram MI scores across six points

Time	Description	n	Mean	*SD*
1	Previsit (May 2011)	27	1.37	.111
2	During visit (Oct 2011)	27	1.30	.110
3	During visit (Feb 2012)	27	1.30	.118
4	During visit (May 2012)	27	1.35	.097
5	After return (Oct 2012)	27	1.31	.145
6	After return (Feb 2013)	27	1.32	.132

effects (collection points) explained 19.7% of the variance in concreteness scores (R^2m = .197) and the combination of fixed and random effects explained 61.1% of the variance R^2c = .611), suggesting a high degree of variation across participants. Post hoc pairwise analyses showed a large, significant increase (p =.002, d = -1.923) between Time 1 (predeparture), and Time 2 (after 2 months abroad). The increase remained significant (p < .001, d = -1.717) until Time 4 (after 9 months abroad). A large, significant decrease (p =.022, d = 0.865) in concreteness scores was observed between Time 4 (the last month abroad) and Time 5 (after being home for 5 months) which remained significant until Time 6 (after 9 months back home). See Table 16 and 17 for a summary of the results and Figure 11 for a visualization of the results.

Correlation analyses

To determine the relationships between indices of L2 Spanish lexical richness used in this study, correlation analyses were conducted. A correlation matrix with the correlation coefficients can be found in Table 18.

The results from the correlational analyses indicate that most indices were not strongly correlated with each other, with a few exceptions. The results show a large,

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

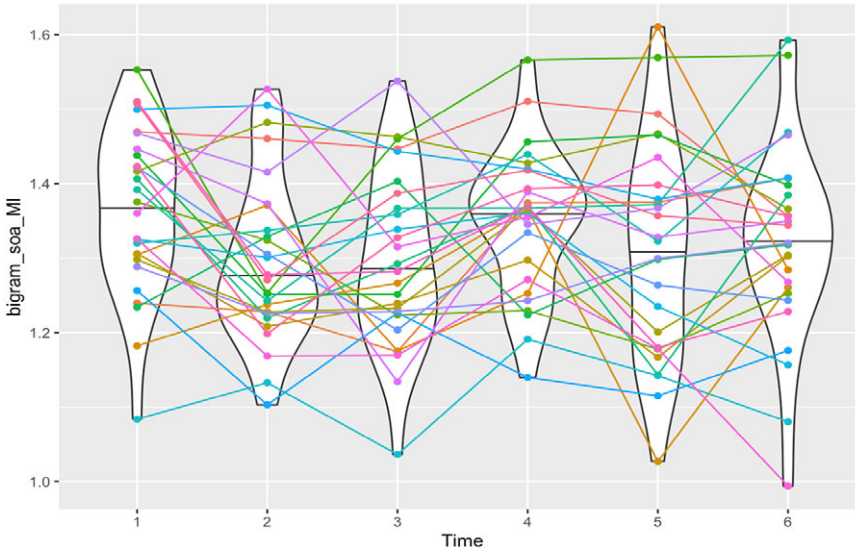

Figure 7. Visualization of group means and individual trajectories.

Table 12. Selected pairwise results for bigram MI

Contrast 1	Contrast 2	estimate	*SE*	df	*t*	*p*	*d*
Time 1	Time 4	0.014	0.023	130	0.611	.990	0.166
Time 1	Time 5	0.055	0.023	130	2.394	.166	0.652
Time 1	Time 6	0.044	0.023	130	1.927	.390	0.525

Table 13. Adjacent pairwise results for bigram MI

Contrast 1	Contrast 2	estimate	*SE*	df	*t*	*p*	*d*
Time 1	Time 2	0.069	0.023	130	3.027	.035	0.824
Time 2	Time 3	0.000	0.023	130	−0.012	1.000	−0.003
Time 3	Time 4	−0.055	0.023	130	−2.404	.162	−0.654
Time 4	Time 5	0.041	0.023	130	1.784	.480	0.485
Time 5	Time 6	−0.010	0.023	130	−0.467	.997	−0.127

positive correlation between T scores and MI scores ($r = .711$). There is also a medium, positive correlation between MATTR and T scores ($r = .523$), and between MATTR and MI scores (I = .483).

Discussion

In this study we investigated the development of lexical richness in L2 oral interviews across multiple dimensions of lexical use using advanced natural language processing tools. First, the results suggest that lexical diversity (as measured by MATTR) sees meaningful growth over the 21-month period. It appears that participants' spoken lexicon slowly increases while being abroad and that it continues after returning home. The findings indicate that it may take some time to incorporate a wider variety of words

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

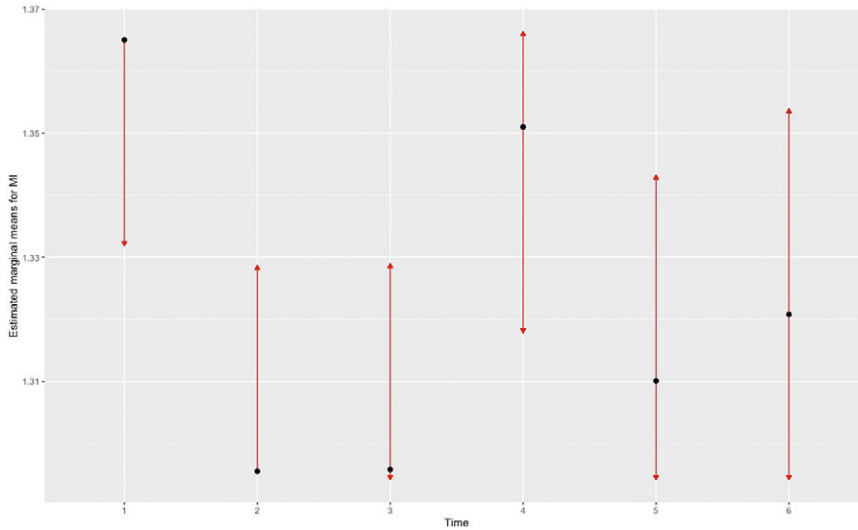

Figure 8. Visualization of pairwise comparisons for bigram MI.

Table 14. Descriptive statistics for bigram T scores across six points

Time	Description	*n*	Mean	*SD*
1	Previsit (May 2011)	27	35.8	35.2
2	During visit (Oct 2011)	27	33.9	33.9
3	During visit (Feb 2012)	27	28.5	31.6
4	During visit (May 2012)	27	35.6	30.0
5	After return (Oct 2012)	27	18.3	42.2
6	After return (Feb 2013)	27	30.1	42.6

in spontaneous speech, as MATTR values only reach significance after participants have been back home for 9 months. Although the results support some of the previous findings of lexical diversity being a good indicator of L2 development (Jarvis, 2017; Mora & Valls-Ferrer, 2012; Serrano et al., 2012; Tavakoli, 2018), this study adds further insight into how advanced learners develop their spontaneous spoken lexicon and that the growth in their vocabulary may take time to show. A related study that used same learner corpus but unlemmatized lexical items and a different index of lexical diversity (Mitchell et al., 2017) had somewhat divergent findings. Mitchell et al. found increases in lexical diversity between predeparture interviews and all times abroad, after which lexical diversity scores decreased. There are at least two issues that warrant further exploration. The first is the degree to which operationalization of lexical items (lemmatized versus unlemmatized orthographic forms) affects measurements of diversity in Spanish L2 texts (see Jarvis & Hashimoto, 2021, for some explorations into this issue for L2 English texts). The second issue is operationalization of lexical diversity. In this study MATTR was used, which has been demonstrated to be particularly independent of text length (Vidal & Jarvis, 2020; Zenker & Kyle, 2021). Mitchell et al. (2017) used *D* to index lexical diversity. The results of research concerning *D* have been mixed, with some studies finding a positive relationship between *D* and text length (Koizumi &

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

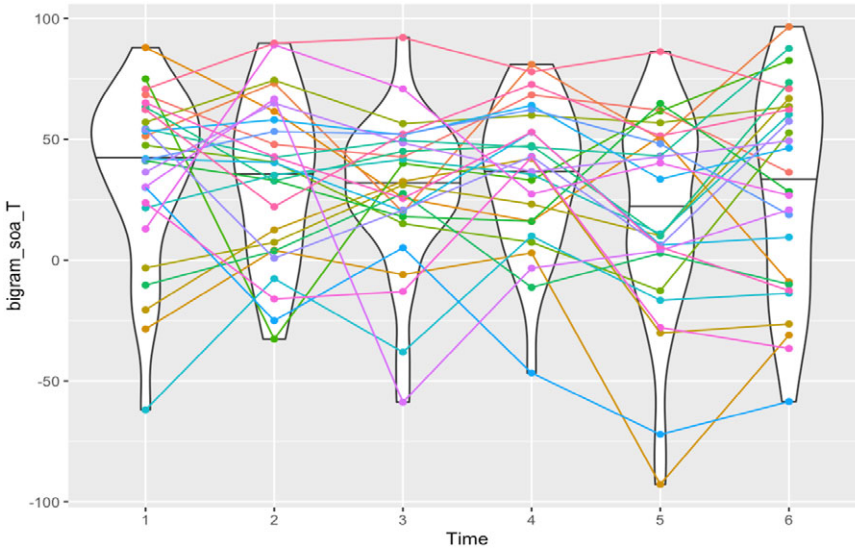

Figure 9. Visualization of group means and individual trajectories.

Table 15. Descriptive statistics for concreteness across six points

Time	Description	*n*	Mean	*SD*
1	Previsit (May 2011)	27	3.53	.249
2	During visit (Oct 2011)	27	4.00	.325
3	During visit (Feb 2012)	27	3.97	.431
4	During visit (May 2012)	27	3.95	.398
5	After return (Oct 2012)	27	3.74	.322
6	After return (Feb 2013)	27	3.68	.354

In'nami, 2012), which suggests that *D* conflates text length and diversity. This may help to explain the differences found between the two studies.

The results also suggest that lexical sophistication, as measured by mean content word frequency, changes meaningfully during the first 2 months abroad. More time in the country does not appear to affect content word frequency, but change does appear to be durable, as content word frequency values do not significantly decrease after returning home. As participants travel abroad and advanced in their proficiency, they incorporate more frequent words into their spontaneous speech. These results are mostly consistent with previous findings of spoken lexical use (Crossley et al., 2010; Eguchi & Kyle, 2020; Kyle & Crossley, 2015; Tracy-Ventura, 2017; cf. Bardel et al., 2012). It may be possible that as learners spend time in the host country, they begin to use more frequent words as they move away from the textbook language they most likely had experienced in their classrooms. These results diverge from the findings of Tracy-Ventura (2017), who measured lexical sophistication in the LANGSNAP corpus but combined written and spoken tasks at each collection point. Using frequency bands for all words across two points (predeparture and end of stay), Tracy-Ventura found that participants used more low-frequency words by the end of their stay than at predeparture. The observed differences between the two studies suggest that register

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

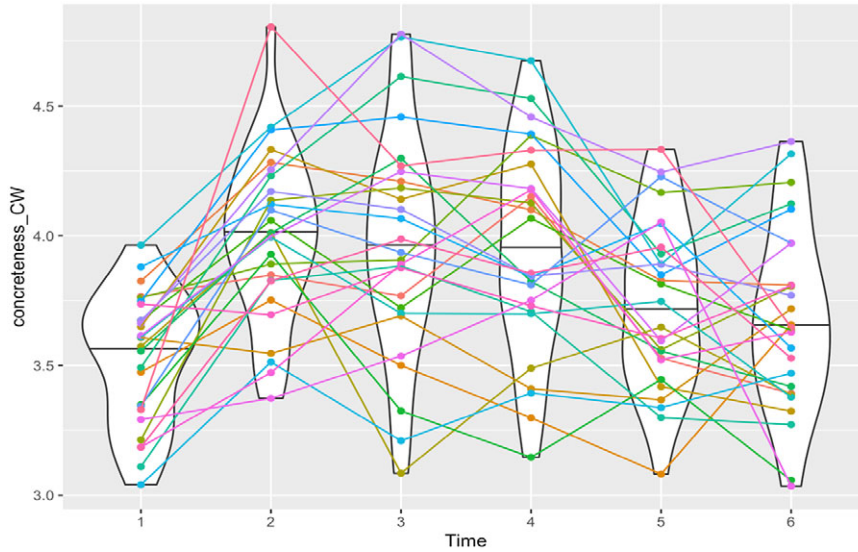

Figure 10. Visualization of group means and individual trajectories.

Table 16. Selected pairwise results for concreteness scores

Contrast 1	Contrast 2	estimate	*SE*	df	*t*	*p*	*d*
Time 1	Time 4	−0.420	0.067	130	−6.309	<.001	−1.717
Time 1	Time 5	−0.208	0.067	130	−3.132	.026	−0.852
Time 1	Time 6	−0.150	0.067	130	−2.255	.221	−0.614

Table 17. Adjacent pairwise results for concreteness scores

Contrast 1	Contrast 2	estimate	*SE*	df	*t*	*p*	*d*
Time 1	Time 2	−0.470	0.067	130	−7.065	<.001	−1.923
Time 2	Time 3	0.032	0.067	130	0.477	.997	0.130
Time 3	Time 4	0.019	0.067	130	0.279	1.000	0.076
Time 4	Time 5	0.211	0.067	130	3.177	.022	0.865
Time 5	Time 6	0.058	0.067	130	0.878	.951	0.239

differences play an important role in the measurement of lexical sophistication. Although previous research has found relatively similar results across band-based and mean-based frequency norms (Crossley et al., 2013), differences in operationalizations may have also contributed to differences across studies.

In contrast to content word frequency, verb frequency values decreased by the end of the SA period, indicating that learners incorporated more sophisticated verbs. The change seems durable, as it remains significant 5 months after returning home. The results suggest that learners need to spend several months abroad until a decrease in verb frequency values reaches significance, but that the change is long-lasting. As participants spend more time abroad and advanced in their L2, they are able to produce more infrequent verbs forms. For example, the present tense of indicative of the verb "to have," *tengo* ([I have], *tener_VERB_Ind_Pres*; 2159.445 per million), would receive a

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

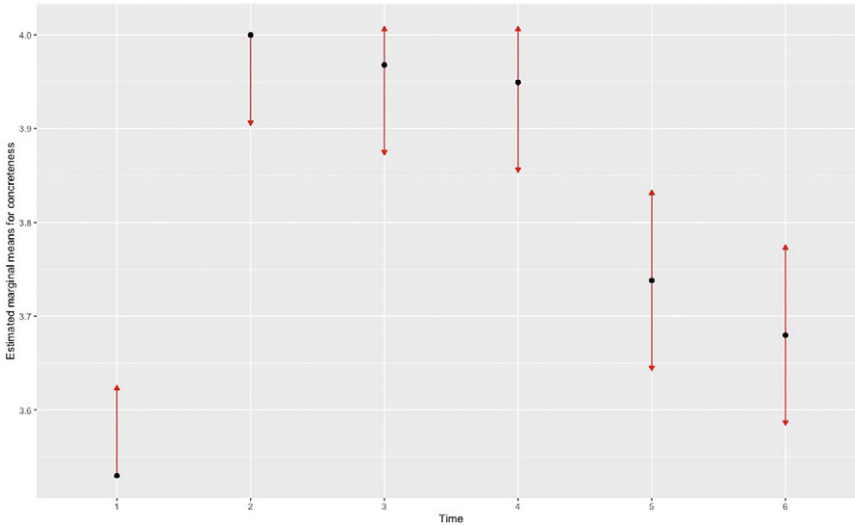

Figure 11. Visualization of pairwise comparisons for concreteness scores.

Table 18. Correlation matrix of lexical indices

	MATTR	CW frequency	Verb frequency	Concreteness	T scores
MATTR					
CW frequency	.034				
Verb frequency	−.131	.312			
Concreteness	−.281	.286	.124		
T scores	.523	−.023	.111	−.014	
MI scores	.483	−.082	.326	−.187	.711

higher frequency score than the subjunctive imperfect form, *tuviera* ([were I to have] tener_VERB_Sub_Imp, 34.683 per million). The findings suggest that the use of more infrequent verb forms may be indicative of advanced spoken L2 Spanish. This implies that certain verb forms in Spanish, such as the past subjunctive, may take longer to learn and may be harder to retrieve in spontaneous speech than other forms. These results shed light into how low-frequency verbs in Spanish may be a good predictor L2 development (Schnur & Rubio, 2021), and that the relationship between lexis and grammar needs to be considered when examining certain features of lexical richness.

Recent research has highlighted the importance of measuring collocation use as a multifaceted construct. In this study, we analyzed two indices that measure the association strength between two words—namely, MI and T score. The results suggest that MI values significantly decrease during the first 2 months of the stay abroad, and this decrease remains significant 5 months later. By the end of the 9 months abroad, participants start using bigrams with higher MI scores again, although the increase is not statistically significant.

These findings may be related to the formulaic language used in a classroom setting that uses highly exclusive n-grams or collocations composed of low-frequency words. As participants get immersed in the target language, they produce a greater proportion of low-MI-scored bigrams. It could be possible that when participants move abroad, the

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

everyday language required from them is not characterized by bigrams formed by infrequent words, as it could be in their previous classroom setting. However, by the end of the stay abroad, bigrams with high MI scores increase. Participants may need more time abroad to incorporate bigrams composed of lower frequency words into their lexical repertoire. Examples of collocations in the learner corpus that receive high MI scores are *habla hispana* ("Hispanic language," 8.8144), *hablante nativo* ("native speaker," 8.26333) or *madres solteras* ("single mothers," 7.695), whereas the collocations *un hablante* ("a speaker," 0.77765) or *las madres* ("the mothers," 1.45042) would receive a lower MI score. As their vocabulary grows abroad, so does the use of highly exclusive collocations. The results show no significant change in T scores during the study. High T score values tend to highlight bigrams composed of frequent words, yet the production of collocations composed of frequent items (high T scores) does not seem to be predictive of development in this study.

The results of these SOA norms partly align with previous research, in which more advanced learners' texts will generally include bigrams with higher MI scores, (Ellis et al., 2008; Gablasova et al., 2017; Garner et al., 2019; Granger & Bestgen, 2014; Kyle et al., 2018; Paquot, 2019). Even though there is a significant decrease of MI scores the first few months abroad, they end up increasing later on throughout the study. There is no change in T scores, unlike what previous studies have found. More research on collocational use of L2 Spanish is needed to better interpret these findings. In particular, it may be useful to investigate bigrams with particular parts of speech (Bestgen & Granger, 2014) and/or dependency bigrams (Kyle & Eguchi, 2021; Paquot, 2019). Nevertheless, this study highlights the importance of a multidimensional approach to collocation use and other indices of lexical sophistication of L2 Spanish.

Studies that have examined psycholinguistic word information investigate the extent to which L1 norms can predict L2 spoken lexical proficiency (Salsbury et al., 2011). These psycholinguistic norms are related to processing, saliency, retrieval, and learnability of a word. The results of this study show that concreteness scores changed meaningfully during the study. Participants' use of concrete words increased after 2 months abroad, which remained significant until the end of their stay abroad. However, as they return home, participants started using fewer concrete words in comparison to the words produced after arrival.

A significant decrease in concreteness scores was observed 5 months after their return home. However, the results suggest that the use of more concrete words throughout the study period is still meaningful, as evidenced by the fact that concreteness scores were significantly higher 5 months after returning home than before departure.

These findings suggest that the use of more salient (more concrete) words may be an indicator of L2 spoken development. However, these changes are not durable after returning home. These findings differ from previous research using L2 corpora, where proficient speech is generally characterized by less salient and more difficult to retrieve lexical items (Crossley et al., 2016; Crossley & Skalicky, 2019; Eguchi & Kyle, 2020; Kyle et al., 2016; Salsbury et al., 2011). A possible explanation could be linked to the oral corpus used in this study, which is conversational in nature, unlike studies using controlled tasks or higher stakes tasks, such as those in testing settings. It may be that being immersed in the language in comparison with taking foreign language courses at a university might affect word processing as well.

As a whole, these findings show how several measures of lexical richness may be indicative of oral L2 development. A valuable finding shown in this study is that development of vocabulary is not linear and that the change in some indices is more durable than others. The results suggest that as learners advanced in their L2 oral skills,

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

they tended to use a more varied vocabulary, more frequent content words but infrequent verb forms. The correlation analyses suggest that there is not a strong relationship between most measures of lexical richness but that bigram SOA measures (T scores and MI scores) are strongly correlated. This finding is not necessarily surprising, given that the indices measure related aspects of the same subconstruct. Lexical diversity also appears to be correlated to SOA indices. However, studies of L2 Spanish collocational features are rare. Research on the development of Spanish productive collocational use in a variety of written and spoken registers is clearly needed to better interpret these findings.

As suggested by the findings, the immersion in the country of residence compared with being at the home university taking advanced language courses (Time 5 and Time 6), considering everything that may influence L2 use in each setting (e.g., host family, social life, exposure, formal instruction), may play a role on the characteristics of participants' productive lexical use, especially in oral conversations. It appears that the change of setting from home university context to immersion might affect certain features of productive vocabulary more than it affects others, especially for association strength indices and psycholinguistic word norms. Future research should explore differences in spontaneous speech development in an instructed foreign language setting and SA programs (Collentine, 2004; Segalowitz & Freed, 2004).

Conclusion

This study examined six dimensions of lexical richness of L2 Spanish using a longitudinal spoken learner corpus. We measured a series of frequency-based indices, association strength measures, and psycholinguistic norms that have been found to be representative of advanced L2 proficiency or lexical use. For the analysis, we used TAALES_ES, a tool for the automatic analysis of lexical sophistication in Spanish text that will allow for replicable analysis in Spanish learner corpora. The results on lexical diversity, frequency, and to a certain degree bigram association strength support previous findings on the characteristics of productive vocabulary use. The results regarding psycholinguistic word norms differ to some extent with those from past studies. However, most studies have investigated written and cross-sectional corpora. Research on longitudinal and oral corpora is needed to understand actual L2 development of spontaneous speech. The reported findings require additional evidence of L2 Spanish and oral corpora studies to provide support to understand how speech develops over time.

The current study has pedagogical implications for teaching L2 speaking and, in particular, for vocabulary instruction sensitive to register. Because advanced spontaneous speech may be characterized by features that differ from written proficiency and other spoken registers (e.g., the use of higher versus lower frequency words), teachers may consider implementing awareness-raising tasks that highlight the type of vocabulary that is appropriate for different types of formal and informal registers (e.g., everyday conversation versus work-place small talk). This may require, for instance, emphasizing the teaching of highly exclusive collocations or highlighting the importance of frequent and salient words during spontaneous everyday conversational tasks. Furthermore, facilitating tasks where students are pushed to produce certain verb forms that take longer to learn but may be associated with more advanced speech could be helpful to automatize the processes behind verb conjugations in L2 Spanish.

A limitation of this study is that the psycholinguistic word lists accounted for only 20% of the content words in the learner texts, so the results should be taken with caution, as these just provide a first look into the psycholinguistic properties of words of

https://doi.org/10.1017/S0272263123000384 Published online by Cambridge University Press

L2 Spanish. Additionally, when calculating average verb frequency, we only tagged for tense and mood. Analyzing accuracy and agreement of verb forms, as well as other aspects of lexico-grammar such as gender agreement, would reveal a more in-depth picture of the characteristics of L2 Spanish productive vocabulary use.

Future research should also explore individual development. The LME models show that much of the variance was explained at the individual level, which could indicate that individuals progress at different rates and follow different paths. Taking a dynamic systems perspective (Cameron & Larsen-Freeman, 2007) to study productive lexical use would give us insight into the complexities of a learner's interlanguage development.

Data availability statement. The experiment in this article earned an Open Data badge for transparent practices. The materials are available at https://osf.io/atbws/?view_only=1b01cfd8aa3c41e8b7bac87288095757

References

Appel, R., Trofimovich, P., Saito, K., Isaacs, T., & Webb, S. (2019). Lexical aspects of comprehensibility and nativeness from the perspective of native-speaking English raters. *ITL—International Journal of Applied Linguistics*, *170*, 24–52. https://doi.org/10.1075/itl.17026.app

Asencion-Delaney, Y., & Collentine, J. (2011). A multidimensional analysis of a written L2 Spanish corpus. *Applied Linguistics*, *32*, 299–322. https://doi.org/10.1093/applin/amq053

Bardel, C., Gudmundson, A., & Lindqvist, C. (2012). Aspects of lexical sophistication in advanced learners' oral production. *Studies in Second Language Acquisition*, *34*, 269–290. https://doi.org/10.1017/S0272263112000058

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Berger, C. M., Crossley, S. A., & Kyle, K. (2019). Using native-speaker psycholinguistic norms to predict lexical proficiency and development in second-language production. *Applied Linguistics*, *40*, 22–42. https://doi.org/10.1093/applin/amx005

Berton, M. (2020). *Riqueza léxica y expresión escrita en aprendices suecos de ELE: Proficiencia general, competencia léxica pasiva, tipo y complejidad de la tarea*. Department of Romance Studies and Classics, Stockholm University.

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*, 28–41. https://doi.org/10.1016/j.jslw.2014.09.004

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber, D., & Conrad, S. (2019). *Register, Genre, and Style* (2nd ed.). Cambridge University Press.

Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus. TOEFL monograph series.* Educational Testing Service. http://www.ets.org/Media/Research/pdf/RM-04-03.pdf

Biber, D., & Gray, B. (2013). Discourse characteristics of writing and speaking task types on the *TOEFL iBT*® test: A lexico-grammatical analysis. *ETS Research Report Series*, *2013*, i–128. https://doi.org/10.1002/j.2333-8504.2013.tb02311.x

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904–911. https://doi.org/10.3758/s13428-013-0403-5

Cameron, L., & Larsen-Freeman, D. (2007). Complex systems and applied linguistics. *International Journal of Applied Linguistics*, *17*, 226–240. https://doi.org/10.1111/j.1473-4192.2007.00148.x

Castañeda-Jimenez, G., & Jarvis, S. (2014). Exploring lexical diversity in second language Spanish. In K. L. Geeslin (Ed.), *Handbook of Spanish second language acquisition* (pp. 498–513). Wiley.

Collentine, J. (2004). The effects of learning contexts on morphosyntactic and lexical development. *Studies in Second Language Acquisition*, *26*, 227–248. https://doi.org/10.1017/S0272263104262040

Collentine, J. (2010). The acquisition and teaching of the Spanish subjunctive: An update on current findings. *Hispania*, *93*, 39–51.

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTRC). *Journal of Quantitative Linguistics*, *17*, 94–100. https://doi.org/10.1080/09296171003643098

Crossley, S. A., Clevinger, A., & Kim, Y. (2014). The role of lexical properties and cohesive devices in text integration and their effect on human ratings of speaking proficiency. *Language Assessment Quarterly*, *11*, 250–270. https://doi.org/10.1080/15434303.2014.926905

Crossley, S. A., Cobb, T., & McNamara, D. S. (2013). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System*, *41*, 965–981. https://doi.org/10.1016/j.system.2013.08.002

Crossley, S. A., Kyle, K., & Salsbury, T. (2016). A usage-based investigation of L2 lexical acquisition: The role of input and output. *The Modern Language Journal*, *100*, 702–715. https://doi.org/10.1111/modl.12344

Crossley, S. A., & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology*, *17*, 171–192.

Crossley, S. A., Salsbury, T., & McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers: Polysemy and frequency use in English L2 speakers. *Language Learning*, *60*, 573–605. https://doi.org/10.1111/j.1467-9922.2010.00568.x

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2013). Validating lexical measures using human scores of lexical proficiency. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 135–156). John Benjamins.

Crossley, S. A., Salsbury, T., & Mcnamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, *36*, 570–590. https://doi.org/10.1093/applin/amt056

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011a). What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly*, *45*, 182–193. https://doi.org/10.5054/tq.2010.244019

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011b). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*, 561–580. https://doi.org/10.1177/0265532210378031

Crossley, S. A., & Skalicky, S. (2019). Examining lexical development in second language learners: An approximate replication of Salsbury, Crossley & McNamara (2011). *Language Teaching*, *52*, 385–405. https://doi.org/10.1017/S0261444817000362

Crossley, S. A., Skalicky, S., Kyle, K., & Monteiro, K. (2019). Absolute frequency effects in second language lexical acquisition. *Studies in Second Language Acquisition*, *41*, 721–744. https://doi.org/10.1017/S0272263118000268

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2018). Linguistic dimensions of L2 accentedness and comprehensibility vary across speaking tasks. *Studies in Second Language Acquisition*, *40*, 443–457. https://doi.org/10.1017/S027226311700016X

Daller, H., van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, *24*, 197–222. https://doi.org/10.1093/applin/24.2.197

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL—International Review of Applied Linguistics in Language Teaching*, *47*, Article 007. https://doi.org/10.1515/iral.2009.007

Eguchi, M., & Kyle, K. (2020). Continuing to explore the multidimensional nature of lexical sophistication: The case of oral proficiency interviews. *The Modern Language Journal*, *104*, 381–400.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*, 143–188. https://doi.org/10.1017/S0272263102002024

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and tesol. *TESOL Quarterly*, *42*, 375–396. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*, 139–155.

Explosion Al. (2022). *Spacy models* [Computer software]. Github. https://github.com/explosion/spacy-models/blob/master/meta/es_core_news_sm-3.4.0.json

Fernández Vítores, D. (2022). *El español: Una lengua viva* (No. 110-21-045–5). Instituto Cervantes. https://cvc.cervantes.es/lengua/espanol_lengua_viva/

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence: Collocations in corpus-based language learning research. *Language Learning*, *67*, 155–179. https://doi.org/10.1111/lang.12225

Garner, J., Crossley, S. A., & Kyle, K. (2019). N-gram measures and L2 writing proficiency. *System*, *80*, 176–187. https://doi.org/10.1016/j.system.2018.12.001

Garner, J., Crossley, S. A., & Kyle, K. (2020). Beginning and intermediate L2 writer's use of n-grams: An association measures study. *International Review of Applied Linguistics in Language Teaching*, *58*, 51–74. https://doi.org/10.1515/iral-2017-0089

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*, 193–202. https://doi.org/10.3758/BF03195564

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching*, *52*, 229–253. https://doi.org/10.1515/iral-2014-0011

Gries, S. Th. (2013). 50-something years of work on collocations: What is or should be next … *International Journal of Corpus Linguistics*, *18*, 137–166. https://doi.org/10.1075/ijcl.18.1.09gri

Gries, S. Th. (2015). Statistics for learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 159-182). Cambridge University Press. https://doi.org/10.1017/cbo9781139649414.008

Guasch, M., Ferré, P., & Fraga, I. (2016). Spanish norms for affective and lexico-semantic variables for 1,400 words. *Behavior Research Methods*, *48*, 1358–1369. https://doi.org/10.3758/s13428-015-0684-y

Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique [Problems and methods of linguistic statistics]*. Reidel.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*, 218–238. https://doi.org/10.1016/j.asw.2013.05.002

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

Jarvis, S. (2013a). Defining and measuring lexical diversity. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (Vol. 47, pp. 13–44). John Benjamins. https://doi.org/10.1075/sibil.47.03ch1

Jarvis, S. (2013b). Capturing the diversity in lexical diversity. *Language Learning*, *63*, 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing*, *34*, 537–553. https://doi.org/10.1177/0265532217710632

Jarvis, S., & Hashimoto, B. J. (2021). How operationalizations of word types affect measures of lexical diversity. *International Journal of Learner Corpus Research*, *7*, 163–194. https://doi.org/10.1075/ijlcr.20004.jar

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal*, *102*, 120–141. https://doi.org/10.1111/modl.12447

Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, *40*, 554–564. https://doi.org/10.1016/j.system.2012.10.012

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). lmerTest: Tests in linear mixed effects models. R package version 2. 0–20. https://cran.r-project.org/web/packages/lmerTest/lmerTest.pdf

Kyle, K. (2020). Measuring lexical richness. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 454–476). Routledge. https://doi.org/10.4324/9780429291586-29

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*, 757–786. https://doi.org/10.1002/tesq.194

Kyle, K., Crossley, S. A., & Berger, C. M. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*, 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, *18*, 154–170. https://doi.org/10.1080/15434303.2020.1844205

Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing*, *33*, 319–340. https://doi.org/10.1177/0265532215587391

Kyle, K., & Eguchi, M. (2021). Automatically assessing lexical sophistication using word, bigram, and dependency indices. In S. Granger (Ed.), *Perspectives on the L2 Phrasicon* (pp. 126–151). Multilingual Matters. https://doi.org/10.21832/9781788924863-007

Kyle, K., Eguchi, M., Choe, A. T., & LaFlair, G. (2022). Register variation in spoken and written language use across technology-mediated and non-technology-mediated learning environments. *Language Testing*, *39*, 618–648. https://doi.org/10.1177/02655322211057868

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*, 307–322. https://doi.org/10.1093/applin/16.3.307

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). Emmeans: Estimated marginal means, aka least-squares means (1.4.7) [R package]. https://cran.r-project.org/web/packages/emmeans/emmeans.pdf

Leonard, K. R., & Shea, C. E. (2017). L2 speaking development during study abroad: Fluency, accuracy, complexity, and underlying cognitive factors. *The Modern Language Journal*, *101*, 179–193. https://doi.org/10.1111/modl.12382

Lindqvist, C., Bardel, C., & Gudmundson, A. (2011). Lexical richness in the advanced learner's oral production of French and Italian L2. *IRAL—International Review of Applied Linguistics in Language Teaching*, *49*, Article 013. https://doi.org/10.1515/iral.2011.013

Lindqvist, C., Gudmundson, A., & Bardel, C. (2013). A new approach to measuring lexical sophistication in L2 oral production. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *Vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 109–126). EUROSLA-the European Second Language Association.

Lozano, C. (2015). Learner corpora as a research tool for the investigation of lexical competence in L2 Spanish. *Journal of Spanish Language Teaching*, *2*, 180–193. https://doi.org/10.1080/23247797.2015.1104035

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*, 190–208.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, *19*, 85–104. https://doi.org/10.1191/0265532202lt221oa

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*, 381–392. https://doi.org/10.3758/BRM.42.2.381

McManus, K., Mitchell, R., & Tracy-Ventura, N. (2021). A longitudinal study of advanced learners' linguistic development before, during, and after study abroad. *Applied Linguistics*, *42*, 136–163. https://doi.org/10.1093/applin/amaa003

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, *16*, 5–19.

Mendikoetxea, A. (2013). Corpus-based research in second language Spanish. In K. I. Geeslin (Ed.), *The handbook of Spanish second language acquisition* (pp. 9–29). Wiley. https://doi.org/10.1002/9781118584347.ch1

Meunier, F. (2015). Developmental patterns in learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 379–400). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.017

Mitchell, R., Tracy-Ventura, N., & McManus, K. (2017). *Anglophone students abroad: Identity, social relationships and language learning.* Routledge. https://doi.org/10.4324/9781315194851

Montrul, S. (2004). *The acquisition of Spanish: Morphosyntactic development in monolingual and bilingual L1 acquisition and adult L2 acquisition.* John Benjamins.

Mora, J. C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly*, *46*, 610–641. https://doi.org/10.1002/tesq.34

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

O'Loughlin, K. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing*, *12*, 217–237.

Ortega, L. (2000). *Understanding syntactic complexity: The measurement of change in the syntax of instructed L2 Spanish learners* [*Unpublished doctoral dissertation*]. University of Hawaii.

Ortega, L., & Byrnes, H. (2009). *The longitudinal study of advanced L2 capacities*. Taylor & Francis. https://doi.org/10.4324/9780203871652

Paivio, A. (1971). *Imagery and verbal processes*. Holt, Rinchart, and Winston.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*, 121–145. https://doi.org/10.1177/0267658317694221

Pérez-Vidal, C., Juan-Garau, M., Mora, J. C., & Valls-Ferrer, M. (2012). Oral and written development in formal instruction and study abroad: Differential effects of learning context. In C. Muñoz (Ed.), *Intensive exposure experiences in second language learning* (pp. 213–233). Multilingual Matters.

R Core Team. (2021). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732942

Saito, K. (2020). Multi- or single-word units? The role of collocation use in comprehensible and contextually appropriate second language speech. *Language Learning*, *70*, 548–588. https://doi.org/10.1111/lang.12387

Saito, K., & Akiyama, Y. (2017). Linguistic correlates of comprehensibility in second language Japanese speech. *Journal of Second Language Pronunciation*, *3*, 199–217. https://doi.org/10.1075/jslp.3.2.02sai

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical profiles of comprehensible second language speech: The role of appropriateness, fluency, variation, sophistication, abstractness, and sense relations. *Studies in Second Language Acquisition*, *38*, 677–701. https://doi.org/10.1017/S0272263115000297

Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research*, *27*, 343–360. https://doi.org/10.1177/0267658310395851

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. In P. Bański, H. Biber, E. Breiteneder, H. Kupietz, H. Lüngen, & A. Witt (Eds.), *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)* (pp. 28–34). Association for Computational Lingquistics. http://rolandschaefer.net/?p=749

Schäfer, R., & Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In N. C. (Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (pp. 486–493). http://rolandschaefer.net/?p=70

Schnur, E., & Rubio, F. (2021). Lexical complexity, writing proficiency, and task effects in Spanish dual language immersion. *Language Learning & Technology*, *25*, 53–72.

Segalowitz, N., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition*, *26*. https://doi.org/10.1017/S0272263104262027

Serrano, R., Tragant, E., & Llanes, À. (2012). A longitudinal analysis of the effects of one year abroad. *Canadian Modern Language Review*, *68*, 138–163. https://doi.org/10.3138/cmlr.68.2.138

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Stadthagen-Gonzalez, H., Imbault, C., Pérez Sánchez, M. A., & Brysbaert, M. (2017). Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods*, *49*, 111–123. https://doi.org/10.3758/s13428-015-0700-2

Tavakoli, P. (2018). L2 development in an intensive study abroad EAP context. *System*, *72*, 62–74. https://doi.org/10.1016/j.system.2017.10.009

Tavakoli, P., & Uchihara, T. (2020). To what extent are multiword sequences associated with oral fluency? *Language Learning*, *70*, 506–547. https://doi.org/10.1111/lang.12384

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. https://doi.org/10.1016/j.system.2017.09.022

Tracy-Ventura, N., Mitchell, R., & McManus, K. (2016). The LANGSNAP longitudinal learner corpus. In M. Alonso-Ramos (Ed.), *Spanish learner corpus research: Current trends and future perspectives* (pp. 117–142). John Benjamins.

Uchihara, T., & Saito, K. (2019). Exploring the relationship between productive vocabulary knowledge and second language oral ability. *The Language Learning Journal*, *47*, 64–75. https://doi.org/10.1080/09571736.2016.1191527

Vidal, K., & Jarvis, S. (2020). Effects of English-medium instruction on Spanish students' proficiency and lexical diversity in English. *Language Teaching Research*, *24*, 568–587. https://doi.org/10.1177/1362168818817945

Vincze, O., García-Salido, M., Orol, A., & Alonso-Ramos, M. (2016). A corpus study of Spanish as a foreign language learners' collocation production. In M. Alonso-Ramos (Ed.), *Spanish learner corpus research: Current trends and future perspectives* (pp. 299–331). John Benjamins.

Wolfe-Quintero, Inagaki S., & Kim, H. (1998). *Second language development in writing: Measures of fluency, accuracy & complexity*. University of Hawai'i Press.

Yule, C. U. (1944). *The statistical study of literary vocabulary*. Cambridge University Press.

Zaytseva, V., Miralpeix, I., & Pérez-Vidal, C. (2021). Because words matter: Investigating vocabulary development across contexts and modalities. *Language Teaching Research*, *25*, 162–184. https://doi.org/10.1177/1362168819852976

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, *47*, Article 100505. https://doi.org/10.1016/j.asw.2020.100505

Cite this article: Díez-Ortega, M., & Kyle, K. (2024). Measuring the development of lexical richness of L2 Spanish: A longitudinal learner corpus study. *Studies in Second Language Acquisition,* 46: 169–199. https://doi.org/10.1017/S0272263123000384