CHAPTER 2

Organic models for measuring Spanish learners' linguistic complexity

Joseph Collentine & Karina Collentine
Northern Arizona University

Second language acquisition (SLA) researchers measure linguistic complexity to assess pedagogical effectiveness and depict development (Norris & Ortega, 2009). Yet, from linguistic and cognitive perspectives, commonly used approaches oversimplify complexity. Furthermore, such approaches do not consider the morphological complexities of a highly inflected L2 like Spanish. Norris and Ortega (2009) encourage SLA researchers to develop empirically and theoretically valid measures of linguistic complexity through an organic (i.e., iterative and data-driven) investigative process; resulting models should be multidimensional and developmentally sensitive. This study delineates three multidimensional models of Spanish L2 linguistic complexity based on a principal component analysis of a corpus of learners participating in a task-based activity at the beginning, intermediate, and advanced levels of university instruction.

Keywords: linguistic complexity, second language learning, Spanish, corpus linguistics, developmental stage

1. Introduction

Second language acquisition (SLA) researchers measure linguistic complexity to assess the effectiveness of pedagogical strategies (e.g., task-based language teaching), to depict second language (L2) development, and to compare empirical results (Norris & Ortega, 2009; Pallotti, 2009). However, commonly used approaches oversimplify linguistic complexity, since they disregard theoretical and corpus linguistic conceptualizations. Such approaches are also incompatible with cognitive and psycholinguistic conceptualizations of linguistic complexity since they do not take into account learners' developmental stages (Norris & Ortega, 2009). Furthermore, commonly employed metrics have been mostly developed in

https://doi.org/10.1075/ihll.27.02col

the context of English as an L2, and they largely do not consider the morphological complexities of a highly inflected L2 like Spanish. Norris and Ortega (2009) encourage SLA researchers to delineate models of complexity through an organic (i.e., iterative and data-driven) investigative process for the purpose of developing research and assessment measurements that are more empirically and theoretically valid. They submit that resulting models should be multidimensional (i.e., employing a variety of morphosyntactic features) and sensitive to where learners are in their L2 development.

The present study represents an initial step in helping SLA researchers, especially those focused on Spanish as an L2, to meet Norris and Ortega's (2009) challenge. Specifically, we present a principal component analysis of a corpus representing the production of L2 learners of Spanish at three levels of instruction (i.e., beginning, intermediate, and advanced). The analysis provides three multidimensional models of linguistic complexity for Spanish as an L2, one for each of the three groups of learners.

2. Literature review

2.1 Shortcomings in current metrics of linguistic complexity

Norris and Ortega (2009) critique the validity and reliability of commonly used metrics of linguistic complexity in SLA research. They argue that research should work towards measuring linguistic complexity in ways that "engage with the construct reality of multidimensionality" (p. 574) and that are "developmentally sensitive" (p. 574). Regarding multidimensionality, most investigations measure linguistic complexity with a single metric, counting t-units, mean length of utterance (MLU), clauses per t-unit, clauses per c-unit, raw tallies of various structures (e.g., tensed verbs, comparatives), and lexical measures (e.g., type/token ratio, lexical density) (cf., Wolfe-Quintero, Inagaki, & Kim, 1998). Norris and Ortega (2009) argue that these metrics are imprecise because they either over- or under-estimate complexity. Theoretical and corpus linguists propose that sentences are complex when they contain various morphosyntactic phenomena (Biber & Gray, 2010; Park, 2017). Additionally, the tendency to assess complexity with syntactic measures ignores the morphological complexity of highly inflected languages like Spanish, e.g., *bonit* + *a* + *s* 'pretty – FEM-PL', *cant* + *a* + *ra* + *n* 'they sang IMP-SUBJ' (Park, 2017). Concerning developmental sensitivity, Jackson and Suethanapornkul (2013) suggest that parsimonious, one-size-fits-all measures of linguistic complexity are not sensitive enough to identify linguistic complexity at different levels of proficiency. Norris and Ortega (2009) argue that a learner's potential for producing sophistication is relative to his

or her developmental stage, such that complexity is "developmental in nature, growing, and changing all the time" (p. 556).

2.2 Conceptualizing complexity

Complexity can be defined from linguistic and philosophical perspectives. Linguistic frameworks, like Government and Binding, assume that syntactic units are complex when they contain composite constituents, such as embedded clauses and words with several morphological inflections. Philosophically speaking, language production often results in 'complex systems', a construct developed in the natural sciences and mathematics (Larsen-Freeman & Cameron, 2008). A complex system has nested hierarchies of simple (e.g., nouns, determiners) or composite (e.g., clauses) elements. A declarative sentence may contain several (arguably simple) elements organized into a hierarchical structure, such as: [$_{S}$ [$_{NP}$ Juan] [$_{VP}$ tiene [$_{NP}$ un libro] [$_{PP}$ en la mochila]]] 'Juan has a book in the backpack'. Sentences may also contain a dependent clause that is embedded into an independent clause or that modifies a simple element (Givón, 2009): [$_{S}$ [$_{VP}$ Vimos [$_{NP}$ la destrucción [$_{PP}$ del puente] [$_{CP}$ que mencionaste]]]] 'We saw the destruction of the bridge that you mentioned'. Spanish verbs are often hierarchically organized, where various inflectional morphemes are subordinate to the radical: [$_{stem}$ cant [$_{class}$ a [$_{tense\text{-}mood\text{-}aspect}$ rá [$_{person\text{-}number}$ s]]]] 'You will sing' (cf., Givón, 2009). Thus, Norris and Ortega's (2009) contention that linguistic complexity is multidimensional is consistent with the tenet that complexity occurs where various simple and composite elements are organized hierarchically.

2.3 Cognitive characteristics of linguistic complexity

Cognitive research can help to derive developmentally sensitive assumptions about linguistic complexity (Housen, Pierrard, & Van Daele, 2005; Housen & Kuiken, 2009). Native speakers require exponentially more time to process complex syntax and morphology than they need to process simpler phenomena (i.e., Ferreira, 1991). Learners are similarly limited, and their processing potential and efficiency changes over time (Bulté & Housen, 2014). For both beginning and advanced learners, a noun phrase such as [$_{NP}$ lunes] 'Monday' would require fewer resources than one containing various modifiers, such as in [$_{NP}$ [$_{DET}$ el] [$_{N}$ [$_{AP}$ próximo] lunes]] 'the following Monday'. However, beginners will find it more challenging to process subordinate clauses efficiently than advanced learners will. Thus, structural elaboration, such as modification and subordination, will be a reasonable measure of linguistic complexity at any level of development, even though the nature of elaboration will change over time.

The processing resources required to process inflectional morphology depends largely on a morpheme's semantic abstractness and typological markedness (Bulté & Housen, 2014; Housen, Kuiken, & Vedder, 2012). For example, masculine adjectives are non-abstract and unmarked since they apply to any noun (e.g., *Juan y María son listos* 'Juan and María are smart'). Essentially, learners use masculine-singular adjectives as automatized default forms. The production of a feminine inflection requires analysis, and so more processing resources. Reichenbach's (1947) well-known tense framework suggests that abstract/marked verb paradigms are conceptually complex because their meaning is relative (e.g., the future to the present, the conditional to the past) and because they are anchored to other paradigms (i.e., preterite to the imperfect, indicative to the subjunctive). First-year learners study a handful of simple (e.g., present indicative) and relatively marked (e.g., preterite/imperfect) verb paradigms. Yet, for a third-year learner, these paradigms presumably require fewer processing resources than, say, the present subjunctive and the conditional. Accordingly, conceptual abstractness and typological markedness also constitute linguistic complexity, and, like syntax, their nature changes over time.

3. An 'organic' approach to operationalizing L2 linguistic complexity

Norris and Ortega (2009) call for an organic approach to derive multidimensional, developmentally sensitive models of linguistic complexity that can detect elaborate L2 production while considering the typological features of the target language and the developmental potential of the learner population in question. They advocate an empiricist approach to achieving this goal, through an iterative process of observation, interpretation, and theory construction. Norris and Ortega (2009, p. 547) argue that an analytical approach to this challenge that can effectively align observations and theoretical predictions would be to employ factor analysis, provided the learner sample is large enough. Factor analysis can identify clusters of complex morphosyntactic features that occur in a corpus (Biber & Gray, 2010). Ultimately, such an approach – combined with theoretical insights into complexity – can yield a profile (i.e., a model) of linguistic complexity for a given sample.

A corpus-based approach and associated analytical tools (e.g., analyses of tagged corpora, factor analysis) allow researchers to understand learner interlanguage as its own system (Bley-Vroman, 1983), permitting us to characterize learner innovations with a diverse set of features (e.g., syntactic, morphological, lexical) and at different points in students' development (Marwa, 2014). A corpus-based approach is also advantageous because a learner complexity

profile can be readily compared to a native-speaker (NS) baseline, which facilitates the assessment of the extent to which a (proposed) learner model is native like or not.

4. Research questions

Following Norris and Ortega (2009), we take an organic approach to identifying models of linguistic complexity in the production of L2 learners of Spanish at three levels of instruction.

– What are the multidimensional characteristics of linguistic complexity of foreign-language learners of Spanish at the first, second, and third years of university instruction in a communicative task?

5. Method

5.1 Participants

University-level, foreign-language learners of Spanish at three instructional levels participated in the study (N = 214): beginning (second-semester) learners (n = 70), intermediate (fourth-semester) learners (n = 69), and advanced (third-year) learners (n = 75). The first- and second-year learners were enrolled in a Spanish program that promoted spoken and written proficiency. Students received explicit instruction (e.g., focusing on grammar, lexis, pragmatics) and participated in various communicative (i.e., reading, writing, conversational, cultural) tasks. The third-year learners were enrolled in upper-division grammar, conversation, and composition courses. The data were collected in the last quarter of a semester. Most beginning level students attain 8 credit hours in Spanish (128 seat hours). Most intermediate learners attain 16 credits (256 seat hours), and most advanced learners attain 21 credits (352 seat hours). Funding limitations did not permit an assessment of the participants' proficiency (e.g., ACTFL interview). Additionally, the researchers concluded that subjecting the participants to a proficiency test in addition to the experimental procedures would overburden the students and instructor. Still, we provide an objective perspective of the learners' performance and abilities by comparing the participants' data to a NS baseline (see *Corpus* below). All learners were L1 English speakers who were raised in monolingual households; none reported speaking any Spanish in the home environment. The study was conducted with the permission of IRB, and all participants provided informed consent to have their data reported anonymously.

5.2 Task

We derived the corpus from a two-segment task. First, learners were charged with finding the thief of a stolen relic in a virtual (i.e., 3D) world designed by the researchers, using Unity 3D (unity3d.com). The virtual world was an open-air marketplace with stands (e.g., a flower stand), stores (e.g., a wine shop), a café and restaurants seeded with customers and store owners. Students explored the marketplace at will, interviewing avatars (i.e., nonplayer characters) in the world to determine who was responsible for the theft. Learners gathered information about the avatars, such as their name, their reason for being at the marketplace, and whether they witnessed anything suspicious or a crime. Once learners approached an avatar, a screen appeared with buttons that learners could click in order to choose the questions they wanted to pose. Upon clicking a question button, learners read an avatar's response. All information was textual. When learners finished collecting clues from an avatar, they could move on to collect clues from another one or return to reread the clues from the original avatar. To help the participants recall avatars' names, anytime a question-answer screen was prompted, head shots and names of each of the avatars in the virtual world appeared at the bottom of the screen. This autonomous exploration phase of the task lasted 15 minutes.

Following, learners participated in a synchronized computer mediated communication (SCMC) segment. Random, pre-determined dyads shared clues they had collected from the avatars in Spanish in a Moodle instant-messaging chatroom. Dyads were to come to a consensus about who the culprit might have been. This SCMC phase lasted 25 minutes. Learners did not previously practice using these technologies since the researchers determined that all of the participants had experience using both virtual-world gaming and instant messaging. Dyads' interactions were archived to a database and converted to text format for part-of-speech tagging and syntactic parsing.[1]

5.3 Corpus

We wrote software routines with Python and the Natural Language Tool Kit (NLTK; <http://www.nltk.org/>) to obtain two analyses: morphological and syntactic. For the morphological analysis, we designed a part-of-speech tagger that annotated every word for word class (e.g., adjective, noun, verb, determiner, preposition) and morphological properties (e.g., plural, preterite).

1. See Collentine & Collentine (2015) for a complete description of how 3D world exploration and SCMC can provide communicative tasks for L2 learners.

(1) Sample part-of-speech tagging:
quien;conj
pienso;v_pres_1s
es;v_pres_3s
la;art_fs
persona;n_fs
que;relative
roba;v_pres_3s
el;art_ms
vaso;n_ms
?;punct
endofline;punct

For the syntactic analysis, we designed a regular-expressions parser to produce syntactic tree objects, which groups tagged words into phrasal constituents.

(2) Sample syntactic tree object:
(@SN los/art pistos/x (@SAdj más/adv interesante/adj))
(@SV están/v)
(@SPrep de/prep (@SN don/det manrique/n))

We also utilized quality control procedures to verify the accuracy of the tagging and parsing.

All told, dyads' data were separated into separate files. Two learners' (one at the intermediate level and another at the advanced level) whose sample were less than four exchanges were discarded. The learner corpus contained 22,681 words: 6,125 from the beginning level (mean words per learner = 87.5, sd = 38.8); 7,383 from the intermediate level (mean words per learner = 107.0, sd = 39.4); 9,173 from the advanced level (mean words per learner = 122.3, sd = 54.1).

We also designed, tagged, and parsed a NS baseline corpus. We randomly sampled the *Corpus del español*, extracting files that involved interviews (i.e., interactive dialogs) that were relatively short and where the number of turns per interlocutor was relatively equal. This yielded a sample of 130 individual interlocutors with 55,640 words (mean words per interlocutor = 428.0, sd = 27.4).

5.4 Dataset

Using Python scripts, we tabulated raw counts (per participant) of syntactic features from the syntactic tree objects and morphological features from the tagged data. The features ranged in their theoretical complexity (see 2.2 and 2.3 above). This tabulation allowed the researchers not only to examine linguistic complexity at the three instructional levels but also to assess the extent to which any group exhibited complexity behaviors.

We divided the syntactic variables into two levels of complexity, based on a structure's potential embedding.

Table 1. Syntactic structures and hypothesized complexity

High potential embedding	**Example**
Clause-Adverbial Cause	Voy [porque [necesito [pan]]] 'I'm going [because [I need bread]]]'
Clause-Adverbial Contingency	Voy [aunque [no debo]] 'I'm going [even though [I shouldn't]]'
Clause-Adverbial Purpose	Voy [para que [sepas [la verdad]]] 'I'm going [so that [you know [the truth]]]'
Clause-Adverbial If	Voy [si [tengo [tiempo]]] 'I'm going [if [I have [time]]]'
Clause-Adverbial Time	Voy [cuando [tengo [tiempo]]] 'I'll go [when [I have [time]]]'
Clause-Nominal	Entiendo [que [vas [ahora]]] 'I understand [that [you're going [now]]]'
Clause-Relative	Tengo una amiga [que [va [a [España]]]]] 'I have a friend [who [is going [to [Spain]]]]]'
VP + clitic	Yo [la [vi]]] 'I [her [saw]]]'
Moderate potential embedding	**Example**
Verb Phrase-Attributive	[es [difícil]] '[it's [difficult]]'
Noun Phrase	[un [hombre]], [casas] '[a [man]], [houses]'
Adjective Phrase	[muy [cansado]], [fácil] '[very [tired]], [easy]'
Verb Phrase	[duerme], [tiene [tiempo]] '[duerme], [tiene [tiempo]]'
Adverbial Phrase-pragmatic agreement	[muy [bien]], [de [acuerdo]] '[very [good]], [of [course]]'
Adverbial Phrase-discourse	[así que], [entonces] '[therefore], [then]'
Adverbial Phrase-intensifier	[muy [bien]] '[very [good]]'
Adverbial Phrase-location	[afuera], [muy [lejos]] '[outside], [very [far]]'
Adverbial Phrase-probability	[tal vez], [probablemente] '[perhaps], [probably]'

Table 1. (*Continued*)

High potential embedding	Example
Adverbial Phrase-time	[mañana], [cada [vez]] '[tomorrow], [each [time]]'
Prepositional Phrase	[en [la puerta]], [a [casa]] '[on [the door]], [to [the house]]'

In general, clausal categories and verb phrases with a clitic pronoun have high potential embedding since they may contain several simpler syntactic structures, as indicated in Table 1.

We divided the morphological variables into three levels of complexity, adopting aspects of markedness theory and Reichenbach's (1947) tense framework to operationalize the conceptual complexity of inflectional morphemes.

Table 2. Morphological structures and hypothesized complexity

Conceptual complexity	
High	Verb – Imperfect subjunctive Verb – Present subjunctive Verb – Imperfect Verb – Preterite
Moderate	Adjective – Marked (feminine and/or plural) Verb – Conditional Verb – Future
Low	Adjective – Unmarked (masculine, singular) Verb – Past participle Verb – Present participle Verb – Present indicative

As shown in Table 2, morphemes with high conceptual complexity have a function that is anchored relative to the present or past in addition to being anchored to another verbal inflection (i.e., preterite to imperfect, indicative to subjunctive). Moderate conceptual complexity is represented by inflections whose function is marked or anchored relative to the present or past. Low conceptual complexity is represented in inflections whose function is unmarked, anchored in the present, or not anchored relative to a time reference (i.e., non-finite).

Regarding syntax, for each participant/interlocutor, we calculated an embedding score, averaging the frequency of each structure's embedded constituents. For example, a nominal clause in the sentence *Entiendo [que [vas [ahora]]]* 'I understand [that [you're going [now]]]' would have 3 embedded constituents (including the complementizer), and the relative clause in *Tengo una amiga [que*

[va [a [España]]]]] 'I have a friend [who [is going [to [Spain]]]]]' would have 4. This approach makes the analysis sensitive to the extent to which learners at different levels – as well as the baseline data – complexified constituents. That is, this approach permits and assessment of the types of syntactic constituents where learners and NSs encoded hierarchical relationships between words. Regarding morphology, while inflectional morphology can be hierarchically organized (see 2.2 above), it is well known that high-frequency verbs are likely to be stored as phonetic segments (e.g., *fui* 'I went', *estuvo* 'he/she/it was', *haga* 'I/he/she/it may do/make') rather than analyzed inflectional composites, which is likely the case for less frequent verbs (e.g., *comience* 'I/he/she/it may begin', *encontré* 'I found'). Thus, in lieu of weighing each verb for its potential cognitive analysis, we elected to tabulate the frequency of each inflectional type (e.g., verbs in the present indicative, marked adjectives) per student/interlocutor. Thus, given that the syntactic analysis is largely structural (e.g., amount of embedding) and the morphological analysis largely conceptual (e.g., abstractness, markedness), we provide separate analyses of the learners' syntactic and morphological complexity (see Section 5.5 *Analysis* below).

5.5 Analysis

Principal component analysis (PCA) is a statistical technique similar to factor analysis that identifies linear combinations of variables (Tabachnick & Fidell, 2012). We employ PCA instead of factor analysis because the latter is most appropriate when the researcher possesses *a priori* a model of how features will cluster together. PCA is entirely exploratory, as is the present study. Each resulting 'principal component' (PC) essentially represents a cluster of variables that reliably co-occur, with each variable given a weight akin to a regression coefficient. For our purposes, each identified cluster contained structures that learners and native speakers *reliably* used in tandem, presumably for some communicative purpose. A cluster is considered reliable if the features (i) correlate in the data set and (ii) co-occur in the production of multiple participants. A PCA analysis might find a PC indicating that adverbial clauses (e.g., *para que*… 'so that…') and causal clauses (e.g., …*porque*… 'because…') reliably cluster together in a corpus.

Since linguistic phenomena naturally co-occur, we employed the oblique rotation promax to maximize the difference between PCs. Deriving clusters of co-occurring variables with PCA is an iterative process. For each of the PCA analyses presented, we started with all possible variables shown in Tables 1 and 2. To identify clusters of features that reliably co-occur in the data set and to avoid spurious results, we sought to maximize the communalities of all analyses, including only variables that share common variance at a minimum of 45%

(Tabachnick & Fidell, 2012). We interpret this cutoff to indicate that L2 language is highly variable, even when a corpus contains reliably co-occurring structures. Nevertheless, while this iterative process tended to reduce any final PCA analysis to a few variables, it had the effect of permitting us to establish load suppression levels (i.e., the point at which a variable's weight is too low for us to consider that variable relevant for a cluster) that are higher than normal for corpus analyses. The general rule of thumb for suppression levels is 0.30, and we were able to employ levels between 0.40 and 0.60, meaning that we can be confident that identified clusters are largely replicable.

Finally, while it is common to interpret PCA analyses by ascribing communicative functions to identified clusters (i.e., PCs), the present analysis focuses on each cluster's complexity. Thus, for example, if a cluster contained features such as adverbial clauses of time and temporal adverbs, one communicative interpretation of such a cluster might be that it represents a learner strategy for framing events chronologically. However, given space limitations, we limit ourselves to assessing the complexity of identified clusters.

The data were screened for univariate outliers. The minimum amount of data for PCA were satisfied for all of the morphological analyses, with a minimum of 10 observations per model variable. The syntactic models averaged 8 observations per variable. However, it is important to note that the lower ratio of sample-size-to-variables is most likely mitigated by the fact that communalities of the syntactic analyses were generally greater than 0.60 and the amount of variation accounted for was greater than 60% (cf., Field, 2013). We extracted the number of PCs per analysis based on an assessment of communalities and scree plots of components' Eigen values.

6. Results

In interpreting the results, it is important to keep in mind that the first PC extracted from a PCA analysis represents the most general cluster of features, accounting for the majority of the observed variance. Successive PCs are less pervasive within a dataset. Thus, the first PC is employed by learners or native speakers most reliably. Additionally, as noted above (see *Dataset*), we divided the variables into groups of relative complexity: the syntactic variables were grouped into those that represent high and moderate potential embedding (see Table 1); the morphological variables were grouped into high, moderate, and low conceptual complexity (see Table 2). In the following, we provide component loadings with a visual codification of the hypothesized complexity of each variable in order to facilitate the comparison of the groups of learners and the baseline according to their use of linguistic complexity.

6.1 Syntax

Figure 1 describes the mean embedding scores for the four groups, and the following 4 tables provide the syntactic PCA analyses for the NS baseline data and the three learner groups. In Tables 3–10, PC1 represents the first principal component, PC2 the second principal component, and so on.

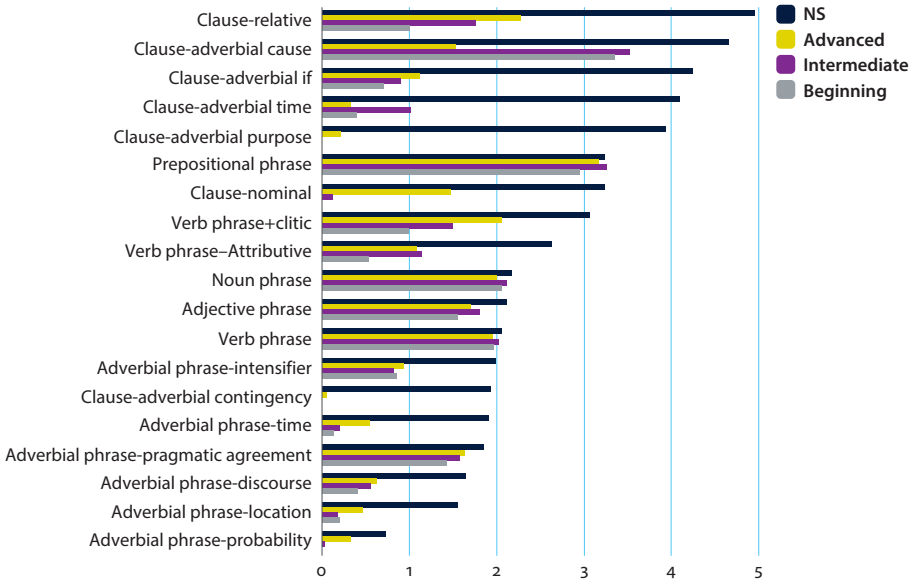

Figure 1. Mean learner/interlocutor syntactic embedding score by group

Table 3. Component matrix: Native speaker syntax

	PC1	PC2	PC3	PC4	PC5
% of variance	**15%**	**12%**	**11%**	**10%**	**8%**
Clause-Adverbial Time	**0.86**				**−0.45**
Adverbial Phrase-intensifier	0.76				
Verb Phrase-Attributive	**0.75**				
Clause-Adverbial If	**0.64**				
Clause-Adverbial Cause	**0.42**				
Prepositional Phrase		0.94			
Clause-Relative		**0.91**			
Noun Phrase		0.71			
Clause-Adverbial Contingency			**0.88**		

Table 3. (*Continued*)

	PC1	PC2	PC3	PC4	PC5
% of variance	**15%**	**12%**	**11%**	**10%**	**8%**
Adverbial Phrase-probability			0.85		
Clause-Nominal			**0.43**		
Verb Phrase + clitic				**0.88**	
Verb Phrase				0.88	
Clause-Adverbial Purpose					**0.90**
Adverbial Phrase-discourse					0.62

Legend – Embedding potential:
High Moderate

Notes:
Variance explained: 56%.
KMO Sampling adequacy: 0.81 (Test of sphericity $p < .001$).
Loadings < ± 0.40 are suppressed.

Table 4. Component matrix: Beginning level L2 syntax

	PC1	PC2	PC3	PC4
% of variance	**24%**	**17%**	**16%**	**15%**
Verb Phrase	0.99			
Noun Phrase	0.96			
Prepositional Phrase	0.46			
Clause-Nominal		**0.89**		
Adverbial Phrase-pragmatic agreement		0.85		
Adjective Phrase			0.64	
Adverbial Phrase-intensifier			0.90	
Clause-Adverbial Cause				**0.94**
Verb Phrase + clitic				**0.54**

Legend – Embedding potential:
High Moderate

Notes:
Variance explained: 72%.
KMO Sampling adequacy: 0.64 (Test of sphericity $p < .001$).
Loadings < ± 0.40 are suppressed.

Table 5. Component matrix: Intermediate level L2 syntax

	PC1	PC2	PC3	PC4
% of variance	**19%**	**17%**	**13%**	**13%**
Adjective Phrase	0.77			
Prepositional Phrase	0.69			
Verb Phrase		0.75		
Verb Phrase + clitic		**0.60**		
Noun Phrase		0.56		
Adverbial Phrase-pragmatic agreement			0.95	
Clause-Adverbial Cause				**0.95**

Legend – Embedding potential:
High Moderate

Notes:
Variance explained: 62%.
KMO Sampling adequacy: 0.64 (Test of sphericity $p < .001$).
Loadings < ± 0.50 are suppressed.

Table 6. Component matrix: Advanced level L2 syntax

	PC1	PC2	PC3	PC4
% of variance	**35%**	**18%**	**14%**	**14%**
Noun Phrase	0.95			
Prepositional Phrase	0.98			
Verb Phrase	0.96			
Clause-Relative		**0.88**		
Adverbial Phrase-pragmatic agreement		0.75		
Adjective Phrase			0.98	
Verb Phrase+clitic				**0.96**

Legend – Embedding potential:
High Moderate

Notes:
Variance explained: 81%.
KMO Sampling adequacy: 0.75 (Test of sphericity $p < .001$).
Loadings < ± 0.60 are suppressed.

Conversely, the learners mostly employed syntactic structures that had moderate embedding potential. Indeed, at all three levels, the first (most reliably produced) component contains not a single syntactic structure with high embedding potential. Throughout the beginning to advanced levels, the complexity of utterances is not found in clausal structures (although NSs generate numerous syntactically complex

utterance) but rather in basic constituents such as noun, verb, and prepositional phrases (i.e., 2 to 3 argument propositions). Syntactic structures with high embedding potential are underutilized relative to the NSs by the learners at all instructional levels. An examination of Figure 1 places the analysis in context. The learners complexified noun, adjective and verb phrases about as much as the NSs, and yet they did not do the same with clausal structures. This gives the appearance that, through advanced levels of instruction, learners hit a complexity wall, such that they can complexify basic syntactic structures as much as NSs do but not an array of clausal structures.

Still, the data give us insights into how complexification might develop. Learners at all 3 levels produced nominal clauses containing complexity (e.g., *Creo que Juan robó el tesoro* 'I think that Juan stole the treasure'). Learners at beginning and intermediate levels produced causal adverbials (e.g., *El criminal es Juan porque no tiene una buena excusa* 'The criminal is Juan because he doesn't have a good excuse') containing complexity, while those at the advanced level also used complexity in relativization (e.g., *Estaba en una casa que tiene tres baños* 'I was in a home that has three bathrooms'). It is interesting to note that these two complex structures were the ones that the NSs used most frequently (see Figure 1).

6.2 Morphology

Figure 2 describes the mean frequency of the morphological variables per group, and the following 4 tables provide the morphological PCA analyses for the NS baseline data and the three learner groups.

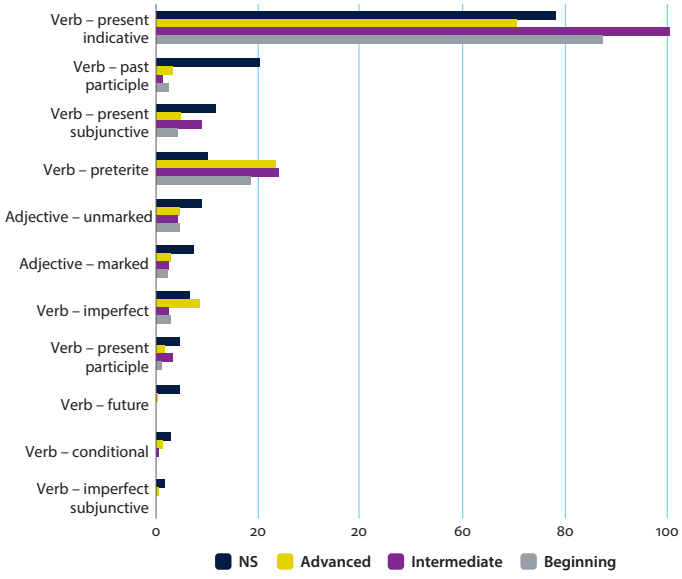

Figure 2. Mean learner/interlocutor morphological structure frequency by group

Table 7. Component matrix: Native speaker inflectional morphology

	PC1	PC2	PC3
% of variance	**21%**	**17%**	**16%**
Verb – Preterite	**0.77**		
Verb – Imperfect	**0.70**		
Verb – Imperfect subjunctive		**0.72**	
Verb – Conditional		0.59	
Verb – Present subjunctive		**0.68**	
Adjective – Marked (feminine and/or plural)			0.85
Verb – Future			0.73
Verb – Past participle			0.51
Adjective – Unmarked (masculine, singular)		−0.69	
Verb – Present indicative	−0.98		

Legend – Conceptual complexity:
High Moderate Low

Notes:
Variance explained: 54%.
KMO Sampling adequacy: 0.62 (Test of sphericity $p < .001$).
Loadings < ± 0.50 are suppressed.

Table 8. Component matrix: Beginning level L2 inflectional morphology

	PC1	PC2
% of variance	**25%**	**21%**
Adjective – Marked (feminine and/or plural)	0.79	
Verb – Imperfect	**0.66**	
Verb – Preterite	**0.65**	
Adjective – Unmarked (masculine, singular)		0.74
Verb – Present subjunctive		**0.68**
Verb – Present indicative		−0.43

Legend – Conceptual complexity:
High Moderate Low

Notes:
Variance explained: 46%.
KMO Sampling adequacy: 0.54 (Test of sphericity $p < .001$).
Loadings < ± 0.60 are suppressed.

Table 9. Component matrix: Intermediate level L2 inflectional morphology

	PC1	PC2
% of variance	**23%**	**18%**
Verb – Imperfect	**0.66**	
Verb – Preterite	**0.58**	**–0.57**
Verb – Present subjunctive		**0.85**
Verb – Present indicative	–0.78	

Legend – Conceptual complexity:
High Moderate Low

Notes:
Variance explained: 41%.
KMO Sampling adequacy: 0.48 (Test of sphericity $p < .001$).
Loadings < ± 0.60 are suppressed.

Table 10. Component matrix: Advanced level L2 inflectional morphology

	PC1	PC2
% of variance	**32%**	**23%**
Verb – Preterite	**0.72**	
Verb – Imperfect	**0.62**	
Verb – Present subjunctive		**0.83**
Adjective – Unmarked (masculine, singular)		–0.55
Verb – Present indicative	–0.75	

Legend – Conceptual complexity:
High Moderate Low

Notes:
Variance explained: 55%.
KMO Sampling adequacy: 0.57 (Test of sphericity $p < .001$).
Loadings < ± 0.60 are suppressed.

Morphologically speaking, an examination of Figure 2 indicates that the NSs mostly employed inflections with low conceptual complexity. Yet, an examination of the PCA data indicates that morphological complexity is interwoven within NS dialogic speech in interesting ways.

First, while Figure 2 and the first PC (see Table 7) indicate that NSs generate dialogic interactions mostly in the present tense, they reliably switch to past time frames. The preterite (0.77) and the imperfect (0.70) have high positive loadings

and the present indicative (−0.98) a high negative loading, indicating that present and past-tense morphology exist on a temporal communicative continuum (cf., Asención-Delaney & Collentine, 2011). (Note that what is important is not the signs of loadings but rather that these verbal inflections oppose each other along a single dimension.) Second, the second PC, which represents a communicative strategy within the more general first PC, is characterized by use of the subjunctive and the conditional. Thus, these data corroborate the findings of Biber, Davies, Jones, and Tracy-Ventura (2006), who found that Spanish assertive discourse is regularly backgrounded with irrealis presuppositions. In any event, while most inflectional morphology in conversational interactions is low in conceptual complexity, NSs reliably utilize inflectional morphology with high conceptual complexity in a supporting role (e.g., past-tense anecdotes/references, hypotheticals).

Before discussing the learner data, it is important to note that the present analysis does not consider errors. Thus, while the analysis indicates that learners used complex morphology to describe past events/states at all three levels in a task such as that employed in this study (i.e., where dyads needed to compare notes on who did what and when), it does not consider the extent to which they use, for example, the preterite and imperfect correctly.

For the most part, learners use inflectional morphology in ways that are similar to that of NSs: more often than not, they utilize inflectional morphology that is low in conceptual complexity. Yet, their interactions are interwoven with inflectional morphology that is high in conceptual complexity. Still, the data indicate that intermediate- and advanced-level learners diverge from beginning-level learners in important ways.

Learners at the beginning level reliably used adjectival inflections to elaborate nouns. The data in Table 8 also indicate that their most consistent strategy entails past-tense morphology. However, Figure 2 reveals that, like NSs, they most frequently use the present tense, and that the majority of the past-tense references are in the preterite. Still, the imperfect is beginning to be associated with the preterite, even though both past tenses are infrequently used. For these learners, the present indicative resides in the second PC probably because it is used in opposition to what appears to be uses of the present subjunctive. A review of the data indicated that none of the subjunctive uses constituted commands or were in subordinate clauses, suggesting that subjunctive use at this level was largely unintentional (see *Corpus samples* below).

For the intermediate- and advanced-level learners, the data indicate that, as in the NS case, the present and past tenses reside on a temporal communicative continuum, with the present tense being most prevalent. First, the preterite and imperfect loadings oppose (because of their signs) the present indicative. Second, the present indicative is the most commonly utilized inflection. Thus, intermediate- and advanced-level learners utilize complex inflectional morphology such as

the preterite and the imperfect in a supporting role to their present-tense assertions and descriptions, as NSs do. Additionally, an examination of the second PCs of the intermediate- and advanced-level PCAs indicates that the (present) subjunctive may begin at this level to assume a supporting irrealis role, as is probably its role for NSs.

6.3 Corpus samples

In the following, we provide samples from the learner corpus to illustrate conclusions presented above. The following provides examples of communicative features and linguistic complexity of two learners at the beginning level of instruction. The discussion references numbers (e.g., {1}, {2}, etc.) with which we have codified one or more segments within any dialog.

(3) Sample dialog from beginning-level learners.
S1: ¡Es Rita!
S2: Yo no [{5} hable] co Rita
S1: Es la mujer en [{1} el tienda de vino]
S1: La chica en el tienda cerda de Rita mira Rita [{3} entra] su tienda muy rapido
S2: Bueno, pero [{6} Sara [{4} dije] que [{1} un muchacho alto [{2} con un camisa blanco y khakis]] [{5} entre] la tienda]
S1: Un hombre [{3} va] en su tienda pero no [{3} mira] Rita esta ayí. No recordo hablar con Sarah.
S1: ¿Quien persona llevar ese ropa?
S2: No conozco pero [{6} yo pienso ser Tito]
S1: It's Rita
S2: I *didn't [{5} speak] *with Rita
S1: She's the woman in [{1} the wine shop]
S1: The girl in the store *near Rita looks at Rita [{3} entering] her store very quickly
S2: Well, but [{6} Sara [{4} *said] that [{1} a tall guy [{2} with a white shirt and khakis]] [{5} enters] the store]
S1: A man [{3} is going] in her store but [{3} doesn't see] Rita is there. I don't remember *speaking with Sarah.
S1: Which person *is wearing that clothing?
S2: I don't *know but [{6} I think *it's Tito]

These beginning-level learners produced some complicated noun phrases, as shown in the {1} segments. Indeed, the segment *un muchacho alto con un camisa blanco y khakis* 'a tall guy with a white shirt and khakis' contains a variety of embedded constituents. The {2} segment demonstrates that S2 can produce prepositional phrases with a noun containing a coordinated adjectival modifier. Yet, the present indicative not only provides conjectures within the dialog (i.e., the speech

situation); it also serves a narrative function (i.e., historical present), as in the {3} segments, which may be a compensatory strategy. Indeed, {4} constitutes an (erred) attempt to use the preterite. There are two possible subjunctive instances, as in the {5} segments, although *hable* 'he/she speaks' could merely lack a written accent and *entre* 'he/she enters' may be a preposition phonologically related to *entrar* 'to enter'. Finally, the {6} segments demonstrate the emergence of the production of nominal clauses, although *yo pienso ser Tito* 'I think to be Tito' is an instance of syntactic simplification.

In contrast, the following two samples provide examples of communicative features and linguistic complexity of two advanced-level dyads.

(4) **S3:** Estoy bien. [{9} Sabes dónde está el tesero]?
S3: tesoro*
S4: No sé, pero hay [{10} un libro que [{9} dice que los lucayanes siempre enterraban sus posesiones debajo de los hamacas]]. Tienes otros "clues"?
S3: Sí! [{9} Pienso que el tesoro está en [{10} la isla se llama San Salvador]]. Esta isla [{7} era] muy importante [{8} porque este lugar es donde Colón [{7} llegó] en 1492].
S3: Tambien, San Salvador [{7} fue] el capital para los Lucayanes.
S4: Muy bien! Para tu información, no tengo tiempo suficiente para hablar con todos las personas. Y por eso, solo tengo información limite. Pero [{9} sé que hay moneda de los lucayanes en la isla de San Salvador]
S3: I'm fine. [{9} Do you know where the *treasure is?]
S3: treasure*
S4: I don't know, but there is [{10} a book that [{9} says that the Lucayans always used to bury their possessions underneath the hammocks]]. Do you have other "clues"?
S3: Yes! [{9} I believe that the treasure is on [{10} the island *called San Salvador]]. This island [{7} was] very important [{8} because this place is where Columbus [{7} arrived] in 1492].
S3: Also, San Salvador [{7} was] the capital for the Lucayans.
S4: Very good! For your information, I don't have enough time to talk with all of the people. And that's why I only have limited information. But [{9} I know that there is *coin of the Lucayans on the island of San Salvador]

As shown in the {7} segments, advanced learners use the preterite and the imperfect to background conjectures framed in the present. There is an instance of a causal adverbial clause in {8}, where S3 uses this syntax to support an observation. The dialogue contains several well-formed nominal clauses, shown in the four {9} segments. Finally, we see the emergence of relativization in the {10} segments. One instance, *un libro que dice* 'a book that says', is well formed although the segment

está en la isla se llama San Salvador 'is on the island called San Salvador' lacks a relative pronoun.

(5) **S5:** Si, [{11} es posible que Sra Mendez [{12} escondió] el oro en el mar].
S5: Es loco, pero todo es posible.
S6: [{12} Tuviera] el equipaje.
S5: Si, estoy de acuerdo, Winston. [{11} Es muy improbable que la ladrona [{12} es] Sra Mendez]. Por eso, [{11} es posible que Rita y Tito [{12} sean] un equipo de ladrones].
S5: *Yes, [{11} it's possible that Sra Mendez [{12} hid] the gold in the sea].
S5: It's crazy, but everything is possible
S6: [{12} Would have] the luggage.
S5: *Yes, I agree, Winston. [{11} It is very improbable that the thief [{12} is] Sra Mendez]. That's why, [{11} it's possible that Rita and Tito [{12} are] a team of thieves].

This segment provides insights into the interaction between syntax and mood. There are various well-formed uses of nominal clauses in the {11} segments. Yet, we see correct and incorrect uses of the subjunctive in both main and subordinate clauses in the {12} segments.

7. Conclusions

This study sought to provide three multidimensional and developmentally sensitive models of linguistic complexity, focusing on L2 learners of Spanish at the beginning, intermediate, and advanced levels of instruction in a communicative task. A decidedly complex picture of linguistic complexity emerges from the analysis. First, the analysis indicates that intermediate- and advanced-level learners can complexify basic phrasal elements (e.g., noun, verb and adjectival phrases) to approximately the same extent that NSs do (see also Ortega, 2000). Beginning-level learners rarely complexify basic elements. Second, while learners seemingly can produce nominal clauses from early on, the ability to complexify an array of clause types may develop quite slowly. At the beginning and intermediate levels, learners generate complexity in the form of causal adverbial clauses (e.g., *porque*… 'because…' , *puesto que*… 'since…'). At the advanced level, learners also complexify relative clauses.

Most inflectional morphology that learners produce is decidedly low in conceptual complexity. Such was also the case for the NS baseline. This result is not entirely surprising since corpus research has found that NSs tend to exhibit little inflectional diversity in spontaneous, conversational interactions, although

planned and displaced discourse (e.g., letters, expository writing) contains inflectional diversity (Biber & Gray, 2010). This pattern may be a function of general processing limitations, such that framing most events/states within a present time frame is easy to manage. What is noteworthy about the results is how learners use morphological complexity alongside simple inflectional morphemes. Complexity occurs when learners interject present tense descriptions and assertions with past-tense anecdotes (i.e., short narratives with the preterite and imperfect) and hypotheticals (i.e., occasional uses of the subjunctive and the conditional).

The morphological analysis compels us to consider the influence of the task on production. The task required that learners conjecture about who did what. This design feature may explain the finding that nominal clauses (e.g., *creo que... 'I believe that…'.*) were found in the repertoire of all instructional levels. However, as Biber and Gray (2010) note, almost all spontaneous NS production, such as conversations and instant messaging (e.g., SCMC), contains little morphological elaboration. Under such production conditions, NS complexity manifests itself in syntax more than in morphology. Thus, the results here are likely to be ecologically valid for interpersonal interactions, especially in light of the fact that the NS baseline contained little morphological diversity. Nonetheless, these observations have implications for future research. A similar corpus study focusing on meaningful displaced writing (e.g., emails, narratives) may more effectively reveal the communicative and processing conditions that foster morphological complexity. More generally, multidimensional and developmentally sensitive complexity profiles should be studied under a variety of tasks types and modalities. In any event, the present study helps researchers to see that the assessment of linguistic complexity requires a consideration of both learners' syntactic and morphological abilities, and how morphosyntactic phenomena interrelate during production rather than the frequency with which learners use such structures.

References

Asención-Delaney, Y., & Collentine, J. (2011). A Multi-dimensional analysis of a written L2 Spanish corpus. *Applied Linguistics*, 32(3), 299–322. https://doi.org/10.1093/applin/amq053

Biber, D., Davies, M., Jones, J., & Tracy-Ventura, N. (2006). Spoken and written register variation in Spanish: A multi-dimensional analysis. *Corpora*, 1, 1–37. https://doi.org/10.3366/cor.2006.1.1.1

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, 9, 2–20. https://doi.org/10.1016/j.jeap.2010.01.001

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning*, 33(1), 1–17. https://doi.org/10.1111/j.1467-1770.1983.tb00983.x

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, 26(December), 42–65. https://doi.org/10.1016/j.jslw.2014.09.005

Collentine, J., & Collentine, K. (2015). Input and output grammar instruction in tutorial CALL with a complex grammatical structure. *CALICO Journal*, 32(2), 273–298. https://doi.org/10.1558/cj.v32i2.24548

Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language*, 30(2), 210–233. https://doi.org/10.1016/0749-596X(91)90004-4

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. London: Sage.

Givón, T. (2009). *Genesis of syntactic complexity: Diachrony, ontogeny, neuro-cognition, evolution*. Amsterdam: John Benjamins. https://doi.org/10.1075/z.146

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4), 461–473. https://doi.org/10.1093/applin/amp048

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 1–20). Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.32

Housen, A., Pierrard, M., & Van Daele, S. (2005). Rule complexity and the efficacy of explicit grammar instruction. In A. Housen & M. Pierrard (Eds.), *Investigations in instructed second language acquisition* (pp. 235–269). Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110197372

Jackson, D., & Suethanapornkul, S. (2013). The cognition hypothesis: A synthesis and meta-analysis of research on second language task complexity. *Language Learning*, 63(2), 330–367. https://doi.org/10.1111/lang.12008

Larsen-Freeman, D., & Cameron, L. (2008). *Complex systems and applied linguistics*. Oxford: Oxford University Press.

Marwa, R. (2014). Building a syntactically-annotated corpus of learner English (Unpublished doctoral dissertation). Indiana University.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4), 555–578. https://doi.org/10.1093/applin/amp044

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

Pallotti, G. (2009). CAF: Defining, refining, and differentiating constructs. *Applied Linguistics*, 30(4), 590–601. https://doi.org/10.1093/applin/amp045

Park, Y. (2017). Syntactic complexity as a predictor of second language writing proficiency and writing quality (Unpublished doctoral dissertation). Michigan State University, East Lansing, MI.

Reichenbach, H. (1947). *Elements of symbolic logic*. New York, NY: Macmillan & Co.

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics*. Boston, MA: Pearson Education.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity*. Honolulu, HI: University of Hawaii Press.