



LA SOFISTICACIÓN LÉXICA EN ESPAÑOL DE HABLANTES DE LENGUAS AFINES (CRIOLLO CABOVERDIANO Y PORTUGUÉS) Y NO AFINES (TELUGU E INGLÉS)¹

LEXICAL SOPHISTICATION IN SPEAKERS OF RELATED (CAPE VERDEAN CREOLE AND PORTUGUESE) AND NON-RELATED (TELUGU AND ENGLISH) LANGUAGES

María Moldes Suárez
Universidad Nebrija
merims222@gmail.com

Anna Doquin de Saint Preux
Universidad Complutense de Madrid
adoquind@ucm.es

RESUMEN

La complejidad lingüística como fenómeno multidimensional cuantificable apenas se ha abordado en la adquisición de español como lengua extranjera (ELE), obviando, además, la afinidad lingüística como una variable clave con efectos empíricamente demostrables. El presente estudio tiene como objetivo determinar la influencia de la afinidad lingüística en el nivel de sofisticación léxica comparando la producción escrita de dos grupos de hablantes de ELE de contextos plurilingües: un grupo de 40 caboverdianos con LM (criollo caboverdiano) y L2 (portugués) afines al español, y un grupo de 40 indios con LM (telugu) y L2 (inglés) no afines al español. Metodológicamente, se emplea el *Lexical Frequency Profile (LFP)* de Laufer & Nation (1995), el *Beyond 2000* de Laufer (1995) y el *Advanced Giraud* como operacionalizaciones a partir de dos listas de clasificación del vocabulario: la lista de lemas del *Corpus de aprendientes de español (CAES)* y la clasificación en niveles del *Marco común europeo de referencia para las lenguas (MCER)* a través de la herramienta *CFER Checker*. Los resultados muestran que la sofisticación léxica del grupo criollo es cuantitativamente superior a la del telugu independientemente de la lista y la operacionalización empleada. La prueba de Mann-Witney permite confirmar la hipótesis de influencia de la afinidad lingüística en el nivel de sofisticación mediante las operacionalizaciones realizadas a partir del *MCER*, siendo los resultados obtenidos a partir de la lista del *CAES* no significativos.

Palabras clave: afinidad lingüística, sofisticación léxica, *Lexical Frequency Profile*, *Beyond 2000*, *Advanced Giraud*, transferencia positiva, ELE.

ABSTRACT

Linguistic complexity as a quantifiable multidimensional phenomenon has hardly been addressed in the acquisition of Spanish as a foreign language, and linguistic affinity as a key variable with empirically demonstrable effects has been neglected. The present study aims to determine the influence of linguistic affinity on the level of lexical sophistication by comparing the written production of two groups of Spanish as a Foreign Language speakers from multilingual contexts: a group of 40 Cape Verdeans with LM (Cape Verdean Creole) and L2 (Portuguese) related to Spanish, and a group of 40 Indians with LM (Telugu) and L2 (English) not related to Spanish. Methodologically, the Lexical Frequency Profile (LFP) of Laufer & Nation (1995), the Beyond 2000 of Laufer (1995) and the Advanced Giraud are used as operationalizations with two vocabulary classification lists: the list of lemmas of the Corpus de aprendientes de español (CAES) and the Common European Framework of Reference for Languages (CEFR) level classification through the CFER Checker tool. The results show that the lexical sophistication of the Creole group is quantitatively superior to that of Telugu regardless of the list and operationalization employed. The Mann-Witney test confirms the hypothesis of the influence of linguistic affinity on the level of sophistication by means of the operationalizations based on the CEFR, while the results obtained from the CAES list are not significant.

Keywords: linguistic affinity, lexical sophistication, Lexical Frequency Profile, Beyond 2000, Advanced Giraud, positive transfer, Spanish as a foreign language.

¹ El presente estudio forma parte de la tesis doctoral en elaboración en la Facultad de Lenguas y Educación de la Universidad Nebrija bajo el título “La adquisición del español en hablantes de lenguas afines y no afines. Análisis de interlengua en hablantes de LM criollo/L2 portugués y hablantes de LM telugu/L2 inglés” y la dirección de Anna Doquin de Saint-Preux.



1. INTRODUCCIÓN

La adquisición de lenguas afines, es decir, aquellas que por estar emparentadas lingüísticamente tienen menor *distancia lingüística objetiva* entre sí (como el español, el francés, el italiano, el portugués, etc.), presenta un buen número de particularidades a las que no se ha prestado el suficiente interés investigador hasta el momento (Calvi, 2004). No obstante, lo relevante no es tanto la distancia lingüística objetiva sino la *subjetiva*, la psicolingüística (Kellermarn, 1977; Ellis, 1994), la *percepción de proximidad lingüística* que determina la tendencia a transferir las formas percibidas como menos marcadas o específicas (Sharwood Smith y Kellerman, 1986; Liceras, 1986). De ahí que la transferencia sea más común entre lenguas próximas o afines (Ringbom, 1987; Odlin, 1989; Fernández 1991; Swan, 1997; Torijano, 2002; Calvi, 2004; Campillos, 2012). En contra de la idea de Ellis (1985, p. 34)², los hablantes de lenguas afines al español, como el portugués, cometen más errores por transferencia negativa de su lengua materna (LM) (interferencia) y tienden en mayor medida a la fosilización de este tipo de errores (Santos Gargallo, 1993; Calvi, 1998, 2004; Torijano, 2002; Sánchez Iglesias, 2003).

Las interferencias se dan de manera más abundante en los estadios iniciales, dada la ausencia de conocimiento de la lengua objeto (LO) (Kellerman, 1983). En los estadios intermedios el aprendiente empieza a ser consciente de las divergencias y la evita, mientras que en los estadios avanzados reaparece como estrategia habitual (aunque cambia el tipo de estructuras transferidas) y se produce una tendencia al estancamiento del aprendizaje y a la fosilización, ya que, alcanzados los objetivos comunicativos, disminuye la motivación (Corder, 1971) y se economiza el “esfuerzo que supone incrementar la competencia” (Calvi, 2004, pp. 15-16).

² “*The greater the difference, the greater the difficulty and the more numerous errors will be*”.

Ahora bien, la transferencia no funciona solo de manera negativa, sino que es también una estrategia de comunicación y aprendizaje (Gass & Selinker, 1983; Ellis, 1985), un mecanismo cognitivo mediante el que el aprendiente de una lengua extranjera (LE) realiza hipótesis acerca de la LO, y que entre lenguas próximas permite el rápido avance de los aprendientes en los primeros niveles de adquisición. El término *influencia interlingüística* (*crosslinguistic influence*) (Sharwood Smith & Kellerman, 1986; Odlin 1989; Jarvis & Pavlenko, 2007) permite dar cuenta de la amplitud de fenómenos relacionados con la transferencia, tanto negativos (los que producen errores) como positivos (que determinan el éxito comunicativo y favorecen la adquisición de LE).

Mientras los errores y la interferencia han sido ampliamente estudiados en la adquisición de lenguas extranjeras, incluyendo el español, escasean los estudios en transferencia positiva e influencia interlingüística en adquisición de ELE.

Más allá de los errores, la concepción de la competencia lingüística como un constructo multidimensional exige la elaboración de índices cuantitativos que permitan medir objetivamente su desarrollo en lo que se ha llamado, por sus siglas inglesas, *CAF: Complexity, Accuracy, Fluency* (complejidad, precisión y fluidez). Este abordaje cuantitativo de la lingüística se inició con la *Ley de Zipf*³ (1932), según la que las palabras más frecuentes de una lengua conforman un reducido número frente a un mayor número de palabras poco frecuentes, estableciendo muchos años más tarde la *Ley de Heap* (1978), que señala que la mayor extensión de un texto implica un menor número de palabras diferentes (*types* o vocablos). Quedó así enunciando lo que se ha llamado, indistintamente, *diversidad, variedad o riqueza léxica* (la capacidad de emplear un vocabulario más variado y, por tanto, menos repetitivo) para la que otro pionero, Guiraud, había diseñado un índice ya en 1954 (López Morales, 2011, p. 17).

³ Se considera a Zipf como el padre de la lingüística cuantitativa (Ferrero García, 2011).

No obstante, la complejidad lingüística va más allá de la diversidad léxica e incluye tanto la *complejidad gramatical* (sintáctica y morfológica) como la *complejidad léxica*⁴ (Wolfe-Quintero, Inagaki & Kim 1998; Skehan, 2009; Bulté & Housen, 2012), dentro de la que se encuentran medidas que hacen referencia a la cantidad o amplitud de vocabulario como la diversidad y la densidad (que distingue las palabras que aportan información léxica de aquellas funcionales de contenido gramatical), y medidas mixtas (cuantitativas y cualitativas) que abordan la profundidad de ese vocabulario, como la sofisticación léxica que es la que nos ocupa en el presente estudio⁵. No obstante, estos subconstructos de la complejidad no funcionan de manera aislada, sino que están relacionados entre sí, por lo que muchas de las medidas son híbridas y miden varias dimensiones simultáneamente (Bulté & Housen, 2012).

Este campo de conocimiento ha sido muy prolífico en la adquisición de otras LE (principalmente inglés y, en menor medida, francés), así como prolíficos son los conceptos, las medidas y las críticas (Bulté & Housen, 2012; Norris & Ortega, 2009; Mavrou, 2016). Sin embargo, los estudios de complejidad sintáctica o léxica de español como lengua extranjera (ELE) son prácticamente inexistentes, de la misma manera que son escasos aquellos que abordan el estudio de la interlengua en ELE desde la perspectiva de la afinidad lingüística que aquí proponemos con participantes, además, pertenecientes a contextos plurilingües. Asimismo, no tenemos conocimiento de ninguna investigación científica en adquisición de ELE en hablantes con la

⁴ El término *riqueza léxica* se ha empleado tanto para referirse a la complejidad léxica en su conjunto (Laufer & Nation, 1995) como sólo a la diversidad o variedad léxica (Fernández, 1991, p. 129; Campillos, 2012, p. 151).

⁵ En la tesis doctoral inédita “La adquisición del español en hablantes de lenguas afines y no afines. Análisis de interlengua en hablantes de LM criollo/L2 portugués y hablantes de LM telugu/L2 inglés” en la que se enmarca el presente estudio se realiza un análisis profundo cuantitativo y cualitativo de la interlengua de los participantes desde una perspectiva holística y multidimensional: complejidad sintáctica, complejidad léxica (densidad, diversidad y sofisticación), análisis contrastivo, de errores y de interlengua.

combinación de lenguas maternas y lenguas segundas (L2) consideradas en este trabajo.

Pese a que la sofisticación léxica en lengua extranjera (LE) ha sido estudiada tanto para el inglés (Laufer & Nation, 1995; Laufer 1995; Naismith, Han, Juffs, Hill & Zheng, 2018) como para el francés (Tidball & Treffers-Daller, 2008), apenas existen estudios aplicados ELE, lo que justifica la pertinencia del presente estudio.

Bajo la hipótesis de que existen diferencias cuantitativas en el nivel de sofisticación léxica basadas en el grado de afinidad lingüística de la LM/L2, comparamos 40 participantes caboverdianos con LM criollo caboverdiano⁶ y L2 portugués⁷ (lenguas afines al español) con 40 participantes del sur de la India (de Anantapur, en el estado de Andhra Pradesh) con LM telugu y L2 inglés⁸ (no afines al español⁹) para determinar empíricamente, mediante diferentes índices, si la afinidad lingüística influye en la sofisticación léxica de su producción escrita.

2. PLANTEAMIENTOS TEÓRICOS

2.1 SOFISTICACIÓN LÉXICA

La *s sofisticación* o *rareza léxica* (*rareness*, Read, 2000, p. 200) está relacionada con la noción de palabras *s sofisticadas*, *avanzadas* (Laufer & Nation, 1995; Tidball & Treffers-Daller, 2008) o *difíciles* (Meara & Bell, 2001), es decir, menos frecuentes en el vocabulario habitual de una lengua. A diferencia de otras medidas exclusivamente cuantitativas, la sofisticación introduce

⁶ Lengua no oficial, pero hablada mayoritariamente por la población caboverdiana tanto en el archipiélago como en la diáspora, de base portuguesa (por lo que también se considera afín al español).

⁷ Cabo Verde fue colonia portuguesa hasta 1975.

⁸ La India fue colonia británica hasta 1947.

⁹ Nos referiremos a los dos grupos de participantes, respectivamente, como *grupo criollo* y *grupo telugu*.

también la variable cualitativa, que aporta una mayor profundidad al análisis (Daller, Van Hout & Treffers-Daller, 2003).

Considerando que las palabras más frecuentes son las más fáciles y las primeras que se aprenden (Cobb & Horst, 2004; Vermeer, 2004), el uso de palabras menos frecuentes es un indicador de una mayor competencia lingüística (Vermeer, 2000; Meara & Bell, 2001), aunque precisamente por esto puede no resultar adecuada para medir niveles de poca competencia (Meara & Bell, 2001) si no se acota al vocabulario de un nivel concreto (Arnaud, 1984; Linnarud, 1986; Laufer & Nation, 1995; Ferrero García de Jalón, 2011).

La validez de las medidas de sofisticación está menos cuestionada entre los investigadores que otras medidas de riqueza léxica (Meara & Bell, 2001, p. 15), si bien es conveniente tener en cuenta una serie de precisiones:

- Se distingue entre *token*, total de ocurrencias o palabras que componen un texto, y *type*, cada lema o palabra diferente (no repetida) dentro de un texto.
- Es necesario definir qué se considera palabra sofisticada o avanzada: para ello se han empleado criterios externos como listas de vocabulario básico o palabras frecuentes (Laufer & Nation, 1995), juicio de profesores (Daller *et al.*, 2003; Daller & Phelan, 2007) o la combinación de ambos (Tidball & Treffers-Daller, 2008).
- La lista empleada define la adecuación de la medida (Tidball & Treffers-Daller, 2008; Naismith *et al.*, 2018): no es adecuado emplear listas que incluyen vocabulario de un nivel superior en niveles más bajos (Arnaud, 1984; Linnarud, 1986), listas basadas en corpus escritos para datos orales (McCarthy, 1998; Tidball & Treffers-Daller, 2008; Lindqvist, Gusmundson & Bardel, 2013) o listas de frecuencia de vocabulario nativo para estudiar la LE, ya que se ha demostrado que se obtienen mejores resultados al emplear corpus de aprendientes (Naismith *et al.*, 2018), juicios de profesores sobre

vocabulario avanzado y la combinación de diferentes tipos de listas de frecuencia (Tidball & Treffers-Daller, 2008).

- La lematización del corpus ofrece mejores resultados (Mavrou, 2016; Naismith *et al.*, 2018). No obstante, los criterios de lematización varían entre los diferentes investigadores, desde los que consideran como un mismo lema las inflexiones básicas de una palabra (ej. género, número, desinencias verbales etc.) (Ávila, 1986; López Morales, 2011; Naismith *et al.*, 2018) hasta los que incluyen todas las derivaciones de la misma familia léxica bajo un único lema (Laufer & Nation, 1995).

La sofisticación se ha operacionalizado como la ratio de palabras sofisticadas entre el total de palabras empleadas, si bien existen otras operacionalizaciones como el *Lexical Frequency Profile (LFP)*¹⁰ de Laufer & Nation (1995), que consiste en clasificar el vocabulario según la banda de frecuencia a la que pertenece (de menor a mayor sofisticación): banda 1 (las 1000 palabras más frecuentes), banda 2 (las siguientes mil más frecuentes), *UWL (University Word List)*, 836 familias léxicas, no contenidas en las bandas anteriores, frecuentes en el discurso académico escrito de diferentes disciplinas), *NOL (Not in lists)*, palabras no contenidas en las listas anteriores).

Se ha postulado que, en el caso del inglés, las 2000 familias léxicas más frecuentes cubren entre el 80% y el 95% de un texto (Sanhueza, Ferreira & Sáez, 2018, p. 283; Schmitt, Schmitt & Clapham, 2001, pp. 55-56), dado que es en esta banda donde se encuentra un gran número de palabras funcionales y proveen de los recursos básicos para la comunicación cotidiana (Schmitt *et al.* 2001). La *UWL* cubre alrededor de un 4% en prensa y entre un 8,5% y

¹⁰ *VocabProfile* en su versión automatizada en <https://www.lex tutor.ca/vp/>. García, Washburn, Glenn y Graham (2013) diseñaron también una versión para el español, el *Spanish Vocabulary Online Profile (SVOP)*, anteriormente disponible en <http://tenney.biz/svop/>, (consultado el 01/11/2020)

10% dentro del discurso académico, y las siguientes 1000 más frecuentes (es decir, la banda de frecuencia de 2001 a 3000), que permitirían a un aprendiente empezar a comprender material auténtico en L2 (Nation, 2001), en torno a un 4,3%. En las siguientes bandas de frecuencia los porcentajes van disminuyendo.

Laufer (1995) propuso posteriormente una versión simplificada del *LFP*, considerando como sofisticadas todas las palabras que no están en la lista de las 2000 más frecuentes (*Beyond 2000*). Esta ha sido la versión más empleada, aunque posteriormente se ha planteado la posibilidad de establecer bandas de frecuencia menores (Tidball & Treffers-Daller, 2008), dado que el incremento de las palabras diferentes en un texto (*types*), contrariamente a la asunción de Malvern, Richards, Chipere & Durán (2004), no produce un incremento de palabras sofisticadas, sino una mayor variedad dentro de las 1000 más frecuentes (Laufer, 1998; Horst & Collins, 2006; Tidball & Treffers-Daller, 2008; Sanhueza, Ferreira & Sáez, 2018).

Anteriormente ya se habían llevado a cabo investigaciones con aproximaciones similares al *LFP*. Arnaud (1984) midió la sofisticación de acuerdo con una lista de 1522 palabras básicas y propuso una fórmula combinando la sofisticación, la variación (*types*) y el índice de errores léxicos, cuya inclusión ha sido cuestionada Laufer & Nation (1995), con el objetivo de discriminar entre la producción de nativos y no nativos. Linnarud (1986) midió la sofisticación considerando las palabras no esperables en un nivel de instrucción concreto, y empleó otras medidas como la variación (*TTR, type token ratio*), la densidad y la individualidad léxica, es decir, palabras dentro de un corpus que han sido usadas solo por uno de los participantes.

Otra medida que se ha mostrado efectiva para medir la sofisticación y discriminar entre grupos y niveles (Daller & Xue, 2007; Naismith *et al.*, 2018; Juffs, 2019), aunque no tanto en diseños longitudinales (Daller, Tulik & Weir, 2013), es la conocida como *Advanced Guiraud (AG)*, que mide la ratio de

types sofisticados respecto al total de ocurrencias (*tokens*) corrigiendo matemáticamente las diferentes longitudes de los textos¹¹ mediante la siguiente fórmula:

$$AG = \frac{\textit{typesofisticados}}{\sqrt{\textit{tokens}}}$$

2.2 ESTUDIOS PREVIOS

Laufer & Nation (1995) determinaron la validez y fiabilidad de su *Lexical Frequency Profile (LFP)* en 130 producciones escritas por 65 estudiantes de inglés LE (dos producciones por estudiante) en tres niveles (intermedio-bajo, intermedio, avanzado). Metodológicamente, se eliminan de los textos nombres propios y palabras con errores de significado, ya que, según los investigadores, no formarían parte del lexicón mental de los participantes (Laufer & Nation, 1995, p. 315) y se corrigen las que contienen errores de forma. La lematización se realiza siguiendo la familia léxica. El *LFP* mostró poder de discriminación en niveles estadísticamente significativo para la banda 1 (1000 familias más frecuentes), con porcentajes de vocabulario del 86,6% y 87,5% para cada producción del nivel más bajo, 79,7% y 79,4% del intermedio y 77% y 74% del avanzado. Sin embargo, las diferencias entre los tres grupos no fueron estadísticamente significativas en la segunda banda (las segundas 1000 más frecuentes), aunque se observó la tendencia a un descenso del porcentaje de vocabulario con el aumento del nivel (7,1% y 7% para el nivel

¹¹ Se ha probado ampliamente en la literatura científica la sensibilidad de las medidas de complejidad a la longitud del texto, especialmente en lo que a diversidad se refiere (Laufer & Nation, 1995; Daller, Van Hout & Treffers-Daller, 2003; Johansson, 2008; Mavrou, 2016), ya que, al producir más se tiende a repetir más palabras. A esta problemática se han aplicado soluciones que van desde reducir todas las muestras a un mismo número de *tokens*, con la consecuente pérdida de información (Laufer & Nation, 1995; Richard & Marlvern, 2000), hasta transformaciones matemáticas de diferentes tipos. Pese a que ninguno de los índices es plenamente satisfactorio (Mavrou, 2016), el índice de *Guiraud* y el *Advanced Giraud* reportan mejores resultados (Van Hout & Vermeer 2007; Vermeer 2000; Broeder, Extra and Von Hout, 1993; Daller, Van Hout & Treffers-Daller 2003; Daller & Xue, 2007; Berton, 2014).

más bajo, 6,7% y 6,8% del intermedio y 6,6% y 5,6% del avanzado). Para el vocabulario más sofisticado (*UWL* y *not on lists*) las diferencias fueron de nuevo significativas, como es lógico, con la tendencia contraria: niveles más bajos produjeron un menor porcentaje de familias pertenecientes a la *UWL* (3,2% y 4,1% del nivel más bajo, 8,1% y 7,8% del intermedio y 8,1% y 10,1% del avanzado) o que no se encuentran en ninguna de las listas (*not in lists*), es decir, más sofisticado (3,3% y 2,8% en el nivel más bajo, 5,6% y 6,6% en el intermedio y 7,5% y 8,7% en el más avanzado). Por otro lado, las correlaciones del *LFP* con el test de vocabulario (positiva para el vocabulario más sofisticado, negativa para el menos sofisticado y no significativa para la banda 2) demuestran su validez, mientras que su fiabilidad se prueba en la estabilidad en diferentes producciones independientemente del tema (excepto para el nivel avanzado).

Por su parte, Tidball y Treffers-Daller (2008) estudiaron la sofisticación en la producción oral de 41 estudiantes de francés LE de dos niveles (1 y 3), que mostraron diferencias significativas en el *C test*, y un grupo de control de 23 nativos. Compararon diferentes operacionalizaciones de la sofisticación (el índice de Guiraud (*G*), el *Advanced Guiraud* (*AG*) y la *D* de Malvern)¹² y diferentes listas de frecuencia: *FF₁* (compuesto por los 1378 lemas básicos del francés), una lista de 246 lemas basada en el juicio de dos expertos y 3 listas extraídas de *Corpaix* (lista de frecuencias del francés oral): una de 1378 lemas y otra de 246, para equiparar la cantidad y poder comparar con las listas anteriores, y otra de 2000 lemas de acuerdo al *Beyond 2000* de Laufer. Todas las operacionalizaciones, independientemente de la lista, mostraron poder de discriminación tanto entre nativos y no nativos como entre los dos niveles de francés LE, aunque estas no fueron significativas aplicando el *AG* basado en *FF₁* y en *Corpaix* (tanto teniendo en cuenta los 1378 lemas más

¹² Para una exposición más amplia sobre las diferentes medidas se puede consultar Mavrou (2016).

frecuentes como los 2000 más frecuentes). El cálculo del tamaño del efecto mediante eta cuadrado (η^2) determinó la superioridad de *D* y *AG* con la lista basada en juicio de expertos. El *Vocabprofil* (versión automatizada del *LFP* para el francés) también distinguió significativamente entre los tres grupos, tanto en la banda 1 (92,77% para el nivel 1, 90,87% para el 3 y 88,83% para los nativos) como en *NOL*¹³ (2,07% en el nivel 1, 4,37% en el 3 y 5,63% en los nativos), siendo este último (*NOL*) el mejor indicador según el tamaño del efecto (η^2). No obstante, como advierten los investigadores, se detecta un uso preferente de cognados de la LM frente a palabras más comunes o frecuentes en la L2 (Laufer y Paribakht, 1998; Horst & Collins, 2006), lo que incrementa el nivel de sofisticación cuando se trata, realmente, de un recurso estratégico (Tidball & Treffers-Daller, 2008: 310).

Otro estudio relevante es el de Naismith *et al.* (2018), que analiza la sofisticación en dos corpus de inglés LE (*Pitt English Language Institute Corpus (PELIC)* y *Corpus of Non-Native Written English (ETS)*¹⁴) con el objetivo de validar listas de frecuencia. Emplearon los 2000 lemas más frecuentes de la *New General Service List (NGSL)*¹⁵, basada en datos de nativos y no nativos, y los 2000 más frecuentes de una lista (*PSL_3*) creada a partir de las producciones escritas del *PELIC*. A diferencia de Laufer & Nation (1995), las palabras mal escritas se eliminan (no se corrigen) y se considera la inflexión (y no la familia léxica) en la lematización. La comparación de ambas listas reveló diferencias entre el vocabulario nativo (*NGSL*) y el no nativo (*PSL_3*), que se traducen en un menor *AG*, pero un mayor rango de diferencia entre niveles al emplear *PSL_3* (ocurriendo lo contrario al emplear *NGSL*). Con ambas listas el *AG* discrimina niveles tanto dentro de la misma LM (aunque no todas las

¹³ Los investigadores consideran como *NOL* las palabras no presentes entre las 3000 más frecuentes, ya que emplean una banda 3 que incluye las terceras 1000 palabras más frecuentes (en lugar de la *UWL* empleada por Laufer & Nation).

¹⁴ Disponible en <https://catalog.ldc.upenn.edu/LDC2014T06> [17/06/2022]

¹⁵ Disponible en <http://www.newgeneralservicelist.org/> [17/06/2022]

diferencias son estadísticamente significativas) como entre diferentes LM (a excepción de los grupos árabe y español del PELIC empleando *NGSL*). Los investigadores concluyen que el tipo de lista de frecuencia afecta los resultados de sofisticación, siendo la lista basada en corpus de aprendientes (*PSL_3*) la que presenta mayor incremento entre diferentes niveles de competencia y que, por tanto, discrimina mejor.

3. MÉTODO

Mediante un estudio descriptivo focalizado sincrónico y transversal, empleando una metodología cuantitativa, se evaluó la hipótesis (H_1) de que existe diferencia en la sofisticación lingüística según la LM/L2, siendo la hipótesis nula (H_0) que no existe diferencia.

Para la recolección de datos¹⁶ se empleó un cuestionario para recabar información sobre el perfil de los participantes y una producción escrita sobre aspectos del país y la cultura de los participantes, de la que se extrae el corpus de datos escritos.

A nivel metodológico, se emplean diferentes operacionalizaciones de la sofisticación léxica basadas en el *Lexical Frequency Profile (LFP)* propuesto por Laufer & Nation y en el *Advanced Giraud (AG)*, partiendo de dos clasificaciones diferentes del vocabulario:

- 1) Para la primera clasificación se emplea la lista total de lemas del *Corpus de aprendices de español (CAES)*, desarrollado por el Instituto Cervantes y la Universidad de Santiago de Compostela, por ajustarse a las características de nuestro corpus: está basado en la producción escrita de aprendientes de español de los niveles A1 a C1 con diferentes LM tanto afines (portugués y francés) como no afines (inglés, chino, ruso y árabe). A partir de esta lista se distribuye el léxico de cada texto de nuestro corpus en bandas de frecuencia: 1-1000 (léxico que

¹⁶ La recolección de datos se efectuó en 2015 para el grupo telugu y en 2019 para el criollo.

figura entre los 1000 lemas más frecuentes de la lista *CAES*); 1001-2000 (léxico que figura entre los segundos 1000 lemas más frecuentes de la lista *CAES*); No en 2000 (léxico que no figura entre los 2000 lemas más frecuentes de la lista *CAES*). Considerando esta clasificación, se establece el *LFP* según la media del porcentaje de *tokens* contenidos en cada banda de frecuencia (Laufer & Nation, 1995), denominado *LFP CAES* en el presente estudio.

Se incluye, asimismo, la especificación del léxico que no figura entre los 1000 lemas más frecuentes (No en 1000), a partir del que se calcula el *AG* 1000, teniendo en cuenta el bajo nivel de significación que ha reportado la banda 1001-2000 en precedentes investigaciones (Laufer & Nation, 1995; Tidball & Treffers-Daller, 2008), contrastándolo con el *AG* 2000 que, siguiendo el *Beyond 2000* de Laufer (1995), considera como sofisticados los *types* no presentes entre los 2000 más frecuentes.

$$AG1000 = \frac{typesNoen1000}{\sqrt{tokens}} \quad AG2000 = \frac{typesNoen2000}{\sqrt{tokens}}$$

- 2) Dado que las listas de frecuencia no aportan información sobre el nivel al que pertenece el vocabulario en términos didácticos de LE, se establece una segunda clasificación empleando el *CEFR Checker*, herramienta *online* gratuita de la plataforma *Duolingo*¹⁷ que clasifica los *types* de un texto según los niveles determinados en el *MCER* (A1, A2, B1, B2, C y desconocido (?)). De la media del porcentaje de *types* contenidos de cada nivel se obtiene una segunda versión del *LFP*, denominada *LFP MCER* en el presente estudio. El *AG* se calcula tanto considerando como sofisticados los *types* iguales o superiores a B1 ($AG \geq B1$), nivel que poseen la mayoría de los participantes del

¹⁷ La aplicación dejó de estar disponible tras la realización del presente estudio.

estudio, como los iguales o superiores a B2 ($AG \geq B2$), nivel superior al de la mayoría de los participantes.

$$AG \geq B1 = \frac{typesB1 + typesB2 + typesC}{\sqrt{tokens}}$$

$$AG \geq B2 = \frac{typesB2 + typesC}{\sqrt{tokens}}$$

Se definen, además, los siguientes criterios:

- La lematización de nuestro corpus se realiza según la inflexión (Ávila, 1986; López Morales, 2011; Naismith *et al.*, 2018), del mismo modo que en un diccionario.
- Se eliminan los elementos que no aportan información sobre la sofisticación léxica y que no están presentes en el CAES: nombres propios, nombres específicos de la cultura de los participantes (ej. *catchupa*, *bindi*, etc.), números escritos con signos matemáticos (ej. *1*, *1.º*, etc.), acuñaciones no identificables (ej. **fissinas*, **percuo*) y formas producidas únicamente como extranjerismos en el corpus y no presentes en CAES¹⁸.
- Se corrigen las deformaciones producidas por motivos ortográficos (Laufer & Nation, 1995).

La variable independiente son las lenguas previamente adquiridas (criollo/portugués o telugu/inglés) y las variables dependientes son las mediciones de sofisticación léxica operacionalizadas como *tokens* y *types* 1-1000, 1001-2000, No en 2000, No en 1000, A1, A2, B1, B2, C, A1+A2, $\geq B1$ y $AG \geq 1000$, $AG \geq 2000$, $AG \geq B1$ y $AG \geq B2$. Los resultados se procesan y contabilizan en Excel y se analizan estadísticamente con el programa libre JASP.

¹⁸ Nótese que hay muchas formas en el corpus que, dada la variabilidad de la interlengua, se producen como extranjerismos, deformaciones y formas correctas. Estas formas fueron corregidas mediante el lema correspondiente y están incluidas en la medida de sofisticación. Asimismo, hay ciertos extranjerismos presentes en CAES (ej. *basketball*, *ficar*, *enquanto*, etc.) que no han sido eliminados de nuestro corpus.

3.1 PERFIL DE LOS PARTICIPANTES

Todos los participantes de Cabo Verde son estudiantes universitarios (31 de género femenino y 9 de masculino), con una media de edad 21,92 años. Los participantes indios (de 37,35 años de media, 30 de género femenino y 10 masculino), sin embargo, no son estudiantes, sino usuarios del español como trabajadores de una ONGD española. Tienen estudios universitarios 28 de ellos, mientras que el resto tiene estudios de bachillerato (19) y secundaria (2).

Además de sus respectivas L2, todos los participantes del grupo criollo conocen el inglés y el francés (dada su obligatoriedad en el sistema educativo caboverdiano), mientras que en el grupo telugu 19 conocen hindi, 8 otras lenguas indias (urdu, kannada y tamil) y 5 tienen nociones de catalán debido a estancias en Cataluña.

Todos los participantes han cursado el nivel A2+, aunque su nivel de competencia puede ser heterogéneo, especialmente en el grupo telugu dado su estatus de usuarios. La instrucción recibida en español y la exposición a este idioma es muy variable entre ambos grupos: los participantes del grupo criollo cursaban el nivel 2 de ELE en el momento de la recolección de datos, acumulando unas 120 horas de instrucción en el idioma. La mayoría de los participantes del grupo telugu había recibido entre 500 y 1200 horas de instrucción en español durante unos 3 años en el pasado y su exposición al idioma es continua, puesto que lo emplean en su trabajo diario.

La producción escrita de los participantes fue evaluada por evaluadores externos (8 evaluadores de media por texto) mediante una escala holística adaptada de los DELE (Instituto Cervantes, 2019a, pp. 13-14; 2019b, pp. 14-15; 2014a, pp. 16; 2014b, p. 15; 2014c, p. 16) y una analítica adaptada del SIELE (2017, pp. 88-94). Estas evaluaciones permitieron determinar un nivel B1 mayoritario (30 textos del grupo criollo y 26 del telugu). Los 10 textos restantes del grupo criollo se distribuyen equitativamente entre los niveles A2 y B2 (5 en cada nivel) y los del grupo telugu en B2 (6 textos), A2 (6 textos) y A1 (2 textos).

4. RESULTADOS

La Tabla 1 recoge los resultados siguiendo la lista de frecuencias del CAES. Partiendo de la propuesta de Laufer & Nation (1995), se calcula la media de *tokens* y *types* en las diferentes bandas de frecuencia tanto en términos absolutos (*M*) como relativos (% *M*), así como el *AG*, considerando como sofisticados los *types* no presentes entre los 1000 más frecuentes (*AG* 1000) y los no presentes entre los 2000 más frecuentes (*AG* 2000), de acuerdo con el *Beyond 2000* de Laufer (1995).

		CRIOLLO				TELUGU				
		<i>M</i>	<i>DT</i>	<i>M</i> %	<i>DT</i> %	<i>M</i>	<i>DT</i>	<i>M</i> %	<i>DT</i> %	
	Tokens	212,90	61,96			235,60	86,55			
	Types	88,60	19,76			103,18	27,19			
LFP CAES	1-1000	<i>tokens</i>	181,20	53,01	85,18%	3,64%	203,5	76,64	86,07%	3,05%
		<i>types</i>	64,55	12,84	73,38%	6,03%	79,95	20,12	77,75%	5,15%
	1001-2000	<i>tokens</i>	10,78	4,94	5,08%	1,96%	13,98	6,45	6,01%	2,20%
		<i>types</i>	8,25	3,23	9,23%	2,61%	9,83	4,46	9,36%	3,16%
	No en 2000	<i>tokens</i>	20,93	9,21	9,73%	2,65%	18,13	7,86	7,93%	2,53%
		<i>types</i>	15,8	6,87	17,40%	4,72%	13,40	5,90	12,89%	3,70%
	No en 1000	<i>tokens</i>	31,70	12,92	14,89%	3,64%	32,10	12,07	13,93%	3,05%
		<i>types</i>	24,05	9,36	26,62%	6,03%	23,23	9,37	22,25%	5,15%
AG	AG 1000	1,64	0,47			1,51	0,44			
	AG 2000	1,07	0,36			0,88	0,29			

TABLA 1. Sofisticación léxica: LFP CAES (*tokens* y *types*) y AG

Como se observa en la Tabla 1 y la Figura 1, el grupo telugu supera al criollo tanto en el total de *tokens* (235,60 de media frente a 212,90) y *types* (103,18 de media frente a 88,60), como en el porcentaje medio de *tokens* pertenecientes a las bandas de mayor frecuencia: 1-1000 (85,18% de media de lemas del grupo criollo y 86,07% del telugu) y 1001-2000 (5,08% del grupo criollo y 6,01% del telugu), lo que significa que el 90,27% del vocabulario del grupo

criollo y el 92,07% del telugu se sitúan entre los 2000 lemas más frecuentes en el corpus *CAES*. Como vocabulario más frecuente de aprendientes de LE, es menos sofisticado o avanzado, más común, por lo que la sofisticación depende del porcentaje que no pertenece a esos 2000 lemas más frecuentes, considerando el *Beyond 2000* de Laufer (1995): 9,73% en el grupo criollo y 7,93% en el telugu. La sofisticación, por tanto, es considerablemente mayor en el grupo criollo.

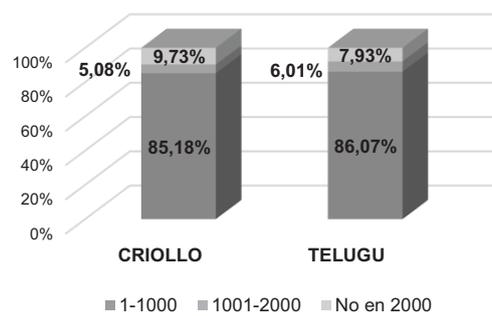


FIGURA 1. Sofisticación léxica: LFP *CAES* (tokens)

En la Figura 2 se aprecia el importante aumento del porcentaje de sofisticación (No en 2000), así como del rango de diferencia entre ambos grupos, al considerar el número de *types* en lugar del de *tokens*, 17,40% para el grupo criollo y 12,89% para el telugu, lo que se debe a la producción de numerosos *tokens* funcionales pertenecientes a la banda de mayor frecuencia (1-1000)¹⁹. De ahí tanto el incremento de la sofisticación como el decrecimiento de la banda de mayor frecuencia (1-1000), que contiene el 73,38% de los *types* empleados por el grupo criollo y el 77,75% del telugu. Los *types* de la banda 1001-2000, en comparación con los *tokens*, también crecen, alcanzando un porcentaje casi idéntico en ambos grupos: 9,23% para el grupo criollo y 9,36% para el telugu.

¹⁹ Recordemos que al establecer el LFP según los *tokens* se cuenta cada ocurrencia del mismo lema dentro de la banda 1-1000, mientras que al tener en cuenta los *types* se cuenta solo una vez.

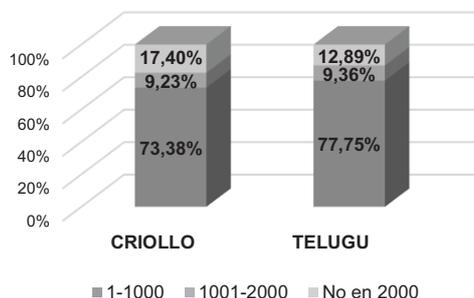


FIGURA 2. Sofisticación léxica: LFP CAES (*types*)

Estos datos implican que el grupo telugu produce tanto más palabras (*tokens*) como más palabras diferentes (*types*), pero que pertenecen a un vocabulario más común y frecuente y, por tanto, menos sofisticado. El grupo criollo emplea menos palabras y palabras menos variadas, pero más avanzadas o, lo que es lo mismo, más sofisticadas.

Tanto los resultados del *AG* (Figura 3) como los obtenidos mediante el *CEFR Checker* confirman el del *LFP CAES*: la mayor sofisticación léxica del grupo criollo. La media del *AG* 1000 y la del *AG* 2000 son mayores en el grupo criollo: 1,64 y 1,07 respectivamente, frente a 1,51 y 0,88 respectivamente en el grupo telugu.

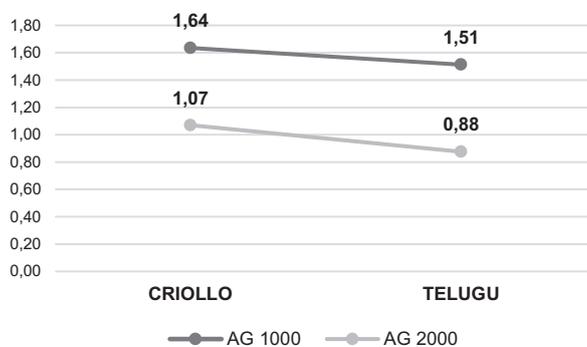


FIGURA 3. Sofisticación: AG 1000 y AG 2000

En cuanto al *LFP MCER* (Tabla 2 y Figura 4)²⁰, el grupo criollo emplea un porcentaje medio menor de *types* del nivel más bajo, A1 (61,63% frente al 72,35%

²⁰ La media de *types* del grupo criollo es ligeramente mayor, dado que el recuento automático del *CEFR Checker* difiere mínimamente del realizado mediante Excel en el *LFP CAES* (88,93 frente a 88,60). Consideramos, no obstante, que la diferencia no afecta a los resultados. No se producen diferencias en el grupo telugu.

del telugu), y porcentajes mayores del resto de los niveles, ampliando el rango respecto al telugu a medida que se incrementa el nivel: A2 (15,9% frente al 14,76% del telugu), B1 (8,15% frente al 6,46% del telugu), B2 (7,93 % frente al 4,43% del telugu) y C (4,49% frente al 1,79% del telugu). El porcentaje de términos no reconocidos (?) también es mayor en el grupo criollo (1,8%)²¹ que en el telugu (0,2%), así como todas las desviaciones típicas, lo que indica mayor variabilidad interna en este grupo.

			CRIOLLO				TELUGU			
			<i>M</i>	<i>DT</i>	<i>M %</i>	<i>DT %</i>	<i>M</i>	<i>DT</i>	<i>M %</i>	<i>DT %</i>
	Tokens		212,90	61,96			235,60	86,55		
	Types		88,93	19,27			103,18	27,19		
LFP	A1	<i>types</i>	54,48	11,08	61,63%	5,71%	73,98	16,99	72,35%	5,12%
MCER	A2	<i>types</i>	14,28	5,05	15,99%	4,40%	15,18	4,95	14,76%	2,85%
	A1+A2	<i>types</i>	68,75	14,37	77,62%	5,94%	89,15	20,61	87,12%	4,29%
	B1	<i>types</i>	7,40	3,90	8,15%	3,39%	6,83	3,74	6,46%	2,65%
	B2	<i>types</i>	7,23	3,64	7,93%	3,09%	4,93	3,44	4,43%	2,37%
	C	<i>types</i>	4,03	2,29	4,49%	2,14%	2,08	2,34	1,79%	1,61%
	≥ B1	<i>types</i>	18,65	8,16	20,57%	6,03%	13,83	7,85	12,67%	4,33%
	?	<i>types</i>	1,53	0,72	1,80%	0,97%	0,20	0,41	0,21%	0,44%
AG	AG ≥ B1		1,27	0,44			0,87	0,37		
	AG ≥ B2		0,77	0,29			0,43	0,25		

TABLA 2. Sofisticación léxica: media del porcentaje de *types* en cada nivel del MCER (LFP MCER) y AG

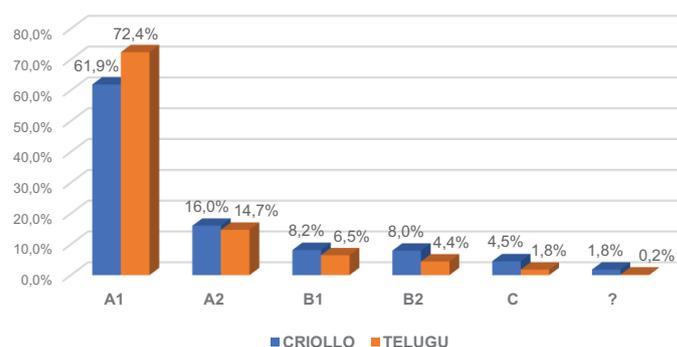


FIGURA 4. Sofisticación léxica: media del porcentaje de *types* en cada nivel del MCER

²¹ Debido, en gran medida, al extenso empleo del adjetivo *caboverdiano/a* (121 ocurrencias), *type* no reconocido en el *CEFR Checker*

El 87,12% de los *types* del grupo telugu y el 77,62% del criollo pertenecen a los niveles iniciales (A1+A2), mientras que el 12,67% del telugu y el 20,57 % del criollo alcanzan o superan el nivel intermedio ($\geq B1$), lo que implica más capacidad por parte del grupo criollo para emplear vocabulario que está por encima de su nivel de competencia, mientras que el grupo telugu tiende a emplear un vocabulario más básico, menos sofisticado, propio de niveles más bajos.

Asimismo, las medias de $AG \geq B1$ y $AG \geq B2$ (Figura 5) del grupo criollo (1,27 y 0,77 respectivamente) superan ampliamente a las del telugu (0,87 y 0,43 respectivamente). Las altas desviaciones típicas en ambos grupos (0,44 y 0,29 en el grupo criollo y 0,37 y 0,25 en el telugu) demuestran, nuevamente, la gran variabilidad interna dentro de los grupos.

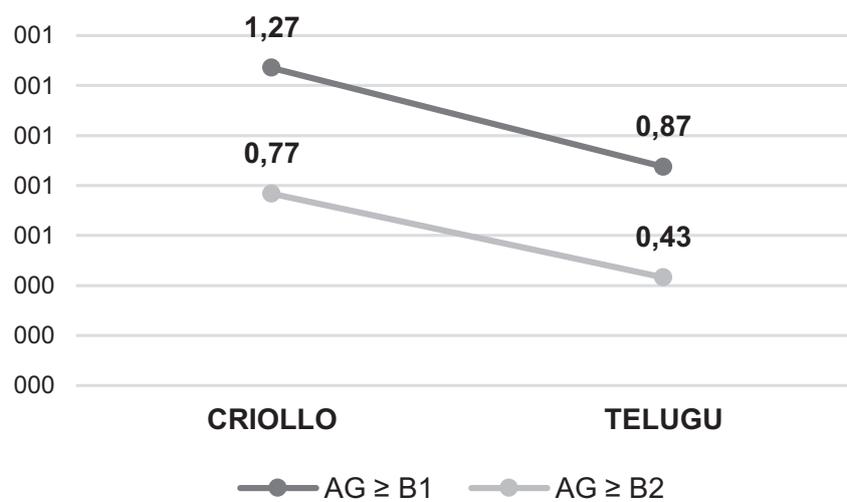


FIGURA 5. Sofisticación: $AG \geq B1$ y $AG \geq B2$

Para determinar estadísticamente la relación entre las diferentes variables, confirmamos en primer lugar el ajuste a la normalidad en la distribución de los datos en cada grupo mediante la prueba Shapiro-Wilk (Tabla 3).

		W	p
Tokens	CRIOLLO	0.953	0.098
	TELUGU	0.941	0.037
Types	CRIOLLO	0.927	0.013
	TELUGU	0.969	0.333
1-1000 tokens	CRIOLLO	0.954	0.102
	TELUGU	0.941	0.038
1-1000 types	CRIOLLO	0.932	0.019
	TELUGU	0.979	0.669
1001-2000 tokens	CRIOLLO	0.935	0.023
	TELUGU	0.978	0.624
1001-2000 types	CRIOLLO	0.901	0.002
	TELUGU	0.973	0.432
No 2000 tokens	CRIOLLO	0.896	0.001
	TELUGU	0.901	0.002
No 2000 types	CRIOLLO	0.886	< .001
	TELUGU	0.867	< .001
A1+A2	CRIOLLO	0.945	0.050
	TELUGU	0.976	0.529
≥ B1	CRIOLLO	0.893	0.001
	TELUGU	0.899	0.002
AG 1000	CRIOLLO	0.925	0.011
	TELUGU	0.953	0.099
AG 2000	CRIOLLO	0.911	0.004
	TELUGU	0.961	0.182
AG ≥ B1	CRIOLLO	0.952	0.087
	TELUGU	0.936	0.025
AG ≥ B2	CRIOLLO	0.965	0.246
	TELUGU	0.924	0.010

TABLA 3. *Shapiro-Wilk: variables de sofisticación léxica*

Dado que no podemos asumir una distribución normal en algunas de las variables de ambos grupos ($p < ,05$) aplicamos la correlación no paramétrica de Spearman (Tabla 4 y Tabla 5), que muestra que todas las correlaciones son significativas (la mayoría con máximo nivel de significatividad) a excepción de:

	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>													
	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>	<i>rs</i>	<i>p</i>													
Types	0.880 ***	—																																							
1-1000 tokens	0.983 ***	0.831 ***	—																																						
1-1000 types	0.818 ***	0.928 ***	0.809 ***	—																																					
1001-2000 tokens	0.644 ***	0.678 ***	0.553 ***	0.529 ***	—																																				
1001-2000 types	0.604 ***	0.703 ***	0.513 ***	0.529 ***	0.903 ***	—																																			
No 2000 tokens	0.746 ***	0.785 ***	0.646 ***	0.620 ***	0.632 ***	0.614 ***	—																																		
No 2000 types	0.734 ***	0.813 ***	0.639 ***	0.607 ***	0.696 ***	0.699 ***	0.916 ***	—																																	
A1+A2	0.834 ***	0.922 ***	0.815 ***	0.943 ***	0.606 ***	0.598 ***	0.647 ***	0.631 ***	—																																
≥ B1	0.582 ***	0.712 ***	0.492 **	0.512 ***	0.605 ***	0.663 ***	0.767 ***	0.886 ***	0.446 **	—																															
AG 1000	0.407 **	0.580 ***	0.276	0.329 *	0.763 ***	0.818 ***	0.722 ***	0.835 ***	0.368 *	0.842 ***	—																														
AG 2000	0.009	0.616 ***	0.085	0.038	0.564 ***	0.579 ***	0.823 ***	0.931 ***	0.386 *	0.893 ***	0.881 ***	—																													
AG ≥ B1	0.309	0.501 ***	0.213	0.296	0.472 **	0.560 ***	0.612 ***	0.763 ***	0.209	0.936 ***	0.841 ***	0.870 ***	—																												
AG ≥ B2	0.282	0.447 **	0.197	0.282	0.510 ***	0.534 ***	0.497 **	0.642 ***	0.187	0.802 ***	0.766 ***	0.717 ***	0.863 ***	—																											
	0.078	0.004	0.224	0.078	< .001	< .001	< .001	< .001	0.247	< .001	< .001	< .001	< .001																												

* $p < .05$, ** $p < .01$, *** $p < .001$

TABLA 4. Correlación de Spearman: sofisticación léxica en el grupo criollo

- $AG \geq B1$ y $AG \geq B2$ en el grupo criollo: no se produce correlación significativa con la producción total (*tokens*), ni con la producción de *tokens* y *types* en la banda más frecuente del *LFP CAES* (1-1000) y del *LFP MCER* (A1+A2).
- $AG 1000$ y $AG 2000$ en el grupo telugu: no correlacionan significativamente con la producción total de *tokens*, ni con la producción de *tokens* y *types* en la banda más frecuente del *LFP CAES* (1-1000). $AG 2000$ tampoco correlaciona en este grupo con la producción total de *types*, de *types* menos sofisticados en el *LFP MCER* (A1+A2), ni de *tokens* en la segunda banda de mayor frecuencia del *LFP CAES* (1001-2000). En el grupo criollo sí se producen relaciones entre estas variables (algunas débiles), a excepción de $AG 1000$, que tampoco correlaciona con los *tokens* más frecuentes del *LFP CAES* (1-1000).

La asunción de Malvern *et al.* (2004) de relación entre diversidad (mayor producción de *types*) y sofisticación (mayor producción de *types* sofisticados) es cierta en la medida en que la mayor producción aumenta la producción de *tokens* y *types* en todas las bandas de frecuencia (Sanhueza, Ferreira & Sáez, 2018), como demuestra el alto nivel de significatividad y la fortaleza de las correlaciones entre la producción (*tokens* y *types*) y las diferentes bandas tanto del *LFP CAES* como del *LFP MCER*. Ahora bien, la asunción contraria de que el incremento de los *types* no produce un mayor incremento de palabras sofisticadas, sino una mayor variedad dentro de las 1000 más frecuentes (Laufer, 1998; Horst & Collins, 2006; Tidball & Treffers-Daller, 2008; Sanhueza, Ferreira & Sáez, 2018) es, asimismo, parcialmente correcta: en el presente estudio el incremento de los *types* produce un incremento de *tokens* y *types* sofisticados en ambos grupos (No en 2000, correlación especialmente fuerte en el grupo criollo, y $\geq B1$, correlación especialmente fuerte en el telugu), si bien, efectivamente, el incremento en las bandas de menor sofisticación (1-1000 y A1+A2) presenta una relación mucho más fuerte (casi perfecta en la mayoría de los casos) con la producción tanto de *tokens*

como de *types*. Es decir, al incrementar la producción se incrementan los *types* y *tokens* en todas las bandas proporcionalmente, por lo que el mayor incremento se produce en la banda de mayor frecuencia o más básica (1-1000 y A1+A2), dado que es la banda que concentra el mayor porcentaje de la producción.

Para confirmar estadísticamente la hipótesis de la influencia de la LM/L2 en las diferencias, y dada la falta de distribución normal, realizamos la prueba no paramétrica de Mann-Witney (Tabla 6). Las diferencias según la LM/L2 son significativas ($p < ,05$) en la mayor producción de *types*, de *types* no sofisticados (tanto 1-1000 como A1+A2) y de *tokens* de la banda 1001-2000 del grupo telugu, así como en la mayor producción de *types* sofisticados ($\geq B1$) y los mayores *AG* (*AG* 2000, *AG* $\geq B1$ y *AG* $\geq B2$) del grupo criollo. Esto nos permite confirmar la hipótesis de que la LM criollo/L2 portugués influye en el mayor nivel de sofisticación medido mediante *LFP MCER* y *AG* (a excepción del *AG* 1000). Sin embargo, en el *LFP CAES* ni los *tokens* de la banda de mayor frecuencia (1-1000) ni los *tokens* y *types* de la de mayor sofisticación (No en 2000) son significativos según la LM/L2, por lo que para estas medidas rechazamos H_1 y aceptamos H_0 ; que no existe diferencia basada en la LM/L2.

		W	p	r_b
	Tokens	687.500	0.281	-0.141
	Types	527.500	0.009	-0.341
LFP CAES	1-1000 tokens	662.000	0.186	-0.172
	1-1000 types	430.000	< .001	-0.463
	1001-2000 tokens	554.500	0.018	-0.307
	1001-2000 types	634.500	0.111	-0.207
	No 2000 tokens	918.500	0.256	0.148
	No 2000 types	959.500	0.125	0.199
	LFP MCER	A1+A2	325.000	< .001
$\geq B1$		1111.500	0.003	0.389
CAES	AG 1000	919.500	0.252	0.149
	AG 2000	1051.000	0.016	0.314
MCER	AG $\geq B1$	1242.000	< .001	0.552
	AG $\geq B2$	1310.000	< .001	0.637

TABLA 6. Mann-Witney: variables de sofisticación léxica

Contrariamente a otros estudios (Sahuenza, Ferreira & Sáez, 2018; Schmitt, Schmitt & Clapham, 2001), los *tokens* de la segunda banda (1001-2000) sí resultan significativos en la discriminación de la LM/L2 para definir la menor sofisticación.

5. DISCUSIÓN Y CONCLUSIONES

Dada la inexistencia de estudios de sofisticación léxica sobre la adquisición del ELE, establecemos la comparación y discusión de los resultados respecto al *LFP* de Laufer & Nation (1995) para el inglés LE y respecto a la versión francesa del *LFP* (*Vocabprofil*) y al *AG* con diferentes operacionalizaciones de Tidball y Treffers-Daller (2008) para el francés LE. Como se observa en la Tabla 7, los resultados varían considerablemente entre los diferentes estudios, lo que demuestra los efectos de la metodología (ej. lematización, lista de frecuencias empleada, etc.) y limita la comparación.

El porcentaje de vocabulario (*tokens*) no sofisticado (1-2000) de nuestro estudio (90,27% para el grupo criollo y 92,07% para el telugu) se sitúa entre el nivel intermedio-bajo (93,7% para la composición 1 y 94,5% para la 2)²² e intermedio (86,4% para la composición 1 y 86,2% para la 2) del estudio de Laufer & Nation, acercándose el grupo telugu más al nivel intermedio-bajo. Sin embargo, tanto el grupo criollo y como el telugu emplean un porcentaje menor de vocabulario frecuente (no sofisticado) respecto a cualquiera de los grupos considerados en el estudio de Tidball & Treffers-Daller, incluyendo el nativo, cuyo porcentaje (93,82%) corresponde al nivel intermedio-bajo no nativo del estudio de Laufer y Nation.

Lo mismo ocurre respecto al vocabulario sofisticado (No en 2000 en nuestro estudio, la suma de *UWL* y *NOL* en el de Laufer & Nation y de las bandas 2001-3000 y *NOL* en el de Tidball y Treffers-Daller): el grupo criollo (9,73%) y el grupo telugu (7,93%) se sitúan entre los niveles intermedio-bajo (6,5% en la composición 1 y 6,9% en la 2) e intermedio (13,7% en la composición 1 y 14,4% en la 2) del estudio de Laufer y Nation, y por encima del grupo nativo de Tidball y Treffers-Daller, cuyo porcentaje (6,55%) es equiparable al nivel intermedio-bajo de este último.

²² Se recogen dos composiciones (1 y 2) por nivel.

Presente estudio: <i>LFP CAES</i>													
Nivel	N	1-1000			1001-2000			1-2000			<i>No en 2000</i>		
		CR	TEL	TEL	CR	TEL	TEL	CR	TEL	CR	TEL	CR	TEL
	40	85,18%	86,07%	5,08%	6,01%	92,07%	90,27%	92,07%	-	-	-	9,73%	7,93%
Laufer & Nation (1995): <i>LFP</i>													
Nivel	N	1-1000			1001-2000			1-2000			<i>UWL + NOL</i>		
		1	2	2	1	2	2	1	2	1	2	1	2
1 (Int-bajo)	22	86,6%	87,50%	7,10%	7,00%	94,50%	93,70%	94,50%	3,20%	4,10%	3,30%	2,80%	6,90%
2 (Int)	20	79,70%	79,40%	6,70%	6,80%	86,20%	86,40%	86,20%	8,10%	7,80%	5,60%	6,60%	13,70%
3 (Int-alto)	23	77%	74%	6,6%	5,6%	79,60%	83,60%	79,60%	8,10%	10,10%	7,50%	8,70%	15,60%
Tidball & Treffers-Daller (2008): <i>Vocabprofil</i>													
Nivel	N	1-1000			1001-2000			1-2000			<i>NOL</i>		
		1	2	2	1	2	2	1	2	1	2	1	2
1	19	92,77%	4,47%	97,24%	4,47%	97,24%	0,69%	2,07%	2,07%	2,76%	2,76%	2,76%	
3	20	90,87%	4,61%	95,48%	4,61%	95,48%	0,15%	4,37%	4,37%	4,52%	4,52%	4,52%	
Nativos	25	88,83%	4,99%	93,82%	4,99%	93,82%	0,92%	5,63%	5,63%	6,55%	6,55%	6,55%	

TABLA 7. Comparación de la sofisticación léxica con los resultados de Laufer & Nation (1995) y Tidball & Treffers-Daller (2008)

Las diferencias cuantitativas del estudio de Tidball & Treffers-Daller se deben a cuestiones metodológicas, principalmente la aplicación de la versión automatizada del *LFP* para el francés (*Vocaprofil*) basada en datos escritos al análisis de la producción oral, más espontánea y con tendencia al uso de un vocabulario más común y frecuente. Además, entre los tres estudios se producen diferencias a nivel de lematización y objetivos que limitan la comparabilidad.

En el objetivo que nos ocupa en el presente estudio, la mayor sofisticación léxica del grupo criollo se confirma independientemente del tipo de operacionalización, si bien el rango entre ambos grupos aumenta al considerar los *types* en lugar de los *tokens*, especialmente si estos están clasificados según los niveles del *MCER*, lo cual también se refleja en los *AG*. La significatividad de la influencia de la LM/L2 en las medidas de sofisticación ($types \geq B1$, $AG \geq B1$, $AG \geq B2$ y $AG \geq 2000$) permite concluir la influencia de la afinidad lingüística en la mayor sofisticación léxica en ELE.

A nivel metodológico, pese a que la lista extraída del CAES corresponde a un corpus de producción escrita de hablantes de ELE, en el presente estudio es la clasificación basada en los niveles del *MCER* la que ha permitido discriminar entre ambos grupos en base a la afinidad lingüística de su LM/L2.

Como han apuntado otros estudios (Laufer & Paribakht, 1998; Horst & Collins, 2006; Naismith *et al.*, 2018), el uso de cognados de la LM/L2 por parte del grupo criollo frente a otras palabras más frecuentes en español (pero más lejanas a su LM/L2), así como de arcaísmos del español que forman parte del vocabulario común del portugués (ej. *acontecer*, *climatérico*, *feriado*, etc.), parecen tener una influencia clara en el nivel de sofisticación. No se puede negar, sin embargo, la ventaja de la que parte el lusófono en el aprendizaje del español, dado el 90% de caudal léxico compartido entre ambas lenguas (Torijano, 2002, p. 389) que favorece la transferencia positiva. No obstante, el empleo estratégico de esta similitud es un arma de doble

filo, como demuestran los abundantes errores de transferencia negativa por el uso de *falsos cognados*, extranjerismos, etc., tanto en la presente investigación como en otros análisis de aprendientes de español con LM portugués (Benedetti, 2001; Sánchez Rodríguez, 2001; Torijano, 2002; Durão, 2004; Campillos, 2012).

El hecho de que el grupo criollo esté compuesto por estudiantes universitarios (con una media de edad menor) y el grupo telugu por usuarios (de mayor edad media) es una variable que influye en los resultados en relación directa con las estrategias empleadas, el estadio de adquisición y el tipo de uso y exposición al idioma: el grupo criollo se encuentra en un estadio de exposición inicial y, ante el desconocimiento de la LO, recurre a la transferencia como estrategia habitual obteniendo tanto resultados positivos (como, por ejemplo, una mayor sofisticación léxica empleando palabras del portugués iguales o similares en español) como negativos (empleo masivo de extranjerismos y gran número de errores de formales y ortográficos producidos mayoritariamente por interferencia); el grupo telugu no recibía instrucción en español en el momento de la recolección de datos, si bien gran parte de su trabajo consistía en traducir cartas de telugu a español (y, en menor medida, también a la inversa) que escriben los niños/as de LM telugu a sus respectivos padrinos y madrinas españoles, por lo que manejan un vocabulario básico, simple, rutinario, estandarizado y propio del discurso infantil. Siendo su trabajo fundamental la producción escrita en español²³, cabría esperar una competencia escrita alta. Sin embargo, la simplicidad del contenido y el alto nivel de estandarización, automatización y repetición provocan lo contrario: ausencia de desarrollo y complejidad en la producción escrita e inhibición de los mecanismos creativos del lenguaje, la capacidad de argumentación, el manejo de géneros discursivos

²³ La mayoría de los miembros del grupo también desenvuelven la función de acompañantes o guías de los españoles que visitan la ONGD donde trabajan. Pese a que en este caso se produce interacción con nativos, el contenido lingüístico sigue siendo rutinario y memorístico, dado que se repiten similares explicaciones y preguntas.

diversos etc., lo que tiene un efecto directo en el nivel de sofisticación léxica y en una competencia lingüística reducida y fosilizada.

Existen, además, variables individuales (edad, nivel de estudios, conocimiento de otras lenguas, competencia en la propia LM/L2, nivel socioeconómico y sociocultural, etc.) y metodológicas (tema y género discursivo de la prueba empleada para la recolección de datos, lematización y criterios empleados, etc.) que influyen en los resultados. Sin embargo, no cabe duda de que la LM/L2 es una variable determinante y de que la transferencia juega un papel clave, tal como demuestran las variaciones interlingüísticas registradas en la adquisición de una misma L2 entre sujetos con diferentes LM debidas, precisamente, a la diferencia en las estructuras y complejidad de la propia LM, incluso entre lenguas con alto grado de afinidad como las romances (Gyllstadt, Granfeldt, Bernardini & Källkvist, 2014; De Clercq & Housen, 2017; Kuiken & Vedder, 2019). La diferencia en el nivel de sofisticación basada en la LM ya había sido demostrada (Naismith *et al.*, 2018) y, atendiendo a nuestros resultados, una mayor afinidad lingüística entre la LM/L2 y la LE determina un mayor nivel de sofisticación. Estos resultados cobran todavía más peso si tenemos en cuenta que la cantidad de instrucción y tiempo de exposición al español ha sido mucho menor en el grupo de lengua afín (aunque con *input* más variado y de mayor calidad). Esto respalda la pertinencia de una metodología específica la adquisición de lenguas en función del nivel de afinidad lingüística y del estadio de aprendizaje. Conviene explotar el potencial del caudal léxico compartido entre lenguas afines empleando la reflexión contrastiva explícita como herramienta didáctica que permita adelantarse a la comparación que inevitablemente realizará el aprendiente para guiarlo en el desarrollo de su conciencia metalingüística, favoreciendo así la transferencia positiva y evitando la interferencia para prevenir la fosilización (Calvi, 2004; Doquin de Saint-Preux, 2008). El desarrollo de esta consciencia metalingüística es igualmente importante en la didáctica de lenguas no afines. El caso del grupo telugu pone en relieve la importancia de que la formación en LE

sea continua o, al menos, periódica, y específica para tratar la fosilización y el desarrollo específico de la competencia léxica y discursiva.

Con este trabajo hemos pretendido contribuir empíricamente a un campo apenas explorado. No obstante, es necesaria más investigación en sofisticación léxica de ELE para corroborar nuestros resultados incluyendo otras LM, niveles de competencia, tipos de discurso y grados de afinidad lingüística, para poder responder a preguntas como ¿en qué medida está determinada por el grado de afinidad lingüística?, ¿y por el tipo de discurso?, ¿cómo evoluciona en los diferentes niveles de competencia?, ¿discrimina entre niveles de competencia?, ¿entre tipos de discurso?, ¿entre diferentes LM?, ¿y entre hablantes de ELE y nativos?, ¿qué otras variables influyen en la sofisticación? Son muchas las preguntas y prácticamente nulas las respuestas empíricas actuales en lo que a español se refiere.

Con base en la literatura revisada, también es necesario definir claramente los criterios respecto a la lematización realizada (dado que emplear la familia léxica o la flexión comporta cambios relevantes en los resultados) y la forma de proceder en cuanto a extranjerismos, acuñaciones, deformaciones²⁴, etc., y homogeneizarlos para posibilitar la comparación científica entre diferentes estudios. La lista empleada, como hemos visto, es determinante, por lo que sería necesario un trabajo previo de validación de listas para ELE, como el realizado por Naismith *et al.* (2018) para el inglés LE, y de comprobación del poder de discriminación de diferentes operacionalizaciones con diferentes listas entre las que se incluyan también las basadas en juicio de expertos, como el de Tidball y Treffers-Daller (2008).

²⁴ Por ejemplo, Laufer y Nation (1995) corrigen las palabras con errores ortográficos (que, de lo contrario, no serían reconocidas y pasarían a formar parte de la banda *NOL*, aumentando engañosamente el nivel de sofisticación), mientras otros autores como Naismith *et al.*, (2018) las eliminan, lo que, como reconocen los propios autores, tiene un impacto determinante en su estudio. Cabría aquí una discusión sobre si el desconocimiento de la forma escrita exacta de una palabra puede ser considerado equivalente al desconocimiento de esta, es decir, no perteneciente al lexicón mental del hablante.

Asimismo, es recomendable combinar diferentes medidas tanto léxicas como sintácticas (híbridas algunas de ellas), correlacionadas y no correlacionadas, para dar cuenta de la complejidad en su multidimensionalidad e incrementar la validez y fiabilidad de los resultados (Norris & Ortega, 2009; Pallotti, 2009; Bulté & Housen, 2012; Mavrou, 2016; Aravena & Quiroga, 2018) y emplear metodologías mixtas que permitan realizar análisis más profundos y completos que consideren también variables cualitativas.

Todos estos son requisitos previos al desarrollo de herramientas automatizadas que faciliten procesos de autoanálisis, autoaprendizaje y autoevaluación de lenguas imprescindibles en la era digital en la que estamos inmersos.

BIBLIOGRAFÍA

- Aravena, S. & Quiroga, R. (2018). Desarrollo de la complejidad léxica en dos géneros escritos por estudiantes de distintos grupos socioeconómicos. *Onomázein*, 42, 197-224. <https://doi.org/10.7764/onomazein.42.03>.
- Arnaud, P. J. (1984). The lexical richness of L2 written productions and the validity of vocabulary tests. En T. Culhane, C. Klein-Braley & D. K. Stevenson (eds.), *Practice and problems in language testing* (Vol. 29) (pp. 14-28). University of Essex Occasional Papers.
- Ávila, R. (1986). Léxico infantil de México: Palabras, tipos, vocablos. En J. G. Moreno de Alba (ed.), *Actas del Congreso del II Congreso Internacional sobre el español de América* (pp. 510-517). Universidad Nacional Autónoma de México.
- Benedetti, A. M. (2001). Interferencias semánticas del portugués en el aprendizaje de español. En J. Gómez Asencio, J. Sánchez Lobato & A. Larrañaga Domínguez (eds.), *Interferencias, cruces y errores* (pp. 9-24). SGEL.
- Berton, M. (2014). *La riqueza léxica en la producción escrita de estudiantes suecos de ELE* [Tesis de maestría. Stockholms Universitet]. DiVA. Digitala Vetenskapliga Arkivet. <https://cutt.ly/ELwX5i0>.
- Broeder, P., Extra, G. & Van Hout, R. (1993). Richness and variety in the developing lexicon. En C. Perdue (ed.), *Adult Language Acquisition. Volume II: The Results* (pp.145-163). Cambridge University Press.
- Bulté, B. & Housen, A. (2012). Defining and operationalising L2 complexity proficiency. Complexity, accuracy and fluency in SLA. En A. Housen, F. Kuiken & I. Vedder (eds.), *Dimensions of L2 Performance and Proficiency. Complexity, Accuracy and Fluency in SLA* (pp. 21-46). John Benjamins. <https://doi.org/10.1075/llt.32.02bul>.
- Calvi, M. V. (1998). La gramática en la enseñanza de lenguas afines. En T. Jiménez Juliá, M. C. Losada Andrey, M. Caneda, J. F. & S. Sotelo Docío (eds.), *Español como Lengua*

- Extranjera: Enfoque Comunicativo y Gramática*. Santiago de Compostela: Actas ASELE IX. Universidad de Santiago de Compostela (pp. 353-360). <https://cutt.ly/BLwCBWg>.
- Calvi, M. V. (2004). Aprendizaje de lenguas afines: español e italiano. *redELE: Revista Electrónica de Didáctica ELE*, 1, s.p. <https://cutt.ly/7LwV8Fx>.
- Campillos Llanos, L. (2012). *La expresión oral en español lengua extranjera: interlengua y análisis de errores basado en corpus*. [Tesis Doctoral. Universidad Autónoma de Madrid]. Biblos-e Archivo. <https://cutt.ly/GLwBkd7>.
- Cobb, T. & Horst, M. (2004). Is there room for an academic wordlist in French? En P. Boogards & B. Laufer (eds.), *Vocabulary in a second language: Selection, acquisition and testing* (pp. 15-38). John Benjamins. <https://doi.org/10.1075/llt.10.04cobb>.
- Consejo de Europa (2021). *Marco Común Europeo de Referencia para las lenguas (MCER): aprendizaje, enseñanza, evaluación*. Instituto Cervantes. Ministerio de Educación, Cultura y Deporte. Anaya. <https://cutt.ly/FLwNrijw>.
- Corder, S. P. (1971). Dialectos idiosincráticos y análisis de errores. En J. M. Liceras (coord.), *La adquisición de lenguas extranjeras* (pp. 63-77). Visor.
- Daller, H. & Phelan, D. (2007). What is in a teachers' mind? En M. T.-D. Daller (ed.), *Modelling and Assessing Vocabulary Knowledge* (pp. 234-244). Cambridge University Press. <https://doi.org/10.1017/CBO9780511667268>.
- Daller, H. & Xue, H. (2007). Lexical richness and the oral proficiency of Chinese EFL students. En M. Daller, J. Milton & J. Treffers-Daller (eds.), *Modelling and assessing vocabulary knowledge* (pp. 150-164). Cambridge University Press. <https://doi.org/10.1017/CBO9780511667268.011>.
- Daller, H., Van Hout, R. & Treffers-Daller, J. (2003). Lexical Richness in the Spontaneous Speech of Bilinguals. *Applied Linguistics*, 24/2, 197-222. <https://doi.org/10.1093/applin/24.2.197>.
- Daller, M., Turlik, J. & Weir, I. (2013). Vocabulary acquisition and the learning curve. En S. Jarvis & M. Daller (eds.), *Vocabulary Knowledge: Human ratings and automated measures* (pp. 185-218). John Benjamins. <https://doi.org/10.1075/sibil.47.09ch7>.
- De Clercq, B. & Housen, A. (2017). A cross-linguistic perspective on syntactic complexity in L2 development: Syntactic elaboration and diversity. *The Modern Language Journal*, 101(2), 315-334. <https://doi.org/10.1111/modl.12396>.
- Doquin Saint-Preux, A. (2008). *L'enseignement du français aux hispanophones: problèmes repérés, études linguistiques, propositions didactiques*. Tesis doctoral. Université Paris IV-Sorbonne.
- Duolingo. *CEFR Checker*. <https://cefr.duolingo.com> [19/11/2020].
- Durão, A. B. (2004). *Análisis de Errores en la interlengua de brasileños aprendices de español y de españoles aprendices de portugués* (2ª ed). Eduel.
- Ellis, R. (1985). *Understanding Second Language Acquisition*. Oxford University Press.
- Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford University Press.

- Fernández López, M. S. (1991). *Análisis de errores en interlengua en el aprendizaje del español lengua extranjera* [Tesis doctoral inédita. Universidad Complutense de Madrid].
- Ferrero García de Jalón, P. (2011). *Definición y análisis de parámetros lingüísticos cuantitativos para herramientas automáticas de evaluación de aplicables al español como lengua extranjera* [Tesis doctoral. Universidad Autónoma de Madrid]. Biblios-e Archivos. <https://repositorio.uam.es/handle/10486/10323>.
- Gass, S. & Selinker, L. (eds.). (1983). *Language Transfer in Language Learning*. Newbury House.
- Gyllstad, H., Granfeldt, J., Bernardini, P. & Källkvist, M. (2014). Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity in written L2 English, L3 French and L4 Italian. En L. Roberts, I. Vedder & J. Hulstijn (eds.), *Eurosla Yearbook*, Vol. 14 (pp. 1–30). John Benjamins.
- Horst, M. & Collins, L. (2006). From Faible to strong: how does their vocabulary grow? *The Canadian Modern Language Review*, 63/1, 83-106. <https://doi.org/10.3138/cmlr.63.1.83>.
- Instituto Cervantes (2014a). *Guía del examen DELE B1*. Consultado el 17 de junio de 2022. <https://cutt.ly/9Lw11NP>.
- Instituto Cervantes (2014b). *Guía del examen DELE B2*. Consultado el 17 de junio de 2022. <https://cutt.ly/ILw0k8e>.
- Instituto Cervantes (2014c). *Guía del examen DELE C1*. Consultado el 17 de junio de 2022. <https://cutt.ly/1Lw0QyW>.
- Instituto Cervantes (2019a). *Guía del examen DELE A1*. Consultado el 17 de junio de 2022. <https://cutt.ly/uLw0YM6>.
- Instituto Cervantes (2019b). *Guía del examen DELE A2*. Consultado el 17 de junio de 2022. <https://cutt.ly/jLw0LqS>.
- Instituto Cervantes; Universidad de Santiago de Compostela (USC), (2020 (2013)). *Corpus de aprendientes de español (CAES). Versión: 1.3 - abril 2020*. Consultado el 18 de julio de 2020. <https://galvan.usc.es/caes>.
- Jarvis, S. & Pavlenko, A. (2007). *Crosslinguistic Influence in Language and Cognition*. Routledge.
- Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: a developmental perspective. *Working Papers*, 53, 61-79. <https://cutt.ly/xLw2e3v>.
- Juffs, A. (2019). Lexical development in the writing of English Language Program Students. En R. DeKeyser & G. Botana (eds.), *Doing SLA Research with Implications for the Classroom: Reconciling pedagogical demands with pedagogical applicability* (pp. 179-200). John Benjamins. <https://doi.org/10.1075/llt.52.09juf>.
- Kellermann, E. (1977). Towards a characterization of the strategy of transfer in second language learning. *Interlanguage Studies Bulletin*, 2, 58-145.
- Kellermann, E. (1983). Now you see it, now you don't. En S. Gass & L. Selinker (eds.), *Language transfer in language learning* (pp. 112-134). Newbury House.

- Kuiken, F. & Vedder, I. (2019). Syntactic complexity across proficiency and languages: L2 and L1 writing in Dutch, Italian and Spanish. *International Journal of Applied Linguistics*, 29, 192-210. <https://doi.org/10.1111/ijal.12256>.
- Laufer, B. (1995). Beyond 2000. A measure of productive lexicon in a second language. En L. Eubank, L. Selinker & M. Sharwood Smith (eds.), *The Current State of Interlanguage. Studies in Honor of William E. Rutherford* (pp. 265-272). John Benjamins. <https://doi.org/10.1075/z.73.21lau>.
- Laufer, B. & Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics*, 16/3, 307-332. <https://doi.org/10.1093/applin/16.3.307>.
- Laufer, B. & Paribakht, T. S. (1998). The Relationship Between Passive and Active Vocabularies: Effects of Language Learning Context. *Language Learning*, 48/3, 365-391. <https://doi.org/10.1111/0023-8333.00046>.
- Liceras, J. M. (1986). Sobre el concepto de permeabilidad. *Revista española de lingüística aplicada*, 2, 49-61. <https://cutt.ly/SLw9oIM>.
- Lindqvist, C., Gudmundson, A. & Bardel, C. (2013). A new approach to measuring lexical sophistication in L2 oral production. En C. Bardel, C. Lindqvist & B. Laufer (eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 109-126). Eurosla. <https://cutt.ly/SLw9E6C>.
- Linnarud, M. (1986). *Lexis in Composition. A Performance Analysis of Swedish Learners' Written English*. Liber Förlag.
- López Morales, H. (2011). Los índices de riqueza léxica y la enseñanza de lenguas. En J. Santiago Guervós, H. Bongaerts, J. J. Sánchez Iglesias & M. Seseña Gómez (eds.), *Del texto a la lengua. La aplicación de los textos a la enseñanza-aprendizaje del español L2-LE.1* (pp. 15-28). Asociación para la Enseñanza del Español como Lengua Extranjera. <https://cutt.ly/0Lw96uz>.
- Malvern, D., Richards, B., Chipere, N. & Durán, P. (2004). *Lexical Diversity and Language Development: Quantification and Assessment*. Palgrave Macmillan. <https://doi.org/10.1057/9780230511804>.
- Mavrou, I. (2016). Complejidad, precisión, fluidez y léxico: Una revisión. *Moderna språk*, 1, 49-69. <https://cutt.ly/NLw3djz>.
- McCarthy, M. (1998). *Spoken language and applied linguistics*. Cambridge University Press.
- Meara, P. & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospects*, 16/3, 5-19. <https://cutt.ly/CLw3nAX>.
- Naismith, B., Han, N. R., Juffs, A., Hill, B. & Zheng, D. (2018). Accurate Measurement of Lexical Sophistication with Reference to ESL Learner Data. En K. E. Boyer, M. Yudelson (eds.) *Proceedings of the 11th International Conference on Educational Data Mining* (pp. 259-265). <https://cutt.ly/yGG6TOK>.
- Nation, P. (2001). *Learning Vocabulary in Another Language*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139524759>.

- Norris, J. N. & 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>.
- Odlin, T. (1989). *Language Transfer. Cross-linguistic influence in language learning*. Cambridge Applied Linguistics.
- Pallotti, G. (2009). CAF: defining, refining and differentiating constructs. *Applied Linguistics*, 30(4), 590-601. <https://doi.org/10.1093/applin/amp045>.
- Read, J. (2000). *Assessing Vocabulary*. Cambridge University Press.
- Richards, B. & Malvern, D. (2000). *Measuring vocabulary richness in teenage learners of French*. *British Educational Research Association Conference*. Cardiff University.
- Ringbom, H. (1987). *The role of the first language in foreign language learning*. Multilingual Matters.
- Sahuenza, C., Ferreira, A. & Sáez, K. (2018). Desarrollo de la competencia léxica a través de estrategias de aprendizaje de vocabulario en aprendientes de inglés como lengua extranjera. *Lexis*, XLII/2, 273-326. <https://doi.org/10.18800/lexis.201802.002>.
- Sánchez Iglesias, J. J. (2003). *Errores, corrección y fosilización en la didáctica de lenguas afines: análisis de errores en la expresión escrita de estudiantes italianos de E/LE* [Tesis Doctoral. Universidad de Salamanca]. Repositorio Documental GREDOS. <https://cutt.ly/ELqfzWk>.
- Sánchez Rodríguez, J. (2001). Interferencias y dificultades en el aprendizaje del español de alumnos portugueses (análisis y comparaciones de dos niveles de aprendizaje). En J. Gómez Asencio & J. Sánchez Lobato (eds.), *Interferencias, cruces y errores* (pp. 25-47). SGEL.
- Santos Gargallo, I. (1993). *Análisis contrastivo, Análisis de errores e interlengua en el marco de la lingüística contrastiva*. Síntesis.
- Schmitt, N., Schmitt, D. & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, 18/1, 55-88. <https://doi.org/10.1177/026553220101800103>.
- Sharwood Smith, M. & Kellerman, E. (1986). *Crosslinguistic Influence in Second Language Acquisition*. Pergamon Press.
- SIELE. (2017). *Guía oficial para preparadores del SIELE*. Consultado el 17 de junio de 2022. <https://cutt.ly/OLw4wJI>.
- Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30(4), 510-532. <https://doi.org/10.1093/applin/amp047>.
- Swan, M. (1997). The influence of the mother tongue on second language vocabulary acquisition and use. En N. Schmitt, & M. McCarthy (eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 156-180). Cambridge University Press.

- Tidball, F. & Treffers-Daller, J. (2008). Analysing lexical richness in French learner language: what frequency lists and teachers judgements can tell us about basic and advanced words. *French Language Studies*, 18/3, 299-313. <https://doi.org/10.1017/S0959269508003463>.
- Torijano Pérez, J. A. (2002). *Análisis teórico-práctico de los errores gramaticales en el aprendizaje del español, L-2: expresión escrita* [Tesis doctoral inédita. Universidad de Salamanca].
- Van Hout, R. & Vermeer, A. (2007). Comparing measures of lexical richness. En H. Daller, J. Milton & Treffers-Daller (eds.), *Modelling and assessing vocabulary knowledge*. (pp. 93-115). Cambridge University Press. <https://doi.org/10.1017/CBO9780511667268>.
- Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, 17/1, 65-83. <https://doi.org/10.1177/026553220001700103>.
- Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. En P. Boogaards & B. Laufer (eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 173-189). John Benjamins. <https://doi.org/10.1075/llt.10.13ver>.
- Wolfe-Quintero, K., Inagaki, S. & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity* (Vol. 17). University of Hawaii Press.