This is an Accepted Manuscript of a book chapter published by Routledge/CRC Press in:
Lozano, C. (2021). Corpus textuales de aprendices para investigar sobre la adquisición del español LE/L2. in M. Cruz Piñol (Ed.), E-Research y español LE/L2: Investigar en la era digital (pp. 138-163). Routledge. http://doi.org/10.4324/9780429433528-9

7

Corpus textuales de aprendices para investigar sobre la adquisición del español LE/L2

Cristóbal Lozano

https://orcid.org/0000-0003-0068-154X

The increasing interest in the research on the acquisition of Spanish as a second language (L2) has led to the creation of large language databases (learner corpora). Such corpora allow researchers to investigate the type of knowledge or competence (interlanguage) that learners can acquire in their L2 Spanish. We will justify the need for good learner corpus design and will illustrate how corpus methods help researchers understand typical L2 Spanish interlanguage phenomena in a systematic way. An overview of freely available L2 Spanish corpora will be presented along with some representative L2 Spanish acquisition studies that investigate several interlanguage phenomena. Finally, we will do hands-on practice with two software tools: Antconc, which allows to do concordance searchers on the corpora, and UAM Corpus Tool, which allows to tag and statistically analyse the corpora. Both tools will be illustrated with samples from the CEDEL2 corpus.

Resumen

El auge de la investigación de la adquisición del español como segunda lengua (L2) ha hecho necesaria la creación de amplias muestras del lenguaje (corpus de

aprendices). Dichos corpus permiten investigar el tipo de conocimiento o competencia que adquieren los aprendices de español L2. En este capítulo, se justificará la necesidad del buen diseño de un corpus de aprendices y se ilustrará cómo los métodos de corpus permiten al investigador comprender de manera sistemática y objetiva los fenómenos propios del lenguaje (interlengua) de los aprendices de español L2. Se presentará asimismo una panorámica de los corpus de español como L2 disponibles gratuitamente en línea, con algunos estudios de investigación representativos de la adquisición del español L2. Finalmente, nos centraremos en casos prácticos del uso de tres paquetes de software gratuito: la interfaz web de búsquedas del corpus CEDEL2 y el software AntConc, que permiten hacer búsquedas de concordancias, y UAM Corpus Tool, que permite etiquetar y analizar estadísticamente los corpus. Estas herramientas se ilustrarán con datos procedentes del corpus CEDEL2.

7.1 Necesidades

7.1.1 Observar la adquisición del español como LE/L2 en contextos naturales

El auge experimentado en la enseñanza del español como lengua extranjera ha venido acompañado por un interés en la investigación del español como segunda lengua (L2) (Montrul 2004; Geeslin 2014). Si bien en la investigación de adquisición de segundas lenguas (ASL) ha sido común emplear métodos experimentales con alto control pero baja naturalidad, recientemente los investigadores están optando por estudiar la L2 en contextos naturales y menos restrictivos, como es el caso de los corpus (Myles 2005, 2015; Granger 2012; Lozano y Mendikoetxea 2013).

7.1.2 Analizar la interlengua de los aprendices a partir de grandes muestras

El creciente interés en la investigación de español LE/L2 ha venido acompañado de la necesidad de analizar amplias muestras del lenguaje (o interlengua) de los aprendices de español L2 como fuente de datos.[1] En ASL, ha sido común basar las conclusiones en datos de naturaleza experimental en los que se investiga, a veces con muestras limitadas, lo que los aprendices aceptan como (a)gramatical. Sin embargo, es necesario también analizar amplias muestras de producción natural. Entre dichos fenómenos de la interlengua (cf. Geeslin 2014) se encuentra la adquisición de: las alternancias del orden de palabras Adjetivo-Nombre (1a) y Sujeto-Verbo (1b); los contrastes *ser* y *estar* (1c, 1d); el orden de los pronombres objeto átonos (1e); el marcado de dativo con preposición *a* en humanos (1f) pero no en inanimados (1g); la distribución de *por/para* (1h); los pronombres personales nulos y plenos (*él*) en posición sujeto (1j); tiempo y aspecto como el pretérito imperfecto/perfecto simple (1k, 1l); el subjuntivo (1m); las colocaciones (1n, 1ñ); etc. Éstos y otros fenómenos lingüísticos pueden ser investigados cómoda y eficazmente con datos de corpusya que contienen datos naturales y contextualizados.[2]

(1) a. La mesa redonda/*La redonda mesa

b. Un fantasma apareció/Apareció un fantasma

c. El examen es fácil/*El examen está fácil

d. *Mi madre es enferma/Mi madre está enferma

e. Juan la persigue/*Juan persigue la

f. Ayer vi a tu hermano/*Ayer vi tu hermano

g. *Ayer vi a tu coche/Ayer vi tu coche

h. La cama es para dormir/*La cama es por dormir

j. Juan es actor y vive en Hollywood/#Juan es actor y él vive en Hollywood

k. El bebé dormía/*durmió cuando llegué

l. El bebé *dormía/durmió durante todo el fin de semana

m. Cuando *nieva/nieve, jugaré en la calle

n. Tomar/*hacer una decisión

ñ. Hacer/*tomar un examen

Este capítulo se centra en diversas técnicas y herramientas para abordar la investigación de la adquisición del español LE/L2: los corpus y su diseño, el software de concordancias y etiquetado. Sólo si se aborda la investigación sistemática y rigurosamente se podrán sacar conclusiones sólidas sobre el proceso de adquisición y recomendaciones fiables para la docencia de ELE. Este enfoque es típico en la investigación basada en datos (*data-driven research*), que se explicaen los capítulos 1 y 6.

7.2 Cómo ayudan las tecnologías

En esta sección exploraremos las ventajas y limitaciones de las tecnologías de corpus de L2 así como los principios fundamentales del diseño de corpus.

7.2.1 Sobre el concepto de corpus de aprendices

El uso de grandes corpus de aprendices en ASL se enmarca dentro de la denominada Investigación de Corpus de Aprendices (*Learner Corpus Research*, LCR), que hunde sus raíces en la lingüística de corpus (Mendikoetxea 2014; Granger, Gilquin y Meunier 2015). Aunque existen varias definiciones de lo que es un corpus de L2, todas ellas coinciden en que es una colección informatizada y sistemática de textos (orales o escritos), producidos en contextos naturales por aprendices de una L2 y cuyo diseño sigue ciertos criterios. Los corpus de L2 tienen un gran número de usuarios y de usos (Díaz-Negrillo y Thompson 2013): investigadores de adquisición de L2, profesores de LE para preparar actividades pedagógicas, lexicógrafos y gramáticos para crear obras de consulta basadas en datos de corpus (diccionarios, gramáticas,

etc.), lingüistas computacionales para modelar el lenguaje de los aprendices, etc. Este capítulo está pensado principalmente (aunque no exclusivamente) para investigadores de español L2/LE.

7.2.2 Ventajas y limitaciones de la investigación con corpus de aprendices

El uso de grandes corpus para la investigación de la L2 ofrece numerosas ventajas (Myles 2005, 2007, 2015; Granger 2009; Lozano y Mendikoetxea 2013; Callies 2015; Granger, Gilquin y Meunier 2015; Mendikoetxea 2014). En primer lugar, permiten analizar de forma sistemática amplias muestras del lenguaje producidas por los aprendices (y también por los nativos). La evidencia lingüística, por tanto, procede de datos de producción natural y no de intuiciones (aunque ambos tipos de datos se complementan, cf. la conclusión en el apartado 7.4). Una ventaja de los corpus de L2 con distintas L1 es el Análisis Contrastivo de la Interlengua (*Contrastive Interlanguage Analysis*, CIA) (Granger 2015). Se pueden contrastar sistemáticamente dos variedades de interlengua para investigar el efecto de la L1, p.ej., L1 inglés-L2 español avanzado vs. L1 chino-L2 español avanzado. Además, si el diseño del corpus es bidireccional (cf. Lozano, Díaz-Negrillo y Callies 2020) es decir, que permita comparar dos lenguas en ambas direcciones (L1 inglés-L2 español vs. L1 español-L2 inglés), podremos determinar si los resultados se deben a la influencia específica de la L1 o tal vez a algún mecanismo cognitivo más general e independiente de la L1.

Metodológicamente, el software de corpus ofrece ventajas, aunque también algunas limitaciones. Como veremos más abajo (7.3.2), el software de concordancias permite estudiar patrones léxicos y léxicosintácticos en su contexto (*KeyWords In Context*, KWIC), amén de listas de frecuencias léxicas, combinaciones de palabras y colocaciones (véanse detalles de esta metodología en Cruz Piñol 2012 y el capítulo de Buyse en este volumen). No obstante, el tipo de software limita el fenómeno a investigar (Lozano 2015). Las investigaciones de español L2/LE

basadas en software de concordancias se han centrado predominantemente en la adquisición del léxico y colocaciones. Esto puede suplirse con corpus etiquetados (cf. sección 7.3.3).

Otra limitación es que la ausencia de evidencia no es evidencia de ausencia. El corpus siempre corresponderá a una pequeña muestra del lenguaje (por muy grande que el corpus sea). El no encontrar el fenómeno investigado en el corpus no implica necesariamente su desconocimiento por parte del aprendiz. Puede deberse a que el corpus no representa fielmente todo el potencial del lenguaje humano, o bien a que el aprendiz esté evitando dicha estructura por determinados motivos, o simplemente a que no se han creado contextos obligatorios para producir dicho fenómeno. Por ello, es imprescindible a veces triangular los datos de corpus con los datos experimentales (cf. 7.4) para llegar a una mejor comprensión del fenómeno en cuestión, como se demuestra en diversos estudios de corpus de L2 (Tracy-Ventura y Myles 2015; Mendikoetxea y Lozano 2018).

Los investigadores principiantes de LE/L2 suelen enfrentarse a dos retos: la escasa formación técnica en metodologías y software de corpus y también en diseño de experimentos (ambos necesarios para entender los fenómenos lingüísticos desde distintas ópticas) y la escasa formación teórica sobre los aspectos (psico)lingüísticos que rigen la adquisición de L2, lo cual induce al investigador a atribuir muchos de los errores del aprendiz a mera influencia de su L1, cuando en realidad pueden deberse a otros procesos cognitivos universales.

7.2.3 El diseño de corpus de aprendices

Una mera colección de textos o exámenes de aprendices de L2 no es necesariamente un corpus, pues puede no contener suficiente información necesaria para la investigación de la adquisición de L2. Antes de la recogida de datos, es imprescindible un buen diseño. Partiendo de principios generales de diseño de corpus (Sinclair 2005; Hincapié Moreno y Bernal Chávez 2018), éstos se aplicarán al diseño de corpus de L2 (Granger 2008; Lozano y Mendikoetxea 2013; Gilquin 2015; Tono 2016) (tabla 7.1), de tal manera que el corpus contenga variables precisas y detalladas

sobre el perfil del aprendiz (tabla 7.2) y de la tarea (tabla 7.3). Sólo de este modo el investigador podrá explorar el efecto de ciertas variables que son clave en la investigación de ASL.

7.2.3.1 Principios básicos del diseño de corpus de L2

El primer paso consiste en establecer los principios generales de diseño de corpus de L2. Partiendo del trabajo de Sinclair (2005), Lozano y Mendikoetxea (2013) y Lozano (enviado 2020) aplicaron diez principios generales al diseño de corpus de aprendices de L2, aunque exploraremos sólo los esenciales (tabla 7.1).

Tabla 7.1 Principios generales de diseño de corpus de L2.

Fuente: adaptado de Lozano (2009a, enviado 2020) y de Lozano y Mendikoetxea (2013).

Principio	**Descripción**
Selección de contenido	El contenido (lenguaje) del corpus ha de atenerse a criterios externos (función comunicativa de los textos) y no a criterios internos (la lengua de los textos *per se*). Es decir, no se ha de diseñar un corpus de L2 para suscitar fenómenos particulares (p.ej., voz pasiva), sino que los textos deberían de ser producidos por los aprendices sin restringir su lenguaje.
Represent atividad	El corpus debería ser lo más representativo posible del lenguaje que se esté muestreando. Si deseamos diseñar un corpus argumentativo, el corpus contendrá sólo textos argumentativos y no de otro tipo (p.ej., descriptivos).

<table>
<tr>
<td>Contraste</td>
<td>
De los diversos componentes/subcorpus del corpus principal, sólo han de contrastarse aquellos que han sido diseñados para serlo. Esto permite el análisis denominado CIA (cf. sección previa). Por tanto, todos los subcorpus seguirán los mismos principios de diseño y recogerán las mismas variables.

Es imprescindible incluir un subcorpus de control (corpus de nativos) para usarlo como corpus normativo o de referencia, y así determinar si el comportamiento atípico de los aprendices es resultado de su desarrollo o tal vez un reflejo de las normas nativas (que pueden o no coincidir con las normas prescriptivas). Finalmente, todos los subcorpus habrán de estar equilibrados, de tal manera que si tenemos tres variables generales de diseño (nivel de proficiencia: principiante/intermedio/avanzado; modalidad de la tarea: oral/escrita; L1 de los aprendices: inglés/chino), cada celda esté equilibrada en términos del número de participantes, textos y palabras (cf. tabla infra). Esto permite hacer contrastes lingüísticos fiables entre subcorpus. Por tanto, un corpus con muchos subcorpus pero con pocas palabras en algunas celdas o con celdas desequilibradas violaría el principio de representatividad.

<table>
<tr><th>L1-L2</th><th>Nivel</th><th>Modalidad</th><th>Cantidad (textos; palabras)</th></tr>
<tr><td>L1 inglés-</td><td>Principiante</td><td>Oral</td><td>50; 20.000</td></tr>
</table>
</td>
</tr>
</table>

	L2 español		Escrita	100; 50.000
		Intermedio	Oral	47; 19.000
			Escrita	102; 51.000
		Avanzado	Oral	51; 21.000
			Escrita	106; 52.000
	L1 chino-L2 español	Principiante	Oral	48; 19.000
			Escrita	100; 51.000
		Intermedio	Oral	49; 20.000
			Escrita	105; 52.000
		Avanzado	Oral	51; 20.000
			Escrita	98; 49.000

7.2.3.2 Variables del aprendiz

En el diseño del corpus han de contemplarse variables del aprendiz que sean relevantes desde un punto de vista de la ASL y que nos permitan hacer una buena valoración lingüística del aprendiz

en cuestión (tabla 7.2). Algunas variables (p.ej., L1, proficiencia) determinarán el diseño general del corpus.

Tabla 7.2 Variables del aprendiz.

Fuente: adaptado de Lozano (2009a, enviado 2020) y Lozano y Mendikoetxea (2013).

Variable	Descripción
Edad	A no ser que se trate de un corpus específico (corpus de niños de L2), recoger el mayor rango de edad posible para así poder testear los efectos de la edad cronológica (p.ej., niños vs adultos).
L1	Recoger la(s) L1(s) del aprendiz, así como la lengua materna de los padres. Un corpus de varias L1 tipológicamente diferentes permite al investigador dilucidar si el comportamiento de los aprendices se debe a influencia de la L1 o a factores universales cognitivos comunes a todos los aprendices independientemente de su L1.
Test de nivel de proficiencia en L2	Es crucial emplear un test de nivel estandarizado para clasificar a los aprendices según nivel de proficiencia y poder hacer estudios de desarrollo fiables. Evitar el uso de medidas clasificatorias *ad hoc* y poco rigurosas tales como el curso académico en que se encuentra el aprendiz (Myles 2015, 316)
Edad de exposición a L2	Las diferentes edades de exposición a la L2 permiten testear efectos de la edad (periodo crítico) en la adquisición de L2.
Entorno	Recoger datos del entorno de aprendizaje: entorno natural (adquisición de L2) vs entorno educativo (adquisición de LE:

	institución de aprendizaje (colegio/instituto/universidad) para testear posibles efectos del entorno).
Duración de exposición	Registrar el número de horas aproximadas/años de exposición a la L2 en el aula y en el entorno natural, pues su efecto ha sido ampliamente demostrado en la literatura.
Duración de residencia	Registrar la duración de las estancias en un país de habla hispana para estudiar posibles efectos adicionales de exposición naturalista, así como las distintas fechas para ver posibles efectos temporales de (proximidad/lejanía de) residencia.
Patrones de uso	Registrar patrones de uso de la L2 para estudiar posibles efectos de dominancia de la L2 (lengua hablada en casa vs en el trabajo, etc.).
Autoevaluación de proficiencia en L2	Emplear medidas adicionales para estimar el nivel de proficiencia como la autoevaluación del propio aprendiz en cada una de las cuatro destrezas (escuchar/hablar/leer/escribir). uele existir una correlación positiva entre las medidas de autoevaluación y de test de proficiencia (Lozano 2016).
Conocimiento de otras lenguas	Registrar el conocimiento de otras lenguas y el nivel alcanzado (certificados o notas oficiales) o bien autoevaluación en las cuatro destrezas, para poder estudiar posibles efectos de otras lenguas no nativas.

7.2.3.3 Variables de la tarea

Finalmente, habremos de registrar ciertas variables sobre la tarea que realizará el aprendiz (tabla 7.3) ya que es bien sabido que la tarea influye en el tipo de lenguaje producido.

Tabla 7.3 Variables de la tarea.

Fuente: adaptado de Lozano (2009a, enviado 2020) y Lozano y Mendikoetxea (2013).

Variable	Descripción
Tipos de tarea y representatividad	Para lograr máxima representatividad, se recomienda administrar distintas tareas que difieran en: grado de dificultad (temas típicos y asequibles para principiantes vs intermedios vs avanzados), naturaleza del lenguaje a producir (amplia gama de posibles estructuras lingüísticas y tiempos verbales, etc.), género (descriptivo, narrativo, argumentativo, epistolar, etc.), restricciones de tarea (tema libre/guiado, descripción de dibujos, etc.).
Uso de herramientas	Recoger el uso de herramientas empleadas para la confección de la tarea (diccionario monolingüe/bilingüe/de sinónimos, gramática, traductor electrónico, ayuda de hablante nativo, etc.) para filtrar o descartar ciertos aprendices del análisis.
Tiempo	Registrar las condiciones temporales de la tarea (tiempo limitado/ilimitado) por si pudieran arrojar información sobre los efectos de la presión temporal.
Entorno	Registrar si la tarea se ha hecho en clase (donde supuestamente existe cierto control por parte del profesor) o fuera de ésta.

7.3 Casos concretos

En esta sección presentaremos algunos corpus de español LE/L2 disponibles gratuitamente en línea y una muestra de estudios empíricos basados en dichos corpus. Aprenderemos a analizar uno de esos corpus con tres herramientas informáticas: Interfaz web de CEDEL2 y software AntConc (análisis de concordancias) y UAM Corpus Tool (etiquetado y estadística).

7.3.1 Estudios empíricos sobre la adquisición del español LE/L2 basados en corpus de aprendices

7.3.1.1 Ejemplos de corpus gratuitos de aprendices de español

En el indexador *Corpus de Aprendices de Español* se podrá encontrar un listado completo de corpus de español LE/L2 (http://repositorios.fdi.ucm.es/corpus_aprendices_español/). Para más información sobre corpus nativos de español y sobre corpus orales, se remite al lector a los capítulos 6 y 8 de este volumen respectivamente. En la última sección de este capítulo (7.6) se podrán consultar los URLs de los corpus que se describen a continuación:

- CEDEL2 (*Corpus Escrito Del Español como L2*) (Lozano 2009a, enviado 2020; Lozano y Mendikoetxea 2013) es un corpus gratuito y disponible *online*. La versión más reciente (CEDEL2 v. 2) contiene algo más de 1.100.00 palabras procedentes de unos 4.300 hablantes que participaron online en diversos países.[3] Consta de 14 tareas (descriptivas, narrativas y argumentativas) y de 11 subcorpus de aprendices de todos los niveles de proficiencia (principiante, intermedio, avanzado) con diversas lenguas maternas (L1: inglés, alemán, neerlandés, italiano, francés, portugués, griego, ruso, chino, japonés, árabe). Consta asimismo de un subcorpus comparable de control de españoles nativos de distintas variedades (peninsular y latinoamerica) de más de 300.000 palabras procedentes de unos 1.100 hablantes. También contiene otros subcorpus nativos de diferentes L1s (inglés, árabe, griego, japonés, portugués).

- CAES (*Corpus de Aprendices del Español*). También es un corpus escrito. La versión 1.2 consta de 574.000 palabras de 1.423 aprendices de diversas L1s (árabe, chino mandarín, francés, inglés, portugués y ruso) en distintos niveles de proficiencia (desde A1 hasta C1).No contiene un corpus de control de nativos de español. Cada aprendiz compuso 2 o 3 redacciones.
- CORESPI (*CORpus del ESPañol de los Italianos*) (Bailini 2016). Es un corpus bidireccional: L1 italiano-L2 español (CORESPI) y L1 español-L2 italiano (CORITE) resultado de los intercambios escritos en e-tándem entre aprendices adultos de niveles A1 a B2. Cada informante de CORESPI produjo entre 5 y 25 textos recogidos a lo largo de 8 meses (total: 457 textos y 124.000 palabras).
- SPLLOC (Spanish Learner Language Oral Corpus) (Mitchell *et al*. 2008). Es un corpus oral de 120 aprendices de L1 inglés-L2 español (principiante, intermedio y avanzado) y 40 nativos de español peninsular como corpus de control. El corpus contiene varias tareas (narrativas, descriptivas y entrevistas).
- LANGSNAP (Languages and Social Networks Abroad Project) (Tracy-Ventura, Mitchell y McManus 2016) es un corpus longitudinal (escrito y oral) de 27 nativos universitarios de inglés británico con español L2 (nivel avanzado) que fueron muestreados antes, durante y después de una estancia Erasmus en España.
- CORELE (Corpus Oral de Español como Lengua Extranjera) (Campillos Llanos 2014). Es un corpus oral de 40 aprendices de español de dos niveles (A2 y B1) y de distintas L1 (sólo 4 aprendices por L1: portugués, italiano, francés, inglés, neerlandés, alemán, polaco, chino y japonés).
- Fono.Ele es un corpus específicamente diseñado para investigar el componente fónico del español como LE/L2. Contiene 96 aprendices, 16 por L1 (alemán, griego, taiwanés,

polaco, portugués, árabe egipcio) y distribuidos en cuatro niveles de proficiencia (A2 hasta C1).

7.3.1.2 Ejemplos de estudios empíricos

En el resto de este subapartado reseñaremos sucintamente algunos estudios empíricos de adquisición de español LE/L2 basados en corpus para ilustrar cómo los corpus pueden arrojar nueva luz sobre la interlengua.

El estudio de los pronombres ha sido investigado con datos de CEDEL2. Por ejemplo, en relación al uso de pronombres personales sujeto plenos vs nulos, los aprendices de español LE/L2 producen redundantemente pronombres plenos como *él* en contextos de mantenimiento de tópico (2a), si bien en contextos de cambio de tópico el pronombre pleno es necesario para desambiguar (2b). Basándose en datos de CEDEL2, Lozano (2009b) concluyó que los pronombres sujeto se usan redundantemente sólo en contextos anafóricos (3ª persona) y no en contextos deícticos (1ª y 2ª persona). Lozano (2016) demostró que, en contextos de cambio de tópico, los nativos y los aprendices no producían pronombres plenos (como se creía en estudios previos), sino más bien SN o nombres propios (e.g., María vio a Juan en el bar mientras Juan bebía cerveza).

(2) a. Juan fue al bar y Ø/#él bebió cerveza

b. María vio a Juan en el bar mientras #Ø/él bebía cerveza

El estudio de tiempo y aspecto (3a, b) ha sido investigado con datos de SPLLOC (Domínguez *et al.* 2013; Tracy-Ventura y Myles 2015; Domínguez, Arche y Myles 2017). Los aprendices cometen inicialmente menos errores con el pretérito perfecto simple (*jugó*) que con el imperfecto (*jugaba*), si bien los datos de corpus revelan, entre otros muchos hallazgos, que (i) los principiantes e intermedios asocian prototípicamente la morfología de pretérito con verbos que denotan aspecto eventivo (*achievements, accomplishments* y *activities*) como (4a), pero la

morfología de imperfecto con el aspecto estativo (4b) y que (ii) dependiendo de la naturaleza de la tarea, los contrastes tempoaspectuales varían, lo cual indica que es imprescindible incorporar varios tipos de tareas en un corpus (véase tabla 7.3).

(3) a. Marta jugó/*jugaba al tenis ayer

b. Marta *jugó/jugaba al tenis de pequeña

(4) a. El trabajador paró al mediodía para almorzar

b. Me gustaba aquel árbol del parque

Otros estudios de corpus han investigado fenómenos como las colocaciones en CEDEL2 (Vincze *et al.* 2016), los clíticos o pronombres átonos en SPLLOC (Arche y Dominguez 2011) y los marcadores discursivos en CORELE (Campillos Llanos y González Gómez 2014). Véase Mendikoetxea (2014) para más estudios de corpus de español L2.

7.3.2 Búsquedas de concordancias en español L2 con CEDEL2 y AntConc

A continuación se mostrará cómo se pueden usar ciertas herramientas informáticas para buscar patrones léxicosintácticos en español L2 que nos permitirán investigar el orden de palabras con verbos intransitivos: Sujeto-Verbo (SV) y Verbo-Sujeto (VS). A diferencia de lenguas de orden rígido como el inglés (SV), en español la alternancia SV/VS es posible, aunque no es totalmente libre ya que está restringida por sutiles factores léxicos y de estructura informativa. Su adquisición en español L2 es compleja (Lozano 2006; Domínguez y Arche 2014). Según la *Hipótesis de la Inacusatividad*, los nativos de español suelen producir VS con verbos inacusativos (de existencia *existir*, aparición *(des)aparecer* y de cambio de lugar *llegar, venir*) (5), pero SV con inergativos (*llorar, gritar, saltar, caminar*, etc) (6). Los aprendices avanzados

de español L2 son sensibles a esta alternancia, aunque no sea ni evidente en el input ni enseñada en clase (Lozano 2014).

(5) a . . . .existe una gran carga de sentimientos encontrados . . . (ES_WR_22_3_RGS.txt)[4]

b . . . .vino la directora del colegio . . . (ES_WR_42_7_EB.txt)

(6) Daniel gritaba en el suelo . . . (ES_WR_42_7_EB.txt)

El corpus CEDEL2 (o cualquier otro corpus similar de español L2 como CAES, SPLLOC, LANGSNAP o CORESPI), combinado con el software de concordancias adecuado, nos permiten contestar la pregunta de investigación en (7).

(7) *Pregunta de investigación:* ¿Son sensibles los anglófonos a la sutil alternancia SV (inergativo)/VS (inacusativo) en español L2?

A continuación, haremos un análisis de concordancias de una muestra de aprendices avanzados de español L2 para contestar a la pregunta de (7). Los datos proceden del corpus CEDEL2 (cf. 7.3.1). No obstante, podremos siempre usar datos de cualquier otro corpus de español L2. Por ello, vamos a hacer el mismo análisis de concordancias usando dos métodos:

- Método 1: Haremos las búsquedas de concordancias en la interfaz web de CEDEL2, ya que está diseñada para hacer búsquedas complejas directamente sobre el corpus CEDEL2.
- Método 2: Nos descargaremos los textos del corpus y usaremos un software gratuito de concordancias denominado *AntCont*. Este software nos permite hacer concordancias sobre los textos de cualquier corpus.

7.3.2.1 Método nº 1 de búsqueda de concordancias: Interfaz de CEDEL2

Usaremos la interfaz de CEDEL2 siguiendo las instrucciones de (8) y de la figura 7.1.

(8) Ir a la interfaz web de CEDEL2 (***http://cedel2.learnercorpora.com***) > Pinchar en la pestaña *Search/Download* > Seleccionar *Advanced search.*

A continuación, nos centraremos en estas tres seccióne (A, B, C):

A. Sección *Results (Output)*: En *Result type* seleccionaremos la opción *Concordances (KWIC)* > En *Result subtype* seleccionaremos *Grammatical elements*.

B. Sección *Filters*: Pinchar en el subcorpus de aprendices deseado (*L1 English-L2 Spanish*) > Seleccionar los datos a analizar según el nivel de proficiencia deseado (*upper advanced*, aunque también podremos seleccionar un rango de proficiencia a medida, por ejemplo: *85%-100%*). Dejamos el resto de los filtros tal cual están.

C. Sección *Grammatical Elements*: Introduciremos el lema a buscar, 'existir', lo cual nos dará todas las formas del verbo existir (p.ej.: *existe, existen, existía, existían, existió, existieron, existirá, existirán,* etc.).

[Insert 15031-4370-PII-007-Figure-001 Here]

Figura 7.1 Interfaz de búsqueda de CEDEL2.

La interfaz de CEDEL generará las concordancias (ver sección inferior de figura 7.1). Podremos buscar concordancias con distintos verbos inacusativos (*aparecer, llegar, venir*, etc.) e inergativos (*gritar, llorar (son)reír, saltar, trabajar*, etc.). Los resultados nos permitirán contestar la pregunta de investigación en (7). Comprobaremos que, efectivamente, los aprendices producen frecuentemente sujetos postverbales con inacusativos (p.ej., *existe un problema aún más grave*), pero nunca con inergativos. Podemos hacer la búsqueda de concordancias también

en el subcorpus de españoles nativos, y observaremos que, como era de esperar, la Hipótesis de la Inacusatividad se cumple en lenguaje nativo.

7.3.2.2 Método nº 2 de búsqueda de concordancias: Software AntConc

AntConc (www.laurenceanthony.net/software.html) es un software gratuito. Permite hacer búsquedas de concordancias, listas de palabras, colocaciones, *clusters* o agrupaciones de palabras, n-gramas, palabras clave, etc. Primero descargaremos AntConc v. 3.5.0 o posterior. Al ser un software autoejecutable, no requiere instalación. En la interfaz principal (figura 7.2) observamos varias áreas: menús (*File, Global Settings*, etc.), *Corpus Files* (donde se mostrarán los ficheros a analizar), la ventana de concordancias (en la que se mostrarán el nº de concordancias encontradas o *hits* con la palabra en contexto o KWIC: *KeyWord In Context*), el fichero (*File*) de donde procede cada concordancia y, abajo, la ventana del término de búsqueda (*Search Term*).

[Insert 15031-4370-PII-007-Figure-002 Here]

Figura 7.2 Interfaz principal de AntConc.

A continuación, descargaremos la muestra de textos de aprendices a analizar. En este caso, analizaremos datos del corpus CEDEL2 (cf. 7.3.1) siguiendo las instrucciones de (9), si bien podríamos descargar y analizar textos de cualesquiera otros corpus de aprendices de español L2. En nuestro caso, recordemos que analizaremos una muestra de aprendices avanzados de español L2 del corpus CEDEL2

(9) Ir a la web de CEDEL2 (*http://cedel2.learnercorpora.com*) > Pinchar en la pestaña *Search/Download* y seguidamente en *Advanced search* > Pinchar en el subcorpus de aprendices deseado (*L1 English-L2 Spanish*) > Filtrar los datos según el nivel de proficiencia deseado (*upper advanced*, aunque también podremos seleccionar un rango

de proficiencia a medida, por ejemplo: *85%-100%*) > Pinchar en el botón *Search* > Nos aparecerá un listado de los diferentes aprendices filtrados > Pincharemos en el botón *Download* para descargar los textos (tenemos la opción de descargar los textos solamente (*texts only*) o los textos con los metadados (*texts with metadata*). Por simplicidad para este ejercicio, seleccionaremos *texts only* > Pincharemos en el botón *Download* para así descargar los 232 textos seleccionados.

Dado que los textos descargados están en una carpeta comprimida, se descomprimirá la carpeta para poder trabajar con los textos en AntCont. A continuación, subiremos los textos a AntConc según las instrucciones en (10).

(10) En AntConc > *File* > *Open Dir[ectory]* > Seleccionar la carpeta descomprimida con los txt > *Aceptar/OK* > Los 232 ficheros cargados aparecerán en el área *corpus files* de Antconc.

Siguiendo (11), introduciremos el término de búsqueda con el uso de comodines como el asterisco (*exist**) para encontrar todas las raíces que comiencen por *exist-* (*existe, existirá, existió,*. . .) pero también concordancias no deseadas *(existencia*). Dado que los textos están en formato TXT y no están etiquetados (cosa que sí ocurre en la interfaz web de CEDEL2), el uso de comodines se hace imprescindible en AntConc para buscar todas las formas posibles de un lema. Nótese que AntConc acepta los siguientes comodines: * (cero o más caracteres), + (cero o un carácter), ? (cualquier carácter), @ (cero o una palabra), # (cualesquiera palabras), | (operador OR).

(11) Introducir en *Search Term* el término con comodín *exist** o bien términos específicos seguidos del operador |, lo que nos permitirá buscar cada una de las palabras introducidas, p.ej.: *existe*|*existen*|*existía*|*existían*|*existió*|*existieron*|*existirá*|*existirán* > Pinchar en *Start* para iniciar la búsqueda de concordancias > Si deseamos una mejor visualización de los

resultados, pincharemos en el botón *Sort* para organizar las concordancias por primera palabra a la derecha del término de búsqueda.

AntConc generará las concordancias y observaremos algo similar a la figura 7.3. Las analizaremos visualmente y descartaremos las no deseadas (p.ej., concordancias que pudieran contener sustantivos como *existencia*). Buscaremos concordancias con otros verbos inacusativos (*aparecer, llegar, venir*, etc.) e inergativos (*gritar, llorar (son)reir, saltar, trabajar*, etc.) para así contestar la pregunta de investigación en (7). Comprobaremos que, efectivamente, los aprendices producen frecuentemente sujetos postverbales con inacusativos pero nunca con inergativos. Podemos repetir todo el proceso anteriormente descrito y analizar otros subcorpus de aprendices (p.ej., intermedios) o incluso el subcorpus de españoles nativos. Comprobaremos que la Hipótesis de la Inacusatividad se cumple tanto en lenguaje nativo como en lenguaje no nativo.

[Insert 15031-4370-PII-007-Figure-003 Here]

Figura 7.3 Concordancias en AntConc.

7.3.2.3 Conclusión: Concordancias y adquisición de segundas lenguas

Como hemos podido observar, las herramientas de concordancias (ya sean las que vienen incorporadas directamente en las interfaces web como el caso de CEDEL2 o bien las herramientas gratuitas como AntCont) nos permiten hacer búsquedas sofisticadas en los corpus para poder responder preguntas de investigación sobre la adquisición del español L2.

Es importante resaltar que un corpus bidireccional (cf. Sección 7.2.2) nos permite verificar si el fenómeno que hemos observado en una dirección (CEDEL2: L1 inglés-L2 español) es observable en la otra dirección (COREFL: L1 español-L2 inglés). El corpus COREFL (http://corefl.learnercorpora.com) (Lozano, Díaz-Negrillo y Callies 2020) fue diseñado precisamente para tal fin. Si analizamos COREFL (en su propia interfaz o con AntConc),

observaremos también que los aprendices producen sujetos postverbales sólo con inacusativos (cf. resultados en Lozano y Mendikoetxea 2010; Mendikoetxea y Lozano 2018).

Finalmente, la interfaz web de CEDEL2 y la herramienta AntConc también permiten explorar patrones de fenómenos clásicos en adquisición tales como la distinción *ser/estar* (12), la sintaxis de *gustar* (13), colocaciones con *hacer* en torno a una base (p.ej. *calor*) (14), etc.

(12) a. *. . . estoy muy pobre . . . (EN_WR_21_43_3_1_AJ.txt)

b. *Soy muy contenta . . . (EN_WR_18_33_2_1_BLW.txt)

(13) a. *Yo me gusta chocolate mucho. (EN_WR_28_20_4_4_JAG.txt)

b. *Halle Barry gusta comer postres. (EN_WR_19_39_1_2_AH.txt)

(14) *. . . en Granada es mucho calor en Verano (EN_WR_13_34_1_1_JW.txt)

7.3.3 Ejemplo de etiquetado manual de CEDEL2 con UAM Corpus Tool

A continuación mostraremos cómo se puede etiquetar manualmente una muestra de CEDEL2 en UAM Corpus Tool para poder estudiar fenómenos complejos que no podrían ser estudiados con software de concordancias.

Estudios de corpus de español L2 (Lozano 2009b, 2016) demuestran que, en el mantenimiento de tópico (es decir, cuando continuamos hablando del mismo sujeto que el de la oración anterior), los aprendices de español L2 producen redundantemente pronombres personales en posición sujeto, p.ej., *ella* en (15).

(15) Hablé para Tyra Banks. Ella a modela de super. Ella es muy importante. Ama a chicas. (EN_WR_14_24_2_2_JR.txt)

Usaremos CEDEL2 y UAM Corpus Tool para poder responder a la pregunta de investigación planteada en (16).

(16) *Pregunta de investigación:* ¿Producen los aprendices principiantes (L1 inglés-L2 español) más pronombres personales en posición sujeto para mantener el tópico que los españoles nativos?

UAM Corpus Tool (O'Donnell 2009; www.corpustool.com/) es un software gratuito que permite etiquetar un corpus según un esquema de etiquetas definido por el usuario y hacer contrastes estadísticos entre dichas etiquetas. Descargaremos la versión 3.3 o posterior (Windows/MacOS). Una vez instalado, crearemos un nuevo proyecto (17), tras lo cual veremos la pantalla principal (Figura 7.4).

(17) **Crear un nuevo proyecto:** *Start New Project* > *Next* > *Name of the project*: Mantenimiento-tópico > *Folder* [indicar en dónde se guardará el proyecto] > *Finalize*.

[Insert 15031-4370-PII-007-Figure-004 Here]

Figura 7.4 Interfaz principal en UAM Corpus Tool.

A continuación añadiremos al proyecto los textos a analizar. Para esta demostración, seleccionaremos textos de un tema que dé pie a contextos de mantenimiento de tópico (nº 2: Habla de una persona famosa). Seleccionaremos sólo dos textos de aprendices y un texto de nativos, según (18). Finalmente, colocaremos los tres textos en una carpeta a la que denominaremos *textos*.

(18) En la web de CEDEL2 (http://cedel2.learnercorpora.com) > Pinchar en la pestaña *Search/Download* y seguidamente en *Advanced search.*

Textos de aprendices: En la caja *Subcorpus*, seleccionar *Learners of L2 Spanish* > En la caja de L1, seleccionar el subcorpus L1 English-L2 Spanish > En la caja de *Filename*, introducir los nombres de los dos ficheros de aprendices a descargar (EN_WR_12_19_3_2_DT.txt, EN_WR_9_16_2_2_RM.txt) > Pinchar en el botón *Search* > Pinchar en el botón *Download* > Descargar los textos con metadatos (para así tener información precisa de cada aprendiz).

Texto de nativo: En la caja *Suborpus*, seleccionar *Natives* Z En la caja de L1, seleccionar el suborpus *Spanish* > En la caja de *Filename*, introducir el nombre del fichero nativo (ES_WR_26_2_AVS.txt) > Pinchar en el botón *Search* > Pinchar en el botón *Download* > Descargar los textos con metadatos.

Descromprimiremos la carpeta que contiene los textos y seguiremos (19) para añadir las carpetas con los tres textos a UAM Corpus Tool y los incorporaremos al proyecto Mantenimiento-tópico ya creado.

(19) **Añadir ficheros al proyecto:** En UAM Corpus Tool > Pinchar en botón *Extend Corpus* > En la ventana *Corpus Location* seleccionar la opción *I want to add a folder of text files* > Pinchar sobre el botón con tres puntos > Seleccionar las carpetas denominada *texts* > Pinchar en botón *Finalize* > Incorporar los textos al proyecto pinchando en el botón *Incorporate All.*

Tras seleccionar e incorporar los textos, crearemos el esquema de etiquetas (*tagset*). Definiremos dos capas (*layers*). Una servirá para etiquetar el grupo al que pertenece cada texto (aprendices/nativos) y la otra contendrá el esquema de anotación del fenómeno lingüístico en cuestión. Comencemos creando la capa ‘Grupo’ según (20).

(20) **Crear una capa (anotación de documentos completos):** Botón *Layers* > *Add Layers* > *Start* > Denominar la capa como 'Grupo' > *Manual Annotation* para anotar manualmente > *Design Your Own* para crear nuestras propias etiquetas > *Whole Document* para anotar cada documento como perteneciente al grupo de aprendices o de nativos > *Create Layer.*

Ahora definiremos las etiquetas de la capa 'Grupo'. Por defecto, el software nos crea dos etiquetas: 'grupo-1' y 'grupo-2', que habremos de renombrar según (21) para obtener el esquema de etiquetas deseado (figura 7.5).

(21) **Renombrar etiqueta:** Pinchar en *Edit Scheme* de la ventana *Layers* > Para renombrar la etiqueta 'Grupo-1', pinchar sobre ella con el botón derecho del ratón > Opción *Rename feature* > Renombrar la etiqueta como 'Aprendices' > *OK* > Seguir el mismo proceso y renombrar 'Grupo-2' como 'Nativos' > Cerrar la capa 'Grupo' pinchando en el botón superior derecho *Close*

[Insert 15031-4370-PII-007-Figure-005 Here]

Figura 7.5 Esquema de la capa 'Grupo'.

Seguidamente crearemos la capa 'Sujeto-mant-top' en la que definiremos las etiquetas para anotar segmentos. Un segmento puede ser una letra, una palabra, un sintagma, una oración o un párrafo. En nuestro caso, cada segmento corresponde a sujeto oracional de 3ª persona singular. Seguiremos los pasos de (22).

(22) **Crear una capa (anotación de segmentos):** Botón *Layers* > *Add Layers* > *Start* > Denominar la capa como 'Sujeto-mant-top' > *Manual Annotation* para anotar manualmente > *Design Your Own* para crear nuestras propias etiquetas > *Segments within a Document* para anotar segmentos > En las opciones de segmento especial, pinchar en *No* (caso típico) o en *Error* (si se desea especificar la forma correcta cuando el aprendiz

produzca un error) > En las opciones de segmentado automático, pinchar en *No* > *Create Layer.*

Ahora definiremos las etiquetas de la capa 'Sujeto-mant-top'. Por defecto UAM CT nos proporciona dos opciones: sujeto-mant-top1/sujeto-mant-top2 (figura 7.6A). Necesitamos etiquetar dos aspectos (figura 7.6B): la sintaxis del sujeto oracional (pronombre nulo/pronombre pleno/SN) y su adecuación pragmática al contexto de mantenimiento de tópico (adecuado/redundante). Es importante señalar que en los esquemas de dos o más aspectos a etiquetar necesitamos usar gráficamente una llave (operador lógico AND), para etiquetar primero la sintaxis y luego la pragmática. Y dentro de cada uno de estos dos nodos, etiquetaremos sólo una opción usando gráficamente los corchetes (operador lógico OR): *pronombre-nulo* o *pronombre-pleno* o *SN*. Igualmente, en pragmática asignaremos la etiqueta *adecuado* o *redundante*. Para crear un operador lógico AND, seguir los pasos en (23); para renombrar un sistema (p.ej., SUJETO-MANT-TOP-TYPE → SINTAXIS), seguir (24); para renombrar una etiqueta, ver (21) más arriba; y para añadir una nueva etiqueta, seguir (25). El resultado final será el esquema en figura 7.6B.

(23) **Crear un sistema de operador lógico AND (llave):** Pinchar con el botón derecho sobre el nodo raíz 'Sujeto-mant-top'> seleccionar *Add System*.

(24) **Renombrar sistema:** Pinchar con el botón derecho del ratón sobre el nombre del sistema > *Rename System* > Proporcionar el nuevo nombre del sistema > *OK*.

(25) **Añadir nueva etiqueta:** Pinchar con el botón derecho sobre el nombre del sistema > Opción *Add Feature* > Proporcionar el nuevo nombre de la etiqueta > *OK*.

[Insert 15031-4370-PII-007-Figure-006 Here]

Figura 7.6 Esquema de la capa 'Sujeto-mant-top'.

Fuente: simplificación del esquema en LOZANO (2019).

Una vez definidos los esquemas de etiquetas (*tagsets*), etiquetaremos los textos incorporados según las instrucciones de (26).

(26) **Etiquetar:** Botón principal *Files* > Pinchar sobre el nombre de la capa a etiquetar (Grupo | Sujeto-mant-top), que aparecerá en color anaranjado (lo cual indica que el texto aún no ha sido etiquetado completamente; una vez etiquetado el texto completamente, el color será sepia).

Etiquetar texto completo (*Grupo*): Hacer doble click sobre la etiqueta correspondiente (*aprendices*|*nativos*) para asignar la etiqueta al texto > Cerrar la ventana y guardar cambios.

Etiquetar segmentos (*Sujeto-mant-top*): Pinchar en el inicio del segmento a etiquetar y arrastrar hasta el final del segmento > El segmento seleccionado quedará subrayado en verde > Asignar las etiquetas correspondientes de sintaxis y de pragmática haciendo doble click sobre cada etiqueta (ver resultado final en figura 7.6) > Cerrar la ventana y guardar cambios.

Nótese que, en caso de necesitar ampliar/reducir el subrayado verde (cf. figura 7.7), debe posicionarse en alguno de los extremos y arrastrar el segmento. Por otro lado, en caso de asignar una etiqueta por error, se desasignará haciendo doble click sobre la etiqueta ya asignada. Para asignar las mismas etiquetas a una palabra (p.ej., *Ella*), pinche en el menú *Options* > Seleccionar *Automatically code repeated segments*.

[Insert 15031-4370-PII-007-Figure-007 Here]

Figura 7.7 Etiquetado de segmentos con UAM Corpus Tool.

Por ejemplo, en el texto EN_WR_12_19_3_2_DT.txt (figura 7.7), *Ella* en *Ella vive en California* es redundante (etiquetas: ‘pronombre-pleno’ y ‘redundante’), dado que se espera un

pronombre nulo para marcar el mantenimiento de tópico, como ocurriría en *Ama a chicas* del texto EN_WR_9_16_2_2_RM.txt, en el que se está hablando de Tyra Banks (etiquetas: 'pronombre-pleno' y 'adecuado'). En caso de pronombres nulos, podemos etiquetar el verbo (p.ej., *Ama*) dado que no podemos etiquetar físicamente un elemento nulo.

Tras etiquetar los textos, procederemos a la estadística (27) para contestar la pregunta de investigación (16). El lector puede explorar las distintas opciones de estadística buscando en internet el manual de UAM Corpus Tool y tutoriales.

(27) **Estadística (comparar dos grupos):** Botón *Statistics* > En *Type of Study* seleccionar *Compare several datasets* para comparar varios grupos o subcorpus (aprendices vs. nativos); en *Aspect of Interest* seleccionar *Feature Coding* para hacer estadística basada en etiquetas > en *Counting* seleccionar *Local* > en *Unit* seleccionar la unidad a analizar: la raíz del esquema de etiquetas, *sujeto-mant-top*, aunque podemos elegir subetiquetas de dicho esquema; en *Set 1* elegiremos el primer subcorpus a comparar ('aprendices') y en *Set 2* el segundo subcorpus ('nativos') > Pinchar en *Show* para mostrar la estadística (figura 7.7).

Como podemos observar (figura 7.8), aunque nuestros dos subcorpus son pequeños (sólo 2 textos para los aprendices y 1 texto para los nativos), la estadística (χ^2: chi cuadrado) nos indica que los aprendices producen significativamente menos pronombres nulos (24.14%) que los españoles nativos (81.25%) para marcar mantenimiento de tópico; asimismo los aprendices sobreproducen pronombres plenos (72.41%) mientras que los nativos no los producen. Estos datos dan respuesta a nuestra pregunta de investigación.

[Insert 15031-4370-PII-007-Figure-008 Here]

Figura 7.8: Análisis estadístico con UAM Corpus Tool.

Finalmente, también observamos (figura 7.8) que los aprendices son significativamente más redundantes (75.86%) que los nativos (18.75%), aunque no sabemos qué formas son

redundantes. UAM Corpus Tool permite análisis más finos partiendo desde cualquier etiqueta como unidad de análisis (*Unit*), p.ej., 'Redundante'. Observamos ahora (figura 7.9) que las formas redundantes son principalmente pronombres plenos en los aprendices pero SN en los nativos (sobreuso del nombre propio *Rafa Nadal* para mantener el tópico). Este tipo de análisis estadístico exploratorio es ideal para generar nuevas preguntas de investigación e hipótesis que hubieran pasado desapercibidas sin el uso de este software.

[Insert 15031-4370-PII-007-Figure-009 Here]

Figura 7.9 Análisis estadístico con UAM Corpus Tool.

Nótese que en la opción *Counting* podemos seleccionar entre *Local* y *Global*. Para entender la diferencia entre ambos, imaginemos un sencillo corpus de 10 formas de sujeto, 5 de las cuales son nombres (N), de los cuales 2 son nombres comunes y 3 son nombres propios. En la tabla 7.4 podemos apreciar las diferencias estadísticas. Si analizamos los N comunes, un conteo global nos proporciona la frecuencia de etiquetas del total de etiquetas del corpus (2 N comunes de un total de 10 formas de sujeto = 20%), mientras que el conteo local nos proporciona la frecuencia dentro de una categoría (2 N comunes dentro de 5 casos de la categoría N = 40%).

Tabla 7.4 Diferencias entre conteo local vs. global en UAM Corpus Tool.

	Global	Local
N común	2/10 = 20%	2/5 = 40%
N propio	3/10 = 30%	3/5 = 60%

7.4 Conclusión

Los corpus de español LE/L2 si están bien diseñados y si se combinan con herramientas informáticas de análisis de corpus, permiten investigar sistemática y rigurosamente la interlengua de los aprendices. Para que este enfoque sea lo más fructífero posible en futuras investigaciones del español como L2, es recomendable invertir recursos en la formación del uso de herramientas

y software informático del tipo descrito en este capítulo. Para ello, el apoyo institucional es esencial pues permite crear las condiciones necesarias para formar a futuros investigadores.

Los corpus ofrecen la ventaja de estudiar la producción natural de ciertos fenómenos lingüísticos con amplias muestras, aunque cuentan con limitaciones que a veces es necesario contrarrestar con el uso adicional de métodos de investigación más restrictivos como los experimentos. Recientemente, algunos investigadores abogan por la triangulación, es decir, el uso combinado de datos de corpus y experimentales para lograr una mejor comprensión de un mismo fenómeno de la L2 (Mendikoetxea y Lozano 2018).

Notas

7.5 Bibliografía

Arche, M. J. y L. Dominguez. 2011. “Morphology and Syntax Dissociation in SLA: A Study on Clitic Acquisition in Spanish”. En *Morphology and Its Interfaces*, ed. A. Galani, G. Hicks y G. Tsoulas, 291–319. Amsterdam: John Benjamins.

Bailini, S. 2016. *La interlengua de las lenguas afines: El español de los italianos, el italiano de los españoles.* Milano: Edizioni Universitarie di Lettere, Economia, Diritto.

Callies, M. 2015. “Learner Corpus Methodology”. En *The Cambridge Handbook of Learner Corpus Research*, ed. S. Granger, G. Gilquin y F. Meunier, 35–55. Cambridge: Cambridge University Press.

Campillos Llanos, L. 2014. “A Spanish Oral Learner Corpus for Computer-Aided Error Analysis”. *Corpora* 9 (2): 207–238. http://doi.org/10.3366/cor.2014.0058

Campillos Llanos, L. y P. González Gómez. 2014. “Oral Production of Discourse Markers by Intermediate Learners of Spanish: A Corpus Perspective”. En *Yearbook of Corpus Linguistics and Pragmatics 2014*, ed. J. Romero-Trillo, 239–259. New York: Springer.

Cruz Piñol, M. 2012. *Lingüística de corpus y enseñanza del español como 2/L*. Madrid: Arco Libros.

Díaz-Negrillo, A. y P. Thompson. 2013. “Learner Corpora: Looking Towards the Future”. En *Automatic Treatment and Analysis of Learner Corpus Data*, ed. A. Díaz-Negrillo, N. Ballier y P. Thompson, 9–29. Amsterdam: John Benjamins.

Domínguez, L. y M. J. Arche. 2014. “Subject Inversion in Non-Native Spanish”. *Lingua* 145: 243–265. https://doi.org/10.1016/j.lingua.2014.04.004

Domínguez, L., M. J. Arche y F. Myles. 2017. “Spanish Imperfect Revisited: Exploring L1 Influence in the Reassembly of Imperfective Features onto New L2 Forms”. *Second Language Research* 33 (4): 431–457. https://doi.org/10.1177/0267658317701991

Domínguez, L., N. Tracy-Ventura, M. J. Arche, R. Mitchell y F. Myles. 2013. “The Role of Dynamic Contrasts in the L2 Acquisition of Spanish Past Tense Morphology”. *Bilingualism: Language and Cognition* 16 (3): 558–577. https://doi.org/10.1017/S1366728912000363

Geeslin, K. L., ed. 2014. *The Handbook of Spanish Second Language Acquisition*. Oxford: Wiley-Blackwell.

Gilquin, G. 2015. “From Design to Collection of Learner Corpora”. En *The Cambridge Handbook of Learner Corpus Research*, ed. S. Granger, G. Gilquin y F. Meunier, 9–34. Cambridge: Cambridge University Press.

Granger, S. 2008. “Learner Corpora”. En *Corpus Linguistics: An International Handbook*, ed. A. Lüdeling y M. Kytoe, 259–275. Berlin: Mouton de Gruyter.

Granger, S. 2009. “The Contribution of Learner Corpora to Second Language Acquisition and Foreign Language Teaching”. En *Corpora and Language Teaching*, ed. K. Aijmer, 13–32. Amsterdam: John Benjamins.

Granger, S. 2012. “How to Use Foreign and Second Language Learner Corpora”. En *Research Methods in Second Language Acquisition: A Practical Guide*, ed. A. Mackey y S. M. Gass, 5–29. Oxford: Wiley-Blackwell.

Granger, S. 2015. “Contrastive Interlanguage Analysis: A Reappraisal”. *International Journal of Learner Corpus Research* 1 (1): 7–24. https://doi.org/10.1075/ijlcr.1.1.01gra

Granger, S., G. Gilquin y F. Meunier, eds. 2015. *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

Hincapié Moreno, D. A. y J. A. Bernal Chávez. 2018. *Lingüística de corpus*. Bogotá: Instituto Caro y Cuervo.

Lozano, C. 2006. "Focus and Split Intransitivity: The Acquisition of Word Order Alternations in Non-Native Spanish". *Second Language Research* 22 (2): 1–43. https://doi.org/10.1191/0267658306sr264oa

Lozano, C. 2009a. "CEDEL2: Corpus Escrito del Español como L2". En *Applied Linguistics Now: Understanding Language and Mind/La Lingüística Aplicada Actual: Comprendiendo el Lenguaje y la Mente*, ed. C. M. Bretones *et al.*, 197–212. Almería: Universidad de Almería.

Lozano, C. 2009b. "Selective Deficits at the Syntax-Discourse Interface: Evidence from the CEDEL2 Corpus". En *Representational Deficits in Second Language Acquisition*, ed. Y. I. Leung, N. Snape y M. Sharwood-Smith, 127–166. Amsterdam: John Benjamins. https://doi.org/10.1075/lald.47.09loz

Lozano, C. 2014. "Word Order in Second Language Spanish". En *The Handbook of Spanish Second Language Acquisition*, ed. K. L. Geeslin, 287–310. Oxford: Wiley-Blackwell.

Lozano, C. 2015. "Learner Corpora as a Research Tool for the Investigation of Lexical Competence in L2 Spanish". *Journal of Spanish Language Teaching* 2 (2): 180–193. https://doi.org/10.1080/23247797.2015.1104035

Lozano, C. 2016. "Pragmatic Principles in Anaphora Resolution at the Syntax-Discourse Interface: Advanced English Learners of Spanish in the CEDEL2 Corpus". En *Spanish Learner Corpus Research: Current Trends and Future Perspectives*, ed. M. Alonso Ramos, 235–265. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.78.09loz

Lozano, C. 2020, enviado. "CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research".

Lozano, C., A. Díaz-Negrillo y M. Callie. 2020. "Designing and Compiling a Learner Corpus of Written and Spoken Narratives: COREFL". En *What's in a Narrative? Variation in Story-Telling at the Interface Between Language and Literacy*, ed. C. Bongartz y J. Torregrossa, 9–32. Frankfurt am Main: Peter Lang.

Lozano, C. y A. Mendikoetxea. 2010. "Interface Conditions on Postverbal Subjects: A Corpus Study of L2 English". *Bilingualism: Language and Cognition* 13 (4): 475–497. https://doi.org/10.1017/S1366728909990538

Lozano, C. y A. Mendikoetxea. 2013. "Learner Corpora and Second Language Acquisition: The Design and Collection of CEDEL2". En *Automatic Treatment and Analysis of Learner Corpus Data*, ed. A. Díaz-Negrillo, N. Ballier y P. Thompson, 65–100. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.59.06loz

Mendikoetxea, A. 2014. "Corpus-Based Research in Second Language Spanish". En *The Handbook of Spanish Second Language Acquisition*, ed. K. L. Geeslin, 11–29. Oxford: Wiley-Blackwell.

Mendikoetxea, A. y C. Lozano. 2018. "From Corpora to Experiments: Methodological Triangulation in the Study of Word Order at the Interfaces in Adult Late Bilinguals (L2 Learners)". *Journal of Psycholinguistic Research* 47 (4): 871–898. https://doi.org/10.1007/s10936-018-9560-0

Mitchell, R., L. Domínguez, M. Arche, F. Myles y E. Marsden. 2008. "SPLLOC: A New Database for Spanish Second Language Acquisition Research". En *EUROSLA Yearbook 8*, ed. L. Roberts, F. Myles y A. David, 287–304. Amsterdam: John Benjamins.

Montrul, S. 2004. *The Acquisition of Spanish: Morphosyntactic Development in Monolingual and Bilingual L1 Acquisition and Adult L2 Acquisition*. Amsterdam: John Benjamins.

Myles, F. 2005. "Interlanguage Corpora and Second Language Acquisition Research". *Second Language Research* 21 (4): 373–391.

Myles, F. 2007. "Using Electronic Corpora in SLA Research". En *Handbook of French Applied Linguistics*, ed. D. Ayoun, 377–400. Amsterdam: John Benjamins.

Myles, F. 2015. “Second Language Acquisition Theory and Learner Corpus Research”. En *The Cambridge Handbook of Learner Corpus Research*, ed. S. Granger, G. Gilquin y F. Meunier, 309–332. Cambridge: Cambridge University Press.

O’Donnell, M. 2009. “The UAM CorpusTool: Software for Corpus Annotation and Exploration”. En *La Lingüística Aplicada actual: Comprendiendo el Lenguaje y la Mente*, ed. C. M. Bretones *et al.*, 1433–1447. Almería: Universidad de Almería.

Sinclair, J. 2005. “How to Build a Corpus”. En *Developing Linguistic Corpora: A Guide to Good Practice*, ed. M. Wynne, 79–83. Oxford: Oxbow Books.

Tono, Y. 2016. “What Is Missing in Learner Corpus Design?”. En *Spanish Learner Corpus Research: Current Trends and Future Perspectives*, ed. M. Alonso Ramos, 33–52. Amsterdam: John Benjamins.

Towell, R. y R. Hawkins. 1994. *Approaches to Second Language Acquisition*. Clevedon: Multilingual Matters.

Tracy-Ventura, N., R. Mitchell y K. McManus. 2016. “The LANGSNAP Longitudinal Learner Corpus: Design and Use”. En *Spanish Learner Corpus Research: State of the Art and Perspectives*, ed. M. Alonso Ramos. Amsterdam: John Benjamins.

Tracy-Ventura, N. y F. Myles. 2015. “The Importance of Task Variability in the Design of Learner Corpora for SLA Research”. *International Journal of Learner Corpus Research* 1 (1): 58–95. https://doi.org/10.1075/ijlcr.1.1.03tra

Vincze, O., M. Gracía-Salido, A. Orol y M. Alonso-Ramos. 2016. “A Corpus Study of Spanish as a Foreign Language Learners’ Collocation Production”. En *Spanish Learner Corpus Research: Current Trends and Future Perspectives*, ed. M. Alonso-Ramos, 299–331. Amsterdam: John Benjamins.

7.6 Recursos en línea

Corpus de aprendices de L2:

- Corpus CEDEL2 (Corpus Escrito Del Español como L2): http://cedel2.learnercorpora.com
- Corpus CAES (Corpus de Aprendices de ESpañol): https://galvan.usc.es/caes/
- Corpus CORESPI (CORpus de ESPañol de los Italianos): http://corespiycorite.altervista.org/
- Corpus SPLLOC (Spanish Learner Language Oral Corpus): http://www.splloc.soton.ac.uk/
- Corpus LANGSNAP: http://langsnap.soton.ac.uk/
- Corpus CORELE (Corpus Oral de Español como Lengua Extranjera): http://www.lllf.uam.es/ESP/CORELE.html
- Corpus Fono.Ele: http://www3.uah.es/fonoele/
- Corpus COREFL (CORpus of English as a Foreign Language): http://corefl.learnercorpora.com
- Corpus WriCLE (Written Corpus of Learner English): http://wricle.learnercorpora.com
- *Indexador de Corpus de Aprendices de Español*: http://repositorios.fdi.ucm.es/corpus_aprendices_español/

Software de corpus:

- AntConc (software de concordancias): http://www.laurenceanthony.net/software.html
- UAM Corpus Tool (etiquetador): http://www.corpustool.com/
-

[1] Usaremos el término estándar "adquisición de L2" para referirnos tanto a la adquisición del español en entornos naturales (L2) como al español como lengua extranjera en el aula (LE). La literatura científica ha demostrado que muchos de los procesos (psico)lingüísticos de adquisición de L2 son similares independientemente del entorno de adquisición (*inter alia*, Towell y Hawkins 1994).

[2] El asterisco "*" indica agramaticalidad mientras que "#" indica anomalía pragmática (pero no agramaticalidad).

[3] A través de listas de distribución como Infoling y LinguistList se hicieron varias llamadas a la participación en línea en CEDEL2, lo que permitió que la muestra de informantes nativos y no nativos fuese variada y diversa en términos de países de origen, de instituciones muestreadas y de entornos de aprendizaje y niveles de competencia (cf. las ventajas de estos aspectos de CEDEL2 en Gilquin 2015). Agradecemos a estas listas su cooperación.

[4] Al final de cada concordancia se indica, entre paréntesis, el nombre del fichero de CEDEL2 de donde procede.