# CLEAE: Corpus Longitudinal de Español de Aprendientes Estonios. Diseño y estudio piloto

Virginia Rapún Mombiela
*Universidad de Tartu*

## 1 Introducción

El proyecto que vamos a presentar en este artículo surge de la necesidad de entender el proceso de adquisición de la lengua española por parte de estudiantes estonios, un campo de investigación que tiene hasta ahora un corto recorrido. Los alumnos estonios, además de su lengua nativa, tienen ya amplios conocimientos de una segunda lengua extranjera cuando empiezan a estudiar español (casi siempre el inglés), a veces, incluso, más de una (ruso, sueco, finés) y son, por lo tanto, plurilingües. Pese a que, en un primer momento, podría pensarse que la condición plurilingüe de muchos estudiantes de español en Estonia tiene que afectar positivamente el aprendizaje de español, es importante comprender que la adquisición de terceras lenguas es un fenómeno complejo y multidisciplinario que se ve afectado por un gran número de factores individuales y contextuales (Cenoz 2008, 220).

Antes de poder hacer cualquier suposición que explique el porqué de algunos fenómenos recurrentes en la producción de nuestros estudiantes, es indispensable conocer el proceso por el que pasa su interlengua. La lingüística de corpus de aprendientes puede ayudarnos a conocer en profundidad cómo ocurre la adquisición de la lengua española por parte de estudiantes estonios y a describir patrones recurrentes que nos permitan hacer hipótesis y comprobar teorías. Hasta ahora, en Estonia, el empleo de corpus de aprendientes de español con un número significativo de muestras no ha sido una práctica habitual. Kruse (2018) fue pionera en utilizar esta metodología, analizando 73 trabajos de grado defendidos en la Sección de Español perteneciente al Departamento de Filología Románica de la Universidad de Tartu. A través de este corpus, Kruse estudió los sistemas complejos que se generan en el hablante plurilingüe estonio para demostrar de manera empírica si, efectivamente, tal y como ella había observado como profesora de español, los alumnos estonios producían una gran cantidad de falsos amigos léxicos provenientes del inglés (Kruse 2018, 121). El corpus permitió, además, comprobar que los falsos amigos no se limitaban a la categoría léxica y aparecían también en las esferas de la sintaxis y la ortografía. Este trabajo abrió el camino a la lingüística de corpus de aprendientes en las investigaciones relacionadas con la adquisición del español por parte de estudiantes de lenguas finoúgrias, un campo de estudio en el que todavía queda mucho por investigar.

Pese al innegable valor del mencionado corpus, su carácter trasversal impide que pueda ser utilizado para analizar el desarrollo de la interlengua y comprender de qué manera los alumnos de español estonios avanzan en los niveles de competencia. Existe la necesidad, pues, de crear un corpus de aprendientes de español nuevo que contenga muestras producidas por los estudiantes a lo largo del tiempo. Con un enfoque longitudinal, se podría fácilmente comprender en qué áreas los alumnos encuentran más dificultades, si existe o no influencia interlingüística, qué errores se fosilizan y si se produce una secuencia de adquisición determinada y generalizable.

Este artículo versa sobre la compilación del nuevo *Corpus Longitudinal de Español de Aprendientes Estonios* (CLEAE) y los pasos previos a su creación. En primer lugar, se realizará una breve revisión de los corpus de aprendientes de español que están

accesibles en la red y que más estudios han generado. A continuación, se hablará de las ventajas de trabajar con un corpus longitudinal. Después, se introducirá el CLEAE y se describirá el corpus piloto que lo hace posible, poniendo ejemplos de su funcionamiento a través de los datos de una de los informantes. Se planteará, también, una breve discusión acerca de los datos recogidos en tiempos de pandemia y, por último, se propondrán las conclusiones.

## 2 Estudios basados en corpus de aprendientes

La investigación basada en corpus de aprendientes comenzó a finales de la década de 1980 "como una rama de la lingüística de corpus" (Granger *et al*. 2015, 1)[1]. Desde entonces, se ha publicado una cantidad considerable de estudios sobre este tema. Paquot y Tracy-Ventura definen el corpus de aprendientes como una "colección de textos informatizados que consta de muestras representativas del idioma escrito y/o hablado de los alumnos de un idioma adicional" (2020, 23). Los corpus de aprendientes se utilizan generalmente para probar teorías de adquisición de segundas y terceras lenguas, proporcionar información para la mejora de los materiales didácticos (Paquot y Tracy-Ventura 2020, 23), desarrollar los planes de estudios institucionales y evaluar la competencia en lenguas adicionales (LA) (Callies y Götz 2015, 1).

Gran parte de la investigación previa basada en corpus de aprendientes se ha centrado en el inglés LA, mientras que el cuerpo de estudios que se ocupan de la adquisición del español LA es relativamente pequeño. Sin embargo, durante las dos últimas décadas ha habido un interés creciente hacia este tema, lo que ha provocado un incremento notable en el número de corpus de estudiantes de español.

La gran mayoría de los corpus de aprendientes de español son transversales. Este es el caso de los tres que más estudios científicos han generado: el CAES, el CEDEL2 y el SPLLOC. El CAES (*Corpus de aprendices de español*)[2], promovido y financiado por el Instituto Cervantes y diseñado y desarrollado por la Universidad de Santiago de Compostela, es un corpus escrito que contiene actualmente un total de 764 796 elementos lingüísticos incluidos en 4 807 tareas realizadas por 1 830 estudiantes (*CAES* 2020). Este corpus ha generado varios estudios que han contribuido a la investigación sobre la adquisición del español como segunda o tercera lengua. Por ejemplo, Rojo y Palacios Martínez (2016) lo utilizaron para detectar falsos amigos y precisar las estrategias de comunicación que los alumnos usaban para compensar las deficiencias en el idioma de destino.

El *Corpus Escrito de Español L2* (CEDEL2)[3] es un corpus escrito de estudiantes de español LA con diferentes primeras lenguas (L1) (inglés, alemán, holandés, portugués, italiano, francés, griego, ruso, japonés, chino y árabe) compilado por un equipo dirigido por el Dr. Cristóbal Lozano, de la Universidad de Granada, que viene recogiendo muestras desde 2005. También contiene "varios subcorpus de control nativos (inglés, portugués, griego, japonés y árabe)" (Lozano 2021, 1). Hasta el momento 4 399 discentes han participado en el CEDEL2. El número total de palabras compiladas asciende a 1 105 936, lo que lo convierte en uno de los corpus más grandes de su tipo (*ib*., 9). Su interfaz, fácil de usar, lo hace muy práctico y muchos investigadores lo han utilizado para sus propias investigaciones. Entre otros, destacan los trabajos sobre la anáfora realizados por Martín-Villena y Lozano (2020).

[1] Todas las traducciones de este artículo son nuestras.
[2] Enlace de acceso: https://galvan.usc.es/caes.
[3] Enlace de acceso: http://cedel2.learnercorpora.com/.

El corpus SPLLOC (*Spanish Learner Language Oral Corpora*)[4] se creó en 2008 para proporcionar una gran cantidad de datos compilados con las producciones orales de estudiantes de español y llenar el vacío existente en esta área. Su objetivo era contribuir a la investigación sobre "pronombres clíticos, desarrollo de la frase verbal y desarrollo léxico" (Mitchell *et al*. 2008, 50). Hasta la fecha varios estudios han investigado estos aspectos utilizando este corpus (Domínguez 2019; Domínguez *et al*. 2011; 2013; 2017).

Por su parte, la investigación es significativamente más escasa si nos centramos en los estudios longitudinales. Hay dos corpus que deben resaltarse en esta sección: en primer lugar, el corpus de aprendientes longitudinal LANGSNAP contiene datos escritos y orales de estudiantes de francés y español cuya L1 es el inglés. Las muestras se recogieron durante un año académico. Se recopilaron más de 300 000 palabras en dos corpus de aprendientes longitudinales paralelos (Tracy-Ventura *et al*. 2016, 1). En segundo lugar, el *Corpus of Written Spanish of L2 and Heritage Speakers* (COWS-L2H) incluye datos de 1905 participantes y contiene 887 418 palabras (Yamada *et al*. 2020, 28). El COWS-L2H es de particular interés para este estudio porque señala la escasez de datos longitudinales en estudios de adquisición de español y se centra en los estudiantes "dentro de un único programa universitario homogéneo de idiomas" (*ib*., 30).

Con respecto a la investigación en adquisición del español basada en corpus de aprendientes, en el contexto de los estudios estonios debemos mencionar un corpus transversal: el *Corpus tartuense de español estudiantil académico* (ELEACTAR). Asciende a un total de 885 000 palabras, uno de los corpus de español más grandes indexados en la página web *Learner Corpora Around the World*. Está compuesto por 73 trabajos de grado redactados y defendidos en el Departamento de Lengua y Literatura Españolas de la Universidad de Tartu (Kruse 2018, 109). Este corpus es, sin duda, una gran contribución al campo de los estudios de adquisición de español en Estonia, dado que constituye una importante fuente de información. Sin embargo, tiene algunas limitaciones en cuanto al grado de espontaneidad de los textos, reducido por tratarse de trabajos académicos. Para complementarlo, es necesario otro corpus que contenga textos que se acerquen lo máximo posible a la interlengua real de los estudiantes y se produzca en varios niveles de competencia (*ib*., 178).

## 3 Enfoque longitudinal

La mayoría de las investigaciones sobre adquisición de segundas y terceras lenguas suele basar sus resultados en corpus transversales de aprendizaje, recopilados en un solo momento, con muestras de estudiantes que poseen diferente nivel de competencia (Granger 2008, 261). Por lo tanto, los conjuntos de datos obtenidos consisten en producciones de estudiantes que han adquirido el idioma de destino en diferentes contextos. Este método podría limitar la información si el objetivo de la investigación es el desarrollo de la interlengua y la relación entre competencia y adquisición, puesto que no permite investigar procesos, sino que solo ofrece cortes transversales del uso de la lengua meta (Hasko 2013, 2). Asimismo, los corpus longitudinales con un elevado número de muestras son indispensables para probar las teorías de adquisición de segundas y terceras lenguas porque permiten asegurar que las conclusiones alcanzadas puedan ser generalizables (Myles 2008, 70).

[4] Enlace de acceso: http://www.splloc.soton.ac.uk/.

La publicación del libro de Ortega y Byrnes (2008) sobre el vínculo entre la investigación longitudinal y la competencia de nivel avanzado de segundas lenguas (L2) contribuyó significativamente al incremento del uso de corpus longitudinales. Sin embargo, la revisión de la bibliografía especializada revela que todavía son escasos los estudios longitudinales de la adquisición del español como LA. Además, es difícil encontrar investigaciones basadas en corpus de aprendientes centradas en niveles de competencia avanzados, ya que hay una tendencia generalizada a concentrarse en las primeras etapas de aprendizaje (Meunier 2010, 1).

Una de las mayores ventajas de usar un corpus longitudinal radica en la posibilidad de comparar la evolución de la interlengua de aprendientes que estudian en un contexto homogéneo, es decir, con un nivel de competencia muy semejante. Permite a los investigadores tener en cuenta las características individuales de aprendizaje y centrarse en los patrones de desarrollo individuales. Asimismo, permite comprender en qué medida la exposición al *input* pedagógico y el conocimiento de otros idiomas afectan las producciones de los estudiantes. En conclusión, confiar en los corpus longitudinales es una forma adecuada de construir programas curriculares que reflejen la evolución real de la producción y adquisición de los alumnos.

## 4 CLEAE: *Corpus Longitudinal de Español de Aprendientes Estonios*

En la Sección de Español perteneciente al Departamento de Filología Románica de la Escuela Universitaria de Lenguas Extranjeras de la Facultad de Artes y Humanidades de la Universidad de Tartu, se está recopilando el *Corpus Longitudinal de Español de Aprendientes Estonios* (CLEAE). Se planea recoger muestras escritas y orales ―durante los tres años de los estudios de Grado en Filología Románica― producidas por los estudiantes que empiezan en 2021 la especialidad de Lengua y Literatura Hispánicas. Así pues, este corpus pretende convertirse en una fuente de estudio de diferentes aspectos relacionados con el proceso de adquisición del español por parte de los estudiantes estonios.

La decisión de elegir estudiantes de grado como informantes para nuestra investigación se ha tomado sobre la base de la necesidad de homogeneidad intragrupal. Queremos demostrar que algunos de los rasgos del proceso de adquisición de una lengua son comunes a todo el grupo (Jarvis y Pavlenko 2010, 229). Asimismo, es difícil encontrar un grupo homogéneo de aprendientes que estudien durante un periodo relativamente largo; por eso los estudiantes de Filología Hispánica se convierten en un tipo de informantes especialmente relevante para un estudio de estas características.

Creemos que el enfoque longitudinal puede ser muy beneficioso para esta investigación, ya que permite "generar conocimientos de vanguardia sobre la naturaleza de las capacidades avanzadas de segundas lenguas y formular prácticas educativas sólidas que permitan que cada vez más de nuestros estudiantes de idiomas las alcancen" (Ortega y Byrnes 2008, 3). Además, las manifestaciones de la influencia interlingüística en la interlengua de los estudiantes pueden detectarse más fácilmente si se toman "diferentes puntos en el desarrollo de la lengua de un solo individuo o de un grupo de individuos" (Jarvis y Pavlenko 2010, 232).

CLEAE sigue un modelo de diseño similar al corpus longitudinal de aprendientes adultos de sueco ―ASU (Hammarberg 2010)―, que recoge muestras escritas y orales. Se centra en el individuo y pretende que sea posible tanto comparar a cada informante consigo mismo a lo largo del tiempo como contrastar las soluciones lingüísticas y los perfiles de desarrollo de diferentes personas. Así pues, el CLEAE se propone compilar

numerosas muestras de relativamente pocos informantes y, al igual que el ASU, pretende describir cómo se desarrolla la interlengua a partir de sus formas más tempranas. Las diferencias que existen entre la producción escrita y la producción oral de los aprendientes también cobran en nuestro proyecto una importancia considerable.

Las tareas del primer curso ya están diseñadas y testadas en el corpus piloto (del que se hablará en el próximo apartado) y se están utilizando para recoger las muestras del CLEAE. Las de segundo y tercer curso están en proceso de diseño y comprobación. Los alumnos de primer curso de Lengua y Literaturas Hispánicas de Tartu tienen cuatro clases de lengua española a la semana. Al finalizar el curso académico, los discentes alcanzan el nivel de competencia A2. El manual utilizado es *Campus Sur. Curso Intensivo. A1-B1* (Rosales Varo *et al*. 2017), de la editorial Difusión, que incluye 10 temas para los niveles A1 y A2. Al final de cada tema, se lleva a cabo un control escrito y otro oral, con intención de mantener una evaluación continua del progreso de los estudiantes. Los controles tienen lugar cada tres semanas. Además, en enero y junio los alumnos tienen que realizar un examen oral y un examen escrito.

Cada control escrito recoge una actividad de expresión escrita libre, relacionada con el tema que se ha estudiado, excepto el control dos que contiene dos tareas. Estas son las actividades que se recogen como muestras para el corpus. Los controles orales contienen una actividad de lectura en voz alta para practicar la pronunciación y la entonación y tres actividades individuales o en grupo en las que hay que resolver algún tipo de tarea o responder a algunas preguntas de forma oral. Todas las actividades del control oral (excepto la de lectura) se graban y se transcriben para el subcorpus oral. Las producciones de los exámenes de febrero y junio también se unen al corpus. Los textos escritos tienen una extensión de entre 60 y 200 palabras y los textos orales, de entre 40 y 120 palabras, así como una temática variada.

Para el diseño de las tareas se han seguido los principios generales de diseño de corpus L2 de Lozano y Mendikoetxea (2013), según los cuales se plantean de acuerdo con criterios externos (función comunicativa de los textos), y no con criterios internos (lenguaje *per se*). Asimismo, se intenta que sean de naturaleza diversa para lograr la máxima representatividad, difiriendo en el grado de dificultad, la naturaleza del lenguaje producido y las restricciones de la tarea. Los textos son anotados con dos tipos de metadatos ―aquellos relacionados con las características del texto y los que se refieren a las particularidades del alumno― para favorecer que otros investigadores puedan utilizarlos en el futuro. Se siguen las pautas usadas por Lozano (2021) en el CEDEL2. Para cada texto se especifica si la muestra es oral o escrita, el día y la hora en la que se ha recogido, los objetivos funcionales del texto, las instrucciones recibidas por el alumno y la ayuda externa permitida (si la hubiera). Sobre cada alumno se apuntan la edad y el sexo, el tiempo que el alumno ha estudiado español, el tiempo que el alumno ha estudiado español en la Universidad de Tartu, el tiempo transcurrido en un país hispanohablante (si se diera el caso), la L1, la L2 o lengua dominante y su nivel de competencia (según una autoevaluación hecha a los propios estudiantes con anterioridad a la recogida de los textos).

Las tareas escritas se completan en la sala de ordenadores para obtener versiones digitales de los resultados de los estudiantes. Las muestras orales se recogen mediante la grabación de conversaciones mantenidas en Microsoft Teams o Zoom y se transcriben manualmente siguiendo las convenciones de CEDEL2 (Lozano 2021). La decisión de usar las convenciones de este corpus ha sido la consecuencia de un largo proceso de búsqueda y testado de diferentes propuestas. Se fueron descartando aquellas que hacían el texto demasiado obtuso, con la intención de que las transcripciones pudieran ser usadas en futuras investigaciones, independientemente de su naturaleza, así como para

fomentar el uso del corpus por parte de los propios alumnos de español. Por motivos de confidencialidad, no se transcriben los nombres ni la información personal de los participantes.

Al mismo tiempo que se recogen las muestras en español, se está compilando un subcorpus de control con muestras en estonio producidas por los mismos estudiantes y con textos que versan sobre los mismos temas. Algunos subcorpus de CEDEL2 (Lozano 2021) también serán utilizados para disponer de muestras de inglés L1 y L2 y muestras de español L1. Con estos subcorpus se pretende aumentar el número de análisis contrastivos. Así, nuestra intención es poder llevar a cabo comparaciones en las siguientes áreas: 1) entre diferentes niveles de competencia de un mismo individuo; 2) entre diferentes individuos; 3) entre discentes y hablantes nativos; 4) entre discurso oral y escrito; 5) entre géneros textuales (Hammarberg 2010, 4); 6) entre la interlengua y la lengua nativa de los informantes.

### 4.1 Corpus piloto

La investigación longitudinal busca mostrar cómo un fenómeno cambia con el tiempo. Este enfoque puede ser descriptivo o explicativo. En el primer caso, el investigador intenta únicamente describir la forma de cambio en el tiempo (por ejemplo, lineal o no lineal); en el segundo, lo que se busca es identificar la causa del proceso de cambio. Es esencial considerar ambas propuestas a la hora de conceptualizar la investigación longitudinal. De hecho, generalmente es necesario tener primero una descripción de la tendencia de cambio antes de intentar explicarla. Por lo tanto, antes de la primera ola de datos, se debe conceptualizar la forma de cambio en las variables de interés y luego formular las causas teóricas del cambio en esas variables (Ployhart y Vandenberg 2010, 99).

Con esta intención, se decidió compilar un corpus piloto que nos permitiera trazar la forma de cambio de la interlengua de nuestros estudiantes. El corpus piloto comenzó a compilarse durante el curso académico 2020-2021 y recoge muestras orales y escritas de los alumnos de Lengua y Literatura Hispánicas que comenzaron sus estudios de grado ese curso. Las ventajas de este corpus son numerosas, ya que nos permite 1) diseñar y testar las tareas del CLEAE; 2) tomar decisiones sobre la evaluación del nivel de los alumnos que ya han estudiado español con anterioridad; 3) elegir los metadatos que deben incluirse en el CLEAE y descartar aquellos que no sean necesarios; 4) elegir las herramientas que se van a utilizar para la recopilación de los textos; 5) escoger las convenciones de transcripción de los textos orales más adecuadas para este estudio; 6) diseñar y testar una taxonomía de errores; 7) tomar decisiones relacionadas con la anotación y el etiquetado de los textos.

Antes de empezar a recoger las muestras, se pidió a los estudiantes que firmaran un consentimiento mediante el que expresaban su conformidad tanto con que los datos fueran utilizados para este estudio como con que los textos fueran publicados para su posterior utilización en futuras investigaciones. Doce estudiantes lo firmaron a tiempo. Algunos alumnos entregaron el consentimiento firmado tras la recolección de datos, así que no fueron tomados en cuenta. Uno de los informantes abandonó el curso, por motivos personales, después de llevar tres meses estudiando, con lo que solo disponemos de las muestras que produjo en esos meses. Los estudiantes completaron las tareas del primer curso, descritas más arriba, y realizaron el examen oral. Por cuestiones relacionadas con la pandemia y, de manera excepcional, se decidió no realizar el examen escrito. Se recogieron 556 textos, de los cuales 430 son orales y están en

proceso de transcripción. Se estima que en noviembre de 2021 el subcorpus oral contaba con aproximadamente 36 600 palabras y 38 100 tokens. El subcorpus escrito posee en la actualidad 126 textos, 22 094 palabras y 26 338 *tokens*.

Cinco de los 12 alumnos no tenían ningún conocimiento de español cuando empezaron a estudiar en la Universidad de Tartu. Los alumnos que afirmaron tener conocimientos iniciales de español realizaron el examen de muestra de DELE A1 que se puede encontrar en la página web del Instituto Cervantes[5]. Esos estudiantes fueron clasificados entre nivel de competencia A1 (4 personas) y falso principiante ―aquellos que, pese a producir ciertos fragmentos de lengua, no llegaron a aprobar el examen con más del 60 % de la nota― (1 persona). Entre los estudiantes había también una alumna que afirmó poseer el certificado del DELE B2 y otra que había vivido en México y que, tras varias entrevistas y observaciones en clase, situamos en un nivel B1.

En esta etapa inicial de nuestra investigación, nuestros textos han seguido el sistema en tres niveles de etiquetado de errores de la *French Interlanguage Database* (FRIDA) (Granger 2003, 468). Los tres niveles del sistema aludido se refieren, primero, al dominio del error; segundo, a la categoría del error; y, por último, se añade la categoría de la palabra. Por razones de espacio no lo reproduciremos, pero intentaremos mostrar su funcionamiento con un ejemplo. Los segmentos de lengua que se alejan de la norma debido a cuestiones gramaticales se marcan con el símbolo <G>. Se registran varias categorías dentro de este domino: clase, género, modo, número, persona, tiempo y voz. En el dominio sintáctico <X> se encuentran los errores relacionados con el orden de palabras, la ausencia de palabras obligatorias, la redundancia de palabras y la cohesión. En el ejemplo siguiente, podemos identificar dos errores relacionados con el dominio gramatical ―uno de persona y otro de número― y un error sintáctico por ausencia de palabra necesaria.

(1) Producción del estudiante: *En Argentina vive cuarenta y cinco millón personas.*
(2) Texto con correcciones: En Argentina *vive [viven] cuarenta y cinco *millón [millones] [de] personas.
(3) Texto con anotaciones: En Argentina <G><PER><VERB> *vive [viven] cuarenta y cinco <G><NMR><SUST> *millón [millones] <X><FAL><PREP> [de] personas.

### 4.2 Informante EST_2020-2021_001

Ofrecemos algunas pinceladas de la interlengua escrita de un alumno a lo largo de un curso académico, para ilustrar la funcionalidad del corpus piloto. El informante catalogado con el nombre EST_2020-2021_001 tiene las siguientes características: es mujer; es estonia; tiene 19 años; antes de empezar la carrera en Filología Románica en la Universidad de Tartu no tuvo ningún contacto con el español; y su L2 dominante es el inglés. Las muestras escritas de esta informante a lo largo del curso 2020-2021 ascienden a un total de 1266 palabras y 1501 *tokens*.

El número de palabras por texto está distribuido de la siguiente manera:

[5] Enlace de acceso: https://examenes.cervantes.es/es/dele/preparar-prueba.

| Número del control | Nombre de la tarea | Número de palabras |
|---|---|---|
| 1 | Información personal | 74 |
| 2 | Habla de un país. | 83 |
| 2 | Habla de tu familia. | 84 |
| 3 | Una semana normal. | 206 |
| 4 | Correo electrónico sobre tu campus. | 143 |
| 5 | Presenta tu candidatura a un trabajo. | 94 |
| 6 | Biografía viajera. | 160 |
| 7 | Describe a una persona en el pasado. | 101 |
| 8 | ¿Cómo os conocisteis? | 105 |
| 9 | Habla sobre tu estilo u otra estética que te guste. | 104 |
| 10 | Escribe una receta. | 112 |

Tabla 1 – Textos y número de palabras

En estos textos se han detectado un total de 251 desviaciones, de las que más del 50 % pertenecen a la categoría de la sintaxis (véase el gráfico 1). Este primer acercamiento a la interlengua de la alumna nos puede servir para comprender cuáles son los aspectos que le provocan más dificultades. Las desviaciones sintácticas son, en su mayoría, provocadas por la ausencia de una palabra necesaria (112 de 139 casos). De los 112 segmentos de lengua en los que falta una palabra, 67 necesitan un artículo, una categoría de la que la lengua estonia carece, por lo que no nos resulta extraño comprobar que su uso puede provocar dificultades.

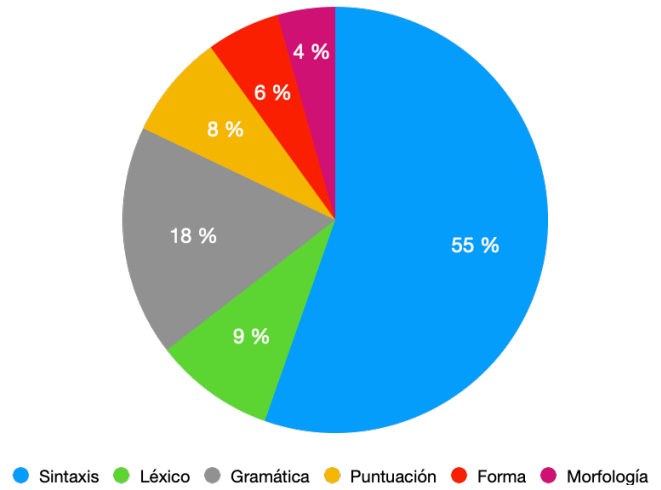


Gráfico 1. Distribución porcentual de desviaciones por categoría

De los 67 casos en los que hay una omisión de artículo, 50 adolecen de artículo determinado, 15 de indeterminado y en 2 ocasiones falta un artículo, pero tanto el

determinado como el indeterminado pueden usarse debido a la falta de contexto específico. Es interesante descubrir que los errores cometidos por omisión del artículo indeterminado empiezan a aparecer a partir del sexto texto, cuando la alumna ya lleva un semestre estudiando español. Son numerosos los estudios que intentan demostrar que existe una secuencia determinada en la adquisición de los artículos de la L2 cuando la L1 de los aprendientes carece de ellos (García Mayo y Hawkins 2009, 15). Consultados los segmentos de lengua producidos por nuestra informante durante el primer semestre, observamos que, efectivamente, apenas usa el artículo indeterminado en sus primeras seis producciones escritas. De hecho, solo lo encontramos registrado en un caso, en la expresión "un poco de" para hablar de las lenguas que conoce: *Yo hablo estonio, inglés, ruso y un poco de alemán y español*.

El uso del artículo determinado por esta informante es igualmente interesante (véase gráfico 2). No lo usa en ningún momento en su primera producción escrita, cometiendo, por ello, un segmento de lengua desviado: **Yo estudio en Universidad de Tartu.* En el segundo texto empieza a utilizarlo, pero el uso correcto fluctúa y se intercala con el uso erróneo, a veces, incluso en los mismos contextos: *El noventa y siete por ciento habla español*; **Cuarenta y dos por ciento habla inglés*. A partir del tercer texto, el uso del artículo determinado continúa variando. Se aprecia un descenso de los errores en algunos textos. Sin embargo, el incremento de errores provocados por la omisión errónea del artículo determinado, en los textos siete y once, y la reducida cantidad de segmentos de lengua con artículo determinado, en general, inducen a pensar que existe una tendencia por parte de la informante a evitar las estructuras con determinantes.

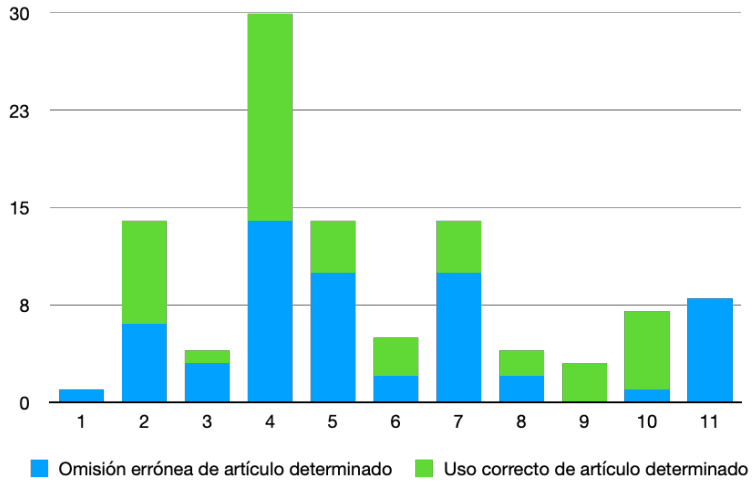


Gráfico 2. Comparación entre usos erróneos por omisión y usos correctos del artículo determinado

Es indispensable contrastar estos resultados con los del resto de informantes del corpus piloto para comprobar si estas tendencias pueden considerarse generalizables. De encontrar un patrón de comportamiento homogéneo podríamos afirmar que la adquisición del artículo por parte de los estudiantes estonios no es lineal y, a partir de allí, se podrían llevar a cabo diferentes estudios relacionados con la adquisición del artículo. La continuación de la compilación de los datos del corpus piloto durante los tres años de grado nos permitirá también comprobar de qué manera esta evolución cambia en los niveles intermedio y avanzado.

### 4.3 El confinamiento por la pandemia COVID-19

En Estonia, a partir de noviembre de 2020 y hasta junio de 2021, se tomó la decisión de estudiar en línea para evitar la propagación de la pandemia. Así pues, la mayor parte de las muestras escritas incluidas en el corpus piloto del curso 2020-2021 se recogió a través de los controles que se llevaron a cabo en la plataforma Moodle, en la que tenía lugar el proceso de enseñanza/aprendizaje sincrónico y asincrónico de la Universidad de Tartu. Los alumnos, en lugar de realizar los controles escritos en la sala de ordenadores, lo hicieron desde sus casas. De esa manera, las profesoras encargadas de esa asignatura no pudieron evitar que los alumnos utilizaran traductores automáticos, así como todo tipo de ayuda externa en la realización de los mismos. Desgraciadamente, esta situación afectó la calidad de los datos, que se vieron seriamente comprometidos, y nos hizo tomar la decisión de no incluirlos en el CLEAE.

De todas formas, no queremos restarle importancia al corpus piloto no solo por todas las decisiones que nos permite tomar y que hemos expuesto previamente, sino también porque nos brinda una gran cantidad de datos con los que podemos testar y aprender a utilizar las herramientas estadísticas y de etiquetado con las que vamos a trabajar en las investigaciones derivadas de la compilación del CLEAE.

## 5 Conclusión

Con este artículo hemos querido describir las bases teóricas y metodológicas sobre las que se asienta el proyecto de compilación del *Corpus Longitudinal de Español de Aprendientes Estonios* (CLEAE) de la Universidad de Tartu, con el que pretendemos llevar a cabo estudios relacionados con adquisición de terceras lenguas. Hemos defendido la propuesta longitudinal como única manera para comprender el proceso de adquisición de una lengua a través del tiempo y relacionarlo con los diferentes niveles de competencia.

Creemos que cualquier estudio longitudinal debe constar de dos partes: una descriptiva y una explicativa y, por ello, hemos justificado la existencia del corpus piloto previo al CLEAE, que sigue compilándose actualmente con datos de los mismos alumnos que ahora ya cursan el segundo curso de Filología Románica. A modo de ejemplo, hemos presentado algunos datos de una de nuestros informantes, con los que hemos comprobado qué esferas del aprendizaje de la lengua española le resultan más difíciles, comprendiendo que los aspectos sintácticos son los más problemáticos para ella, especialmente, el uso del artículo. Hemos descubierto que el número de errores más elevado ha sido provocado por la omisión del artículo determinado y que, en general, parece que la estudiante muestra una tendencia a evitar el uso del artículo. También hemos comprobado que el artículo indeterminado casi no aparece hasta las producciones del segundo semestre, con lo que nos hemos preguntado si tal vez existe realmente un orden de adquisición de artículos para aquellos alumnos cuya L1 carece de ellos.

Para finalizar, hemos discutido brevemente sobre la fiabilidad de los datos recogidos en tiempos de pandemia, llegando a la conclusión de que, pese a la falta de genuinidad de los textos recogidos durante el confinamiento, el corpus piloto sigue constituyendo una herramienta muy ventajosa para poder asegurar la calidad del CLEAE.

**REFERENCIAS BIBLIOGRÁFICAS**


Callies, M. y Götz, S. (Eds.) (2015). *Learner Corpora in Language Testing and Assessment*. Ámsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/scl.70.

Cenoz, J. (2008). The acquisition of additional languages. *Elia: Estudios de Lingüística Inglesa Aplicada*, 8, 219-224.

Domínguez, L. (2019). A 'mixed methods' approach for investigating aspect in a second language: Evidence from the SPLLOC project. *Dutch Journal of Applied Linguistics*, 8(1), 41–66. https://doi.org/10.1075/dujal.19011.dom.

Domínguez, L., Arche, M.J. y Myles, F. (2011). Testing the predictions of the feature-assembly hypothesis: Evidence from the L2 acquisition of Spanish aspect morphology. En N. Danis, K. Mesh y H. Sung (Eds.), *Proceedings of the 35th Annual Boston University Conference on Language Development* (pp. 183-196). Somerville, Mass.: Cascadilla Press.

Domínguez, L., Arche, M.J. y Myles, F. (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., Tracy-Ventura, N., Arche, M.J., Mitchell, R. y Myles, F. (2013). The role of dynamic contrasts in the L2 acquisition of Spanish past tense morphology. *Bilingualism: Language and Cognition*, 16(3), 558-577. https://doi.org/10.1017/S1366728912000363.

García Mayo, M.P. y Hawkins, R. (2009). Introduction: The interest of article acquisition for theories of SLA. En M.P. García Mayo y R. Hawkins (Eds.), *Language Acquisition and Language Disorders* (pp. 1–10). Ámsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/lald.49.03pil.

Granger, S. (2003). Error-tagged learner corpora and CALL: A Promising Synergy. *CALICO Journal*, 20(3), 465–480. https://doi.org/10.1558/cj.v20i3.465-480.

Granger, S. (2008). Learner Corpora. En A. Lüdeling y M. Kytö (Eds.), *Corpus linguistics. An International Handbook* (pp. 259–275). Berlín: de Gruyter. https://doi.org/10.1002/9781405198431.wbeal0669

Granger, S., Gilquin, G. y Meunier, F. (Eds.). (2015). *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.

Hammarberg, B. (2010). The language of the multilingual: Some conceptual and terminological issues. *International Review of Applied Linguistics in Language Teaching (IRAL)*, 48(2-3), 91-114. https://doi.org/10.1515/iral.2010.005.

Hasko, V. (2013). Capturing the Dynamics of Second Language Development via Learner Corpus Research: A Very Long Engagement: Capturing the Dynamics of L2 Development via Learner Corpus. *The Modern Language Journal*, 97(S1), 1–10. https://doi.org/10.1111/j.1540-4781.2012.01425.x.

Instituto Cervantes (2020). *Corpus de aprendices de español* (CAES). Recuperado de: http://galvan.usc.es/caes.

Jarvis, S. y Pavlenko, A. (2010). *Crosslinguistic Influence in Language and Cognition*. Nueva York/Abingdon: Routledge/Taylor y Francis.

Kruse, M. (2018). *La transferencia en personas plurilingües. Los falsos amigos como un obstáculo y una oportunidad en la enseñanza y aprendizaje de lenguas extranjeras.* [Tesis doctoral]. Tartu: Universidad de Tartu.

Lozano, C. (2021). CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research. *Second Language Research*, 38(4), 965-983. https://doi.org/10.1177/02676583211050522.

Lozano, C. y Mendikoetxea, A. (2013). Corpus and experimental data: Subjects in second language research. En S. Granger, G. Gilquin, y F. Meunier (Eds.), *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead. Proceedings of the First Learner Corpus Research Conference (LCR 2011)* (pp. 313–323). Lovaina la Nueva: Presses Universitaires de Louvain.

Martín-Villena, F. y Lozano, C. (2020). Anaphora resolution in topic continuity: Evidence from L1 English–L2 Spanish data in the CEDEL2 corpus. En J. Ryan y P. Crosthwaite (Eds.),

*Referring in a Second Language* (pp. 119-141). Londres/Nueva York: Routledge/Taylor y Francis.
Meunier, F. (2010). Learner corpora and English language teaching: Checkup time. *Anglistik: International Journal of English Studies*, 21(1), 209–220.
Mitchell, R., Domínguez, L., Arche, M.J., Myles, F. y Marsden, E. (2008). SPLLOC: A new database for Spanish second language acquisition research. *EUROSLA Yearbook*, 8, 287–304. https://doi.org/10.1075/eurosla.8.15smit.
Myles, F. (2008). Investigating learner language development with electronic longitudinal corpora: Theoretical and methodological issues. En L. Ortega y H. Byrnes (Eds.), *The longitudinal study of advanced L2 capacities* (pp. 58–72.) Nueva York/Londres: Routledge/ Taylor y Francis.
Ortega, L. y Byrnes, H. (2008). *The Longitudinal Study of Advanced L2 Capacities*. Nueva York/Londres: Routledge/Taylor y Francis.
Paquot, M. y Tracy-Ventura, N. (Eds.) (2020). *The Routledge handbook of second language acquisition and corpora*. Nueva York/Londres: Routledge/Taylor y Francis.
Ployhart, R.E., y Vandenberg, R.J. (2010). Longitudinal Research: The Theory, Design, and Analysis of Change. *Journal of Management*, 36(1), 94–120. https://doi.org/10.1177/0149206309352110.
Rojo, G. y Palacios Martínez, I.M. (2016). Learner Spanish on computer: The CAES 'Corpus de Aprendices de Español' project. En M. Alonso-Ramos (Ed.), *Spanish Learner Corpus Research*. *Studies in Corpus Linguistics* (pp. 55–87). Ámsterdan: John Benjamins Publishing Company. https://doi.org/10.1075/scl.78.03roj.
Rosales Varo, F. Moreno, T., Martínez-Lara, A., Salamanca, P., Buyse, K., Martínez, M., Murillo, N y Garrido, P. (2017). *Campus Sur. Curso intensivo de español. A1-B1*. Barcelona: Difusión.
Tracy-Ventura, N., Mitchell, R. y McManus, K. (2016). The LANGSNAP longitudinal learner corpus: Design and use. En M. Alonso-Ramos (Ed.), *Spanish Learner Corpus Research*. *Studies in Corpus Linguistics* (pp. 117–142). Ámsterdan: John Benjamins Publishing Company. https://doi.org/10.1075/scl.78.05tra.
Yamada, A., Davidson, S., Fernández-Mira, P., Carando, A., Sagae, K. y Sánchez-Gutiérrez, C. (2020). COWS-L2H: A corpus of Spanish learner writing. *Research in Corpus Linguistics*, 8(1), 17–32. https://doi.org/10.32714/ricl.08.01.02.