DIPLOMOVÁ PRÁCE

2021 SARA GARCÍA FERNÁNDEZ

JIHOČESKÁ UNIVERZITAV ČESKÝCH BUDĚJOVICÍCH
FILOZOFICKÁ FAKULTA
Ústav romanistiky

DIPLOMOVÁ PRÁCE

Los corpus de aprendices de español: análisis contrastivo de la interlengua de estudiantes extranjeros respecto a fenómenos gramaticales conflictivos

Vedoucí práce: PhDr. Jana Pešková, Ph.D.
Autor práce: Sara García Fernández
Studijní obor: Románská filologie – Španělská filologie

2021

Declaro que soy autor(a) de este trabajo académico final y que lo he elaborado tan solo a raíz de las fuentes y de la literatura presentada en las referencias bibliográficas.

Sara García Fernández, 7 de mayo de 2021

Resumen

En este trabajo hemos llevado a cabo un estudio contrastivo entre tres lenguas (español, checo e inglés) en lo relativo al uso explícito o tácito de los pronombres personales sujeto, ya que puede suponer un foco de conflicto para estudiantes de español checoparlantes o angloparlantes, en cuyas lenguas maternas el funcionamiento de este fenómeno gramatical no es idéntico ni exactamente equivalente. Para ello nos basaremos, a raíz de un proyecto de colaboración en el que hemos participado, en el proceso de revisión, corrección y etiquetado de determinado tipo de errores en una sección limitada de textos procedentes del corpus de aprendices estadounidenses denominado *Corpus of Written Spanish of L2 and Heritage Speakers* (COWS-L2H), y, por otro lado, en un pequeño corpus de aprendices checoparlantes de creación propia y analizado a través de la herramienta Analec. El estudio realizado se contextualiza en el marco de la evolución de la lingüística de corpus como metodología en pleno auge, aplicable a múltiples áreas de investigación en lenguas a través del desarrollo de distintos tipos de corpus. En este caso concreto, nos interesan especialmente aquellas bases de datos lingüísticos compilados a partir de redacciones de aprendices de español, ya que permiten analizar la interlengua de los estudiantes para identificar y determinar las facilidades y dificultades a las que deberán hacer frente a lo largo de su proceso de aprendizaje. Siguiendo esta línea, el profesor podrá elaborar una metodología más efectiva y apropiada, acorde a las características concretas de su(s) grupo(s) meta. No obstante, para ello resulta esencial fomentar la exposición de la comunidad docente y estudiantil a este tipo de recursos.

PALABRAS CLAVE: corpus de aprendices, interlengua, estudio contrastivo, pronombres personales sujeto, angloparlantes, checoparlantes, lingüística de corpus, COWS-L2H.

Abstract

In this paper, we conduct a contrastive analysis among three languages (Spanish, Czech and English) as far as the explicit or implicit use of subject personal pronouns is concerned, given that it can entail a source of linguistic conflict from the standpoint of English and Czech speakers who learn Spanish, as this particular grammatical phenomenon works differently in their mother tongues. The issue of this study is firstly and mostly grounded on a collaboration project in which we have participated along this past academic year, focused on the revision, correction and annotation of certain mistakes detected in a section of texts compiled in an American learner corpus called *Corpus of Written Spanish of L2 and Heritage Speakers* (COWS-L2H). The second basis for our research is a small, own compiled corpus of Czech-speaking learners of Spanish, which we analysed using a special software for linguistic annotation called Analec. Our study can be contextualized within the framework of corpus linguistics as a growing and evolving methodology, applied to multiple language research fields via the development of different types of corpora. In this case, our main concern are those linguistic databases that contain essays from students of Spanish, since these materials allow for the analysis of their interlanguage in order to identify and determine the prime faculties and difficulties that they will face when learning this language. Following this path, teachers will be able to develop a more effective and suitable pedagogical approach, in accordance with the specific characteristics of the target group(s). However, to attain this goal, it is vital to foster the exposure of educators and students to these linguistic tools.

KEY WORDS: learner corpora, interlanguage, contrastive analysis, subject personal pronouns, English speakers, Czech speakers, corpus linguistics, COWS-L2H.

Índice

Índice de ilustraciones

1. Introducción

Existe una realidad en el proceso de aprendizaje de una lengua extranjera que rara vez recibe la consideración de la que es meritoria: cuando entramos en contacto y estudiamos una lengua que no es la nuestra, los aprendices nos encontramos en una especie de limbo, de "tierra de nadie" entre dos idiomas, cada uno con sus respectivas culturas, y por mucho que lleguemos a dominar la lengua extranjera que estamos aprendiendo, siempre requeriremos de las dos para comunicarnos de manera efectiva (Valverde Ibáñez, 2018). La razón es que, interna y emocionalmente, nuestra capacidad expresiva está vinculada tanto a la lengua materna en la que hemos crecido y que nos ha formado como seres humanos como al idioma que necesitamos utilizar como vehículo de comunicación con aquellos que no comparten nuestra lengua nativa. De esta realidad nace el concepto de interlengua, una noción clave en los tres ejes temáticos en torno a los que gira el presente trabajo: los corpus de aprendices, la didáctica de idiomas y los estudios contrastivos. Una triada de conceptos estrechamente relacionados en la actualidad, aunque por desgracia su vínculo no capte la atención que debería por parte de muchos profesionales de la lingüística. Y es que de esta relación triangular se pueden obtener muchos beneficios para el progreso en el estudio de las lenguas, pues estos elementos se influyen positivamente entre sí. De ahí el interés por explorar el potencial de cada uno por separado y de su conexión.

Si hace cuatro años, cuando fuimos expuestos a un corpus por primera vez, nos hubieran vaticinado el futuro y nos hubieran predicho lo útiles y fundamentales que iban a resultar en nuestra ulterior vida académica (y, casi con toda seguridad, profesional), probablemente habríamos pecado de escepticismo. No por suspicacia hacia el potencial de estas herramientas, sino por desconocimiento de todas sus aplicaciones y posibilidades. Porque lo cierto es que, a lo largo de nuestros casi seis años de estudios superiores relacionados con las lenguas, hemos tenido un contacto limitado con estos recursos, sus utilidades y objetivos. Por tanto, nunca los hemos utilizado y aplicado de manera directa… hasta ahora. Y ha sido entonces cuando nos hemos percatado de su auténtico valor para docentes, aprendices y cualquier profesional o estudiante vinculado a los idiomas. Resulta desalentador observar cómo herramientas tan meritorias y provechosas pasan desapercibidas o se abordan de manera superficial en la formación de futuros lingüistas, profesores, traductores, filólogos… Si pretendemos que la ciencia lingüística progrese, hemos de abrir nuestras puertas a las innovaciones, a recursos

revolucionarios que sacudan los cimientos de la tradición purista, lo preestablecido, lo "clásico" … solo así el ser humano ha conseguido evolucionar a lo largo de la historia. En este sentido, los corpus entraron muy tímidamente en el mundo de la lingüística y paulatinamente han logrado hacerse un hueco que, sin embargo, aún tiene que expandirse mucho más si aspiramos a aprovechar todo su potencial.

El desarrollo de los corpus y de su diversidad tipológica muestra el interés de la comunidad lingüística por avanzar y por diseñar herramientas que no solo faciliten el trabajo de los profesionales de las lenguas en la medida de lo posible, sino que les permitan desarrollar, adaptar y aplicar metodologías de trabajo de mayor calidad, más eficientes, más acordes a su función y objetivos laborales. Por ello, resulta sorprendente que los esfuerzos de los lingüistas por progresar no tengan siempre un reflejo directo en la instrucción de las futuras generaciones de lingüistas, docentes, traductores, etc. El tratamiento de los corpus en titulaciones de grados y másteres ligados al estudio de las lenguas resulta, a nuestro juicio, insuficiente, puesto que apenas se estudian en profundidad su proceso de creación y diseño, su funcionamiento y sus utilidades. Es cierto que la comunidad académica universitaria está comenzando a adaptarse a la nueva realidad lingüística generada por este tipo de progresos, con asignaturas enfocadas exclusivamente al estudio directo de la lingüística de corpus. Sin embargo, consideramos que aún quedan muchas cuestiones por abordar. Por supuesto, somos conscientes de que no se pueden abarcar todos los conocimientos teóricos y prácticos existentes sobre uno o varios idiomas en el proceso de formación de profesionales de las lenguas. No obstante, en nuestra opinión, los corpus constituyen una de las herramientas básicas del procesamiento y análisis de datos lingüísticos actualmente, por lo que su estudio debería ocupar un lugar más preeminente en cualquier contexto de instrucción lingüística. Por ello, en el presente trabajo trataremos de abogar por una mayor exposición y por la aplicación de un enfoque multidisciplinar o multifunción en lo que a corpus se refiere, mostrando el inmenso potencial de explotación que este tipo de recursos tiene en el ámbito concreto de la didáctica de idiomas, a modo de ejemplo ilustrativo de su utilidad.

Como docentes de español en formación, siempre tendemos a buscar las claves de la enseñanza en manuales, artículos, estudios… sin reparar en que todos esos materiales basan su contenido en datos lingüísticos extraídos de alguna parte. Entre esas fuentes de información, consulta e inspiración se encuentran los corpus y, concretamente en el caso de la didáctica de idiomas, un tipo de corpus muy concretos: los corpus de aprendices. Si un corpus general contiene, de acuerdo con Tognini Bonelli (2010 citado en Rojo

Sánchez, 2016), textos o realizaciones tangibles (*parole*) de una lengua que nos permiten establecer patrones para estudiar su funcionamiento, reglas y fenómenos particulares (*langue*), un corpus de aprendices nos faculta, a través de muestras auténticas, fiables y objetivas (*parole*), para analizar un tipo de *langue* muy concreta: la interlengua de los estudiantes, es decir, el conocimiento y uso real que demuestran en el idioma que están aprendiendo. Por tanto, el objetivo fundamental de este tipo de corpus, más allá de otras posibles metas más específicas que puedan perseguir, consiste en examinar la actuación lingüística real de los estudiantes para evaluar su nivel de competencia y los obstáculos a los que se enfrenta respecto a la lengua extranjera, en lugar de presuponer un desempeño interactivo concreto en función del presunto estadio de dominio del idioma (ya sabemos que, en muchas ocasiones, el conocimiento teórico no tiene un reflejo directo en la práctica y, al fin y al cabo, a la hora de comunicarse, la aplicación práctica de la teoría es lo esencial).

A nuestro juicio, el análisis de la interlengua de los aprendices de español (o de cualquier otro idioma) constituye un procedimiento elemental para todo profesor y debería formar parte de su proceso de preparación para las clases si quiere desarrollar un enfoque didáctico realmente adaptado a las necesidades, facilidad, y dificultades de aprendizaje de su(s) grupo(s) meta(s). De este modo, podrá elaborar un plan curricular, materiales, explicaciones e incluso pruebas de manera más informada y eficiente, teniendo la seguridad, y no la intuición, de que van a resultar útiles para sus estudiantes. En este sentido, consideramos que los corpus de aprendices pueden revelarse una herramienta clave para conocer y estudiar esa interlengua que permitirá abordar de manera más conveniente, eficaz y fructífera la enseñanza de la lengua. Constituyen una base de datos de calidad, con un contenido auténtico, sólido y fiable que facilita la detección y análisis de aquellas estructuras lingüísticas empleadas con más frecuencia y de aquellos errores cometidos con mayor asiduidad en composiciones reales de diversos aprendices. Por tanto, estos recursos son el aliado ideal e imprescindible para el docente de idiomas.

Con el objetivo de apoyar esta postura, en el presente trabajo haremos un recorrido a través de la historia de la lingüística de corpus, su evolución y sus aplicaciones, tanto en términos generales como en lo relativo al español, analizando los criterios que determinan la calidad y adecuación de un corpus, así como los distintos tipos de corpus que existen, para pasar después a comentar en profundidad una clase de corpus muy concreta vinculada al ámbito de la didáctica de idiomas (del español en este caso, nuestro

foco de interés y trabajo): los corpus de aprendices. Analizaremos en qué consisten, cómo se han desarrollado, cuáles son los factores particulares de diseño de esta clase específica de corpus (más allá de los generales), cuáles son sus utilidades fundamentales y modos de aplicación en los ámbitos de la adquisición y enseñanza de idiomas, con ejemplos de investigaciones concretas, y cuáles son las ventajas y desventajas del uso de corpus de aprendices en ambos campos de estudio. Todo ello irá acompañado de una relación de los corpus de aprendices de español existentes en la actualidad, clasificados en función de diversos criterios: lengua(s) materna(s), nivel de competencia en español, tamaño, tipo de textos, etc. (Anexo I). Este conte.5nido teórico servirá como introducción a la sección de corte más práctico de nuestro trabajo, en la que expondremos el proceso y resultados de un estudio contrastivo español-inglés-checo elaborado a partir de la contraposición de dos corpus de aprendices, uno de informantes estadounidenses y otro de informantes checos. Esta investigación se ha inspirado en nuestra participación en un proyecto de corrección y etiquetado de errores de una fracción de textos contenidos en el *Corpus of Written Spanish of L2 and Heritage Speakers* (COWS-L2H). A partir de uno de los fallos que, como revisoras, teníamos que corregir y anotar en este proyecto de colaboración, la presencia/ausencia indebida de pronombres personales sujeto, hemos decidido analizar el fenómeno del uso/omisión de estos elementos en español y hemos establecido una comparativa del empleo que hacen estudiantes checoparlantes y angloparlantes de dichos componentes gramaticales a lo largo de diversos niveles de instrucción y dominio del idioma. A partir de este estudio contrastivo basado en corpus de aprendices hemos podido deducir cuáles son las facilidades y dificultades que muestra cada uno de los dos grupos en términos de asimilación del uso explícito y la omisión de los pronombres personales sujeto en función del estadio de aprendizaje en que se encuentras los estudiantes, de las transferencias positivas y negativas desde sus lenguas maternas, del contexto formativo, etc.

El potencial de aplicación didáctica de estudios como el que hemos desarrollado en este trabajo no hace sino confirmar las grandes posibilidades de explotación de los corpus de aprendices desde un punto de vista pedagógico. Ahora bien, para que podamos aprovechar de manera realmente útil estos recursos, resulta esencial que nos familiaricemos con ellos, que nos expongamos a su uso a lo largo de nuestro proceso de formación como profesores de lenguas extranjeras, ya sea de manera autónoma o a través de una educación reglada. De estas reflexiones nacerán nuestras conclusiones, en las que abogaremos por una instrucción docente específica vinculada a este tipo de herramientas,

así como por la exposición de los alumnos a los corpus de aprendices para que entren en contacto con sus propias dificultades y facilidades, y las de otro tipo de estudiantes. Esto fomentará la empatía, el mayor conocimiento de la propia interlengua, la capacidad de autocrítica, autoanálisis e introspección lingüística, y, sobre todo, avances en el aprendizaje. En definitiva, como profesores de una lengua extranjera, es fundamental que tengamos siempre presente que "si pensamos en la explotación de los corpus de aprendices, una cuestión es qué datos podemos extraer de ellos y otra muy diferente es cómo usar de manera eficiente, y con qué finalidad, la información que obtengamos en los análisis" (Mas Álvarez & Gil Martínez, 2018, p. 47). A lo largo de las siguientes páginas trataremos de ilustrar el amplio abanico de posibilidades que nos brindan los corpus de aprendices y cómo se pueden aprovechar de manera útil y efectiva, demostrando así que la escasa familiaridad y exposición a estas herramientas quizá pueda ser motivo de desconocimiento, pero no de suspicacia y mucho menos de escepticismo… más bien todo lo contrario: la innovación debe fomentar nuestra confianza en el progreso y nuestro interés por conocer, por saber, por descubrir. Los corpus de aprendices podrían constituir uno de esos pasos hacia la evolución lingüística, por lo que merecen su hueco en todo estudio vinculado a las lenguas, incluida la enseñanza de idiomas. Esta es la opinión fundamental que pretendemos defender a lo largo del presente trabajo.

2. Lingüística de corpus: una ventana a un mundo de posibilidades

2.1. Definición, historia y desarrollo de la lingüística de corpus

Según apunta el Centro Virtual Cervantes en su *Diccionario de términos clave de ELE*, la lingüística de corpus es "una rama de la lingüística que basa sus investigaciones en datos obtenidos a partir de corpus, esto es, muestras reales de uso de la lengua" (Centro Virtual Cervantes, n.d. citado en Núñez Nogueroles, 2019, p. 177). Por tanto, este término hace referencia a un determinado conjunto de principios que conforman un enfoque metodológico que se puede adoptar para analizar estadísticamente diversas áreas de la estructura y composición de una lengua: morfología, sintaxis, ortografía, léxico…

Asimismo, como indica el Instituto Cervantes, este procedimiento de estudio lingüístico "se contrapone a una metodología basada fundamentalmente en la introspección" (Centro Virtual Cervantes, n.d.), lo cual implica que basa sus conclusiones e investigaciones en los resultados extraídos de análisis empíricos de datos auténticos, objetivos y habitualmente representativos del fenómeno o realidad examinados (en este caso, una lengua). Ese tipo concreto de datos o ejemplos de uso de un idioma es lo que conforma lo que se define como *corpus*.

En el presente trabajo, ampliando esa definición sintética de "muestras reales de uso de la lengua" elaborada por el Instituto Cervantes (Centro Virtual Cervantes, n.d.), entenderemos el concepto de *corpus* como "un conjunto de textos informatizados producidos en situaciones reales, que se han seleccionado siguiendo una serie de criterios lingüísticos explícitos que garantizan que dicho corpus pueda ser usado como muestra representativa de la lengua" (Alonso Pérez-Ávila, 2007 citado en Buyse & González Melón, 2013, p. 247). Es decir, un corpus constituye una colección de volumen más o menos amplio de producciones orales o escritas (o ambas) en formato digital que se han producido en contextos comunicativos reales, y que constituyen ejemplos de uso representativos de una variedad lingüística determinada. Para construir un corpus provechoso desde el punto de vista de su diseño y su uso intuitivo, fácilmente aplicable a la investigación, esos textos son almacenados y sometidos a un tratamiento informático concreto, adecuado y pertinente de codificación y anotación, lo cual posibilita la explotación fructífera de sus datos en estudios y proyectos de corte lingüístico y/o pragmático. Asimismo, es fundamental tener en cuenta que, durante la primera fase de creación de un corpus, esto es, la recopilación de muestras, se establecen determinados parámetros lingüísticos que servirán como "filtro" de selección de la información, con el objetivo de uniformizar el contenido e integrar en el corpus solo aquellos datos y fenómenos que resulten realmente representativos y/o interesantes en función de la meta que se pretenda alcanzar con la compilación del corpus.

Por tanto, podemos decir que las características principales que constituyen un buen corpus son la autenticidad, la naturalidad, la representatividad, la homogeneidad, la recopilación selectiva de información y un tratamiento informático apropiado de los datos. En línea con estas nociones, es fundamental tener presente que no todos los fenómenos y ejemplos que incluya un corpus serán gramaticales, pero no por ello resultarán menos válidos o significativos, ya que representarán de manera fiel y objetiva el uso real y (relativamente) espontáneo que se hace de la lengua o variedad estudiadas.

Esta precisión resulta especialmente pertinente en el presente trabajo, dado que nuestro objeto de estudio son los corpus de aprendices de español, recopilaciones de textos que habitualmente contienen usos agramaticales, síntoma típico del proceso de aprendizaje de un idioma.

En cualquier caso, cabe destacar que esta definición y estos rasgos asociados a lo que se entiende como un "buen corpus" en la actualidad se basan en los ejemplos de corpus contemporáneos con los que contamos, de los que ha nacido una considerable, interesante y novedosa producción científica. Sin embargo, la noción de *corpus*, así como los criterios para su creación, que contemplamos en este trabajo no siempre se han aplicado a la compilación de este tipo de recursos. Basta con hacer un repaso a la evolución de la lingüística de corpus para percatarse del cambio paradigmático y metodológico que ha experimentado esta rama de la lingüística en un periodo considerablemente breve de tiempo. La historia de la lingüística de corpus como enfoque metodológico propiamente dicho es relativamente reciente, pese a que la práctica de recopilar textos para estudiar determinados aspectos estructurales, lexicográficos y gramaticales de una lengua data ya del siglo XX (Barroso Jiménez, 2011). Por consiguiente, podríamos concebir la lingüística de corpus como un enfoque metodológico considerablemente antiguo cuyo proceder no se había "estandarizado" y "parametrizado" con el objetivo de explotar su potencial de análisis lingüístico lo máximo posible hasta finales del siglo pasado. La razón que motivó este progreso paradigmático es, fundamentalmente, la evolución y expansión de la ciencia informática en los años setenta y ochenta. El desarrollo tecnológico y computacional, y los avances en la metodología de la lingüística de corpus fueron fenómenos que se produjeron de manera paralela. En este sentido, cabe establecer una distinción entre la lingüística de corpus de los primeros años y la lingüística computacional, una denominación más contemporánea y cada vez más extendida, basada en el instrumento fundamental que ha permitido la (re)evolución y actualización de esta metodología: la computadora u ordenador, cuyo uso se vuelve más intuitivo y su utilidad, por tanto, más explotable. El tratamiento de los datos de un corpus a través de un ordenador permite aprovechar de manera más rápida, automática y eficiente todas las posibilidades que estos recursos y todas sus herramientas ofrecen para la investigación en el ámbito del análisis lingüístico, además de que se agiliza considerablemente el proceso de compilación de los textos que conforman dicho corpus.

Por otro lado, es fundamental tener en cuenta que el desarrollo de los recursos electrónicos no solo ha tenido un impacto directo en el tamaño de los corpus compilados,

sino también en el tipo de textos que estos incluyen en términos de género, variedad, contenido, registro…, lo cual provoca el desarrollo de una perspectiva más representativa en la lingüística de corpus. A modo de resumen, podríamos concluir que, gracias al progreso tecnológico y de los ordenadores, que ofrecen la posibilidad de acceder a textos en versión electrónica de forma rápida y efectiva, y almacenarlos en un formato que permita extraer información de ellos o añadírsela, la lingüística de corpus progresa desde una perspectiva cuantitativa a una cuantitativo-cualitativa en la que la calidad y el carácter significativo y "emblemático" de las muestras priman por encima de su volumen, una tendencia que se ha mantenido hasta la actualidad. De hecho, son aquellos corpus que prestan una menor atención a la envergadura de su contenido y se interesan más (o, al menos, de manera pareja) por la aptitud de las muestras en cuanto a los objetivos que persiguen con su creación los que resultan realmente provechosos para la investigación en el ámbito de la lingüística.

Este enfoque cualitativo de la lingüística de corpus ha originado, asimismo, el desarrollo de herramientas adicionales que permitan explotar todas las posibilidades que una compilación representativa de muestras lingüísticas (escritas u orales) puede ofrecer. En efecto, de esta replanteamiento y actualización metodológicos han nacido los recursos para el análisis de frecuencias, concordancias, colocaciones, distribución (orden) y combinaciones de palabras en sintagmas… (Labrador de la Cruz, 1997), así como aplicaciones y programas para el etiquetado y la lematización automáticas de las diferentes formas lingüísticas que permiten la creación de interfaces para una consulta rápida de datos a través de una recuperación selectiva acorde a los criterios de búsqueda que se establezcan en cada caso. Todo ello sumado al hecho de que, hoy en día, los corpus pueden ser consultados en línea desde prácticamente cualquier lugar del mundo con acceso a Internet no hace sino predecir el potencial que pueden tener estos recursos como motor de diferentes investigaciones de índole lingüística que pueden revelarse muy interesantes, innovadoras y fructíferas. A lo largo de estas páginas, intentaremos demostrar que, efectivamente, la aplicabilidad de estas herramientas es muy heterogénea y provechosa.

2.2. Criterios para la creación de corpus

Como se puede comprobar en el ámbito de la lingüística actual, el nacimiento y evolución de la lingüística de corpus en las últimas décadas como un enfoque metodológico con unos objetivos, un proceso de compilación, unos requisitos y unos procedimientos concretos ha iniciado un proceso de desarrollo y actualización en la investigación y/o análisis de las lenguas, ya que ofrece acceso a muestras reales de uso del idioma objeto de estudio en cada caso. Pero, ¿qué requisitos deben cumplir estas producciones textuales para construir un "buen corpus"? ¿A qué criterios debemos atenernos para que nuestro corpus sea de la mayor utilidad y calidad posibles? Para determinar cómo compilar un corpus eficiente, práctico y fructífero, John McHardy Sinclair (2005) propuso diez principios básicos que relacionaremos y explicaremos de manera breve a continuación, ya que en ellos se basa el diseño de distintos y numerosos tipos de corpus actuales (entre ellos, el CEDEL2, que abordaremos más adelante en este trabajo). Estos principios son los siguientes (Sinclair, 2005 citado en Lozano, 2009):

1. Contenido del corpus. Si queremos que nuestro corpus sea lo suficientemente abarcador y contenga ejemplos que huyan de la artificialidad, hemos de primar la comunicación frente a la presencia de determinados elementos lingüísticos para garantizar un uso natural de la lengua (aunque no sea del todo correcto), así como la aparición de una variedad de fenómenos morfológicos, sintácticos, léxicos…
2. Representatividad. El contenido del corpus debe ser una muestra fiel y característica de la lengua o variedad que deseamos estudiar y de las estructuras lingüísticas que le son propias.
3. Contraste. Los datos deben estar verificados y comprobados respecto a su validez (no tanto respecto a su corrección lingüística, como ya hemos comentado).
4. Criterios estructurales. El diseño del corpus debe seguir una estructura considerablemente intuitiva cuyos principios sean "reducidos en número y claramente separables los unos de los otros" (Sinclair, 2005 citado en Lozano 2009, p. 201): temática, autoría, tipología textual, niveles de competencia…
5. Etiquetado. Si queremos analizar el potencial científico que tiene un corpus realmente, además de los metadatos relacionados con el contexto en que se escribió cada texto y sus características generales, y la anotación no lingüística de cada texto (codificación y formato, marcas textuales que indiquen el inicio y fin de párrafos y frases…), lo ideal sería que cada muestra fuera acompañada de

anotaciones lingüísticas (etiquetas) basadas en el análisis de los textos, su morfología y su sintaxis: lematización, clasificación de las diferentes "partes del discurso" (categorías gramaticales), etc. En este sentido, como afirma el lingüista británico Geoffrey Leech,

> […] no se puede acceder a la información de un corpus si este no ha sido preparado para buscar datos de manera rápida, sistemática y automática. Además de digitalizarlo, es necesario codificarlo y etiquetarlo, y la forma en que se realicen estas tres tareas condicionará indefectiblemente las aplicaciones que dicho corpus pueda tener y el tipo de explotación que pueda acometerse (Leech, 1993 citado en Calero Fernández, Serrano Zapata & Gómez-Devís, 2020, pp. 207-208).

6. Muestra. El contenido del corpus debe basarse en ejemplos de uso completos, pese a la diferencia de longitud entre los textos, ya que las producciones incompletas pueden generar problemas de contextualización y/o comprensión de la información que contienen. La calidad del contenido debe primar siempre sobre la cantidad.
7. Documentación. A lo largo del proceso de compilación y diseño del corpus, es importante describir el contexto en que se producen los textos recopilados en el corpus para que resulte más sencillo solucionar posibles confusiones o analizar imprevistos en los resultados que se obtengan en el marco de una investigación.
8. Equilibrio. El contenido del corpus debe ser característico y representativos de todo lo que la variedad estudiada engloba, así como de los objetivos que se persigan con su creación, lo cual afecta a la tipología textual, la temática, los niveles de competencia, etc. considerados.
9. Tema. La temática debe tener un carácter genérico y estar vinculada a funciones comunicativas, no a la elicitación de estructuras sintácticas, morfológicas o léxicas concretas, ya que solo de este modo pueden obtenerse muestras realmente representativas, objetivas y genuinas del uso de un idioma o variedad.
10. Homogeneidad. La elaboración de un corpus debe tener por meta la uniformidad generalizada de su contenido, independientemente de su envergadura y sin caer por ella en la artificialidad.

Atendiendo a todos estos principios y a lo descrito previamente, podríamos concluir que un corpus debe contener muestras auténticas, objetivas, contextualizadas, completas, homogéneas temática y tipológicamente, anotadas y debidamente documentadas, organizadas, representativas y habitualmente sincrónicas del idioma

objeto de estudio. Estas cualidades son las que, en principio, constituyen lo que entenderíamos como "un buen corpus" en la actualidad: herramientas metodológicas que proporcionan un panorama preciso del estado actual de una lengua y constituyen un recurso más exacto, eficaz, natural, fiable y sistemático de análisis lingüístico con numerosas posibilidades de explotación. Sin embargo, para compilar una colección de textos verdaderamente real, representativa y pertinente para la investigación, el análisis lingüístico y la caracterización de una lengua o una variedad concretas, no debemos olvidarnos de una cuestión fundamental que se ha venido repitiendo sutilmente a lo largo de los párrafos anteriores: ¿qué objetivos se persiguen con la creación del corpus? Para explotar al máximo el potencial de estas herramientas en términos de producción científica, hemos de determinar de manera clara la función y finalidad de nuestro corpus desde el inicio con vistas a poder seleccionar adecuadamente su contenido y tomar otras decisiones procedimentales necesarias en el proceso de elaboración de la manera más provechosa y conveniente posible: tipología de textos, anotación, diseño, temática, tipo de informantes, etc. Y es que, definitivamente, "los criterios con los que se diseña, se cataloga y se etiqueta un corpus lingüístico determinan las aplicaciones que dicho corpus podrá tener" (Leech 1993 en Calero Fernández et al., 2020, p. 206). Por tanto, si tenemos la intención de crear un corpus para desarrollar nuestro estudio y este sigue una línea de investigación muy definida, enfocada a analizar un fenómeno lingüístico o pragmático determinado, aunque debamos tener siempre en cuenta los diez principios expuestos por Sinclair para constituir un corpus eficiente, no debemos olvidar las metas que perseguimos con la creación de esa colección de muestras, ya que de lo contrario su contenido no resultará pertinente ni práctico para nuestra investigación.

En cualquier caso, ahora que ya conocemos la evolución y las motivaciones de la lingüística de corpus, qué es un corpus y cómo se debe construir, y una vez que ya hemos esbozado el potencial o posibilidades que ofrecen estas bases de datos, podríamos afirmar casi con toda seguridad que esta herramienta puede revelarse (y, en efecto, se está revelando) como un recurso renovador, eficiente, útil y revolucionario en el ámbito de la investigación lingüística. De hecho, como asegura Mª Belén Labrador (1997), de la Universidad de León, "basarse en la observación de datos, en la evidencia textual, aporta muchas más ventajas que tomar la intuición como único criterio para discernir fenómenos lingüísticos y pronunciar aseveraciones sobre el lenguaje" (p. 64). Estas ventajas se incrementan aún más, como ya hemos indicado, si esos datos se encuentran disponibles en línea y pueden ser consultados a través de herramientas de búsqueda que permiten

filtrar o seleccionar con precisión la información de manera que recuperemos los datos y estadísticas que nos interesan o que necesitamos en cuanto al uso de fenómenos lingüísticos y/o pragmáticos, así como a su frecuencia, su gramaticalidad, sus contextos de empleo, su distribución geográfica, etc. Por consiguiente, comenzamos a percibir claramente que las posibilidades que ofrece un corpus para el progreso de la investigación lingüística son incalculables. A este respecto, no debemos olvidar, asimismo, que al uso y contenido de un corpus se le atribuyen una gran versatilidad y, en consecuencia, una considerable capacidad de explotación científica como características fundamentales, puesto que se pueden abordar los datos compilados en él desde diferentes perspectivas y enfoques metodológicos y/o analíticos (Biber, Conrad & Reppen, 1998 citado en Rojo Sánchez & Palacios Martínez, 2016). Por otro lado, esta terna de autores afirma también lo siguiente:

> Las características principales de un análisis basado en corpus pueden describirse del siguiente modo: a) es empírico, ya que se requiere de un análisis y una compilación de datos. Se presta atención a patrones de uso extraídos de textos naturales […]; b) se basa en muestras de textos o en un "corpus", compilado con un objetivo concreto en mente y concebido como colección representativa de una lengua en particular; c) se utilizan principalmente ordenadores para el análisis; se recurre a técnicas y herramientas tanto automáticas como interactivas; y d) se pueden aplicar técnicas cualitativas y cuantitativas para extraer conclusiones sólidas[1] (Biber et al., 1998 citado en Rojo Sánchez & Palacios Martínez, 2016, p. 56).

De este fragmento, nos interesa centrarnos concretamente en el apartado b), en el que se insiste explícitamente en la naturaleza "definida" de los corpus en lo relativo a sus objetivos, los cuales resultan determinantes en el proceso de compilación, selección de la información y anotación (no) lingüística de las muestras. Esto reafirma la importancia que aquí consideramos que tienen los propósitos de un corpus en cuanto a su elaboración. Asimismo, esta diversidad de finalidades implicaría la existencia de distintos tipos de corpus dependiendo de las metas que se desee alcanzar o de la rama de la lingüística a la que se pretenda servir de base y herramienta en términos de análisis, investigación y producción científica. En efecto, de manera simultánea a la evolución de la lingüística de corpus como método lingüístico-estadístico a partir, sobre todo, del desarrollo tecnológico y la incorporación del ordenador a su proceso de compilación, codificación, etiquetado y consulta, hemos sido testigos de un incremento considerable en la variedad y diversidad de corpus, que va más allá de meras consideraciones de tamaño. Esto ha sido la consecuencia directa, precisamente, de "las posibilidades que esta forma de acceder al

[1] Traducción propia.

análisis de los fenómenos lingüísticos brinda a cultivadores de diferentes subdisciplinas lingüísticas, con intereses muy variados y orientaciones diversas" (Rojo Sánchez, 2016, p. 286). Procederemos ahora a hacer un repaso de la amplia tipología de corpus existentes con la que contamos en la actualidad, con el objetivo de confirmar definitivamente la versatilidad y el potencial científico del que hacen gala estos recursos.

2.3. Tipología de corpus

La diversidad tipológica en el ámbito de la lingüística de corpus responde a las crecientes, innovadoras y, sobre todo, distintas exigencias procedentes de la investigación lingüística y su rápido desarrollo. Cabe deducir, por tanto, que los distintos tipos de corpus se conformarán en torno a y con base en dichas necesidades científicas. El modo más sencillo de clasificar el amplio abanico de posibilidades que ofrecen las distintas clases de corpus de manera gráfica y comprensible es a través de un conjunto considerablemente numeroso de dicotomías. A continuación, destacaremos las más importantes y frecuentes (Sierra, Bel & Lázaro Hernández, 2018).

- Modalidad de lengua. En primer lugar, es habitual dividir las clases de corpus en función del medio o canal en el que se expresan los textos que lo conforman, es decir, la modalidad de lengua en la que se transmiten. Atendiendo a este criterio, los corpus pueden ser orales o escritos (o multimodales en aquellos casos en los que recojan muestras en ambos modos).
- Tiempo. Los corpus pueden ser sincrónicos o transversales, es decir, centrados en la lengua o variedad lingüística de un momento concreto, normalmente contemporáneo a la creación del corpus (aunque también pueden enfocarse en una época anterior específica), o diacrónicos o longitudinales, que contienen muestras lingüísticas de los mismos informantes (habitualmente) en diferentes períodos de tiempo para analizarlas, compararlas, relacionarlas y estudiar su evolución y sus cambios. Asimismo, los corpus diacrónicos pueden ser históricos, es decir, centrados en un momento pasado en particular, o cronológicos, cuyos textos abarcan varios períodos de tiempo en orden sucesivo o lineal. Los corpus longitudinales pueden permitir también el análisis transversal de sus datos, por lo que se pueden revelar más versátiles y explotables desde un punto de vista científico, aunque su recopilación supone un arduo y complicado trabajo.

- Objetivos. Según el propósito que se persiga con la creación del corpus, este puede ser general o multipropósito, que incluye textos de muchos géneros diferentes y esboza el panorama de una lengua en concreto en un registro estándar normalmente, o puede ser especializado y tener fines específicos. Como cabría esperar, los primeros son explotables para estudios e investigaciones lingüísticas de diversa índole, mientras que los segundos se enfocan a una rama o cuestión lingüística concretas, por lo que son útiles para todo análisis que se mueva dentro de esos límites determinados.
- Autores. Un corpus es genérico si contiene textos pertenecientes a un mismo género, pero escritos por múltiples autores, o canónico si se compone de las obras de un solo autor. Si el corpus no se ajusta a un mismo género ni a un mismo autor, es un corpus "de autoría variada" (Sierra et al., 2018).
- Codificación y anotación. Si un corpus no contiene ningún tipo de codificación ni marca para su contenido se denomina corpus simple. Si, por el contrario, la información compilada en el corpus está debidamente etiquetada con el objetivo de identificar de alguna manera los componentes de los distintos textos que lo conforman, entonces se trata de un corpus codificado o anotado. Esta anotación puede ser textual, morfológica, morfosintáctica, sintáctica, semántica, pragmática… o de más de uno de estos tipos. Asimismo, esta anotación "puede referirse a la clase de palabra (sustantivo, adjetivo, verbo…), al tipo de error (de adición, de simplificación, global, local…) o a cualquier otra característica que queramos señalar…" (Núñez Nogueroles, 2019, pp. 178-179).
- Espontaneidad. Un corpus espontáneo o no premeditado es el que contiene muestras totalmente naturales del uso del idioma, ya sean orales o escritas, producidas de manera más o menos desenvuelta y con la soltura y aleatoriedad propias del habla/escritura libres. Por el contrario, un corpus premeditado incluye textos procedentes de un planteamiento muy medido y unos criterios de producción concretos que actúan como guía para los informantes.
- Representatividad. Aunque no sea lo habitual, no todos los corpus son representativos de una lengua, variedad o fenómeno lingüístico concretos. Esto dependerá de los objetivos que se persigan con él, así como de la precisión en el proceso de selección y "criba" de las muestras que se incluyen en él y en el método empleado para obtenerlas. Por tanto, un corpus puede ser representativo o no.

- Acceso. Un corpus puede tener un acceso libre o restringido.
- Inclusividad. Es posible que un corpus adopte una perspectiva abarcadora, lo cual implica que incluirá muestras procedentes de las diversas variedades geográficas con las que cuente la lengua estudiada. No obstante, es también posible que los corpus se enfoquen a una variedad o norma concretas del idioma en cuestión para describirla y facilitar su análisis en mayor profundidad.
- Lengua. Los corpus monolingües son aquellos que contienen muestras pertenecientes a una única lengua, mientras que los multilingües incluyen textos en diferentes idiomas. Estos últimos pueden comprender muestras extraídas de un proceso de selección no demasiado minucioso (accesibilidad de textos, similitudes aparentes, etc.) con fines de corte más general, cuantitativos y o estadísticos (Sierra et al., 2018) o pueden dividirse, a su vez, en otros dos tipos. En primer lugar, pueden ser corpus paralelos, que cuentan con los mismos textos en diversas lenguas (originales y traducciones). En segundo lugar, pueden ser comparables, en los que se introducen muestras originales de diferentes idiomas pertenecientes a un mismo género o que comparten otro tipo de similitudes (tema, procedencia, longitud…), pero que no constituyen traducciones unas de otras.
- Procedencia de los informantes. Existen dos tipos de corpus atendiendo a este criterio: los corpus de hablantes nativos y los corpus de aprendices de un idioma (estos últimos pueden ir acompañados de un subcorpus de control de informantes nativos dependiendo de sus objetivos). En los segundos se centra, precisamente, el presente trabajo.

Como se puede comprobar, tenemos acceso a un amplio abanico de posibilidades en lo que a tipos de corpus se refiere. Esto no hace sino enriquecer y facilitar la investigación lingüística en todos los campos de estudio en que se aplique una metodología basada en corpus. Asimismo, la tipología de este tipo de bases de datos no constituye un conjunto cerrado: los avances en el análisis de las lenguas a lo largo de las últimas y próximas décadas son la materia prima de la que se nutren estas herramientas, por lo que es muy probable que el progreso lingüístico y sus novedades en cuanto a objetivos traigan consigo nuevas modalidades y tipos de corpus. Este hecho puede suponer un avance significativo en la lingüística de corpus en términos generales y, a

nuestro juico, también en la lingüística de corpus en español, lengua en torno a la que giran nuestro trabajo, estudio y reflexiones.

2.4. Lingüística de corpus en español

Hasta ahora hemos elaborado una introducción teórica sobre la lingüística de corpus en términos generales: historia y desarrollo, criterios de creación de un corpus y distintas clases que existen en la actualidad, con una propuesta de clasificación dicotómica para su amplia diversidad tipológica. Antes de enfocarnos en el paradigma concreto de corpus que nos interesa en este caso, los de aprendices de español, consideramos que es necesario realizar una breve revisión histórico-cronológica de la evolución de este enfoque metodológico en el ámbito hispanohablante para contextualizar en mayor profundidad el contenido que se expondrá a continuación en relación con los corpus de aprendices. Como resume de manera muy acertada Guillermo Rojo Sánchez, uno de los lingüistas españoles más involucrados en el desarrollo de corpus en español,

> La LC [lingüística de corpus] en español comenzó de forma relativamente tardía, pero ha experimentado un desarrollo muy rápido e intenso que, hasta cierto punto, se explica precisamente por el hecho de haber iniciado su desarrollo en un marco tecnológico más evolucionado, lo cual permite entender también algunos factores característicos de la LC en español frente a lo habitual en otras tradiciones (Rojo Sánchez, 2016, p. 286).

Por tanto, el hecho de que el progreso la lingüística de corpus se haya atrasado en el mundo hispanohablante ha constituido una ventaja en términos de desarrollo gracias a la solidez y estabilidad que los avances tecnológicos útiles y aplicables a la evolución de esta metodología habían alcanzado. Esto parece implicar que la lingüística de corpus española presenta ciertos rasgos muy particulares y distintivos, que son esencialmente dos; por un lado, la abundancia e influencia de grandes corpus de referencia monolingües, uniformes, tanto sincrónicos como diacrónicos y con una perspectiva panhispánica y multi-normativa pero uniforme, como el CREA, el CORDE o el CORPES (todos promovidos por la Real Academia Española); por otro lado, el acceso público y libre a estas valiosas bases de datos de consulta lingüística prácticamente desde su creación. En conclusión, podríamos decir que la evolución tardía de este método de análisis de datos provocó que los corpus españoles permitieran con mayor prontitud "obtener la visión general de un determinado fenómeno no solo en un momento determinado, sino también a lo largo del tiempo o del espacio" (Rojo Sánchez, 2016, p. 286). Teniendo en cuenta la

expansión del español y la multiplicidad de normas que rigen su uso alrededor del globo, esta celeridad en la innovación metodológica supuso un gran impulso para el desarrollo de la lingüística de corpus en el ámbito hispanohablante.

No obstante, no por haber nacido y madurado posteriormente en un contexto tecnológico, investigador y lingüístico más favorable debemos pensar que la evolución de la lingüística de corpus en español se produjo de manera rápida y casi espontánea. Al igual que sucedió en otros idiomas, esta rama de estudio siguió un proceso de desarrollo pausado y un tanto incierto y vacilante antes de reafirmarse como metodología "de pleno derecho". De hecho, en sus orígenes (en torno a los años sesenta), la lingüística de corpus española se basaba en estudios, investigaciones o métodos que no empleaban herramientas tecnológicas, pero compartían enfoques teóricos con ella y, por otro lado, en planteamientos y perspectivas que favorecían el uso de la automaticidad de la tecnología y de sus prestaciones como apoyo, pero cuyas metas y objetos de estudio nada tenían que ver con la lingüística de corpus más allá de la creación de listas de frecuencia y/o de concordancias a partir de una colección de textos (Rojo Sánchez, 2016). Paulatinamente, con el paso de los años, nacieron proyectos desarrollados en distintas universidades centrados en el español literario y que basaban sus estudios en la elaboración de listas de frecuencias, concordancias y diccionarios inversos a partir del análisis de conjuntos textuales de diversa índole. Las características de estos estudios los acercarían más a lo que hoy conocemos como *corpus*. Es más, la mayoría de los resultados de todos estos proyectos pasarían después a formar parte de corpus o archivos lingüísticos actuales.

Siguiendo esta línea evolutiva, llegados los años ochenta y a lo largo de los noventa, la creación, características y aplicación de los corpus comienza a responder más a la noción que tenemos de este tipo de recursos en la actualidad y se desarrollan hasta alcanzar el nivel y la calidad de los que ya existían en otras lenguas, en las que la evolución de la lingüística de corpus había sido más precoz. Según una propuesta de división de Guillermo Rojo Sánchez (2016), podríamos clasificar la tipología de corpus de esas dos primeras décadas, significativos en el desarrollo de este enfoque metodológico basado en corpus, en cinco bloques.

1. Corpus de poco volumen nacidos de proyectos desarrollados por particulares o por grupos de investigación. Su objetivo es de corte generalista: funcionan como bases de datos para estudios de muy diversa índole. Los ejemplos de este tipo de corpus han sido promovidos esencialmente por distintas universidades de Europa:

el corpus de la Universidad de Lovaina de los años noventa o los corpus ENTREVIS90 y ENTREVIS95 de la Universidad de Århus.

2. Corpus diseñados con fines lexicográficos incentivados por la aparición del COBUILD en el ámbito angloparlante. Algunos de ellos son el CUMBRE (para la elaboración del conocido diccionario que lleva el mismo nombre y otros recursos del estilo orientados a las frecuencias) y el *Corpus del español mexicano contemporáneo* (CEMC), del cual también nacieron diferentes diccionarios sobre esta variedad del español.
3. Corpus de volumen reducido desarrollados como parte de proyectos europeos, como el corpus multilingüe para la traducción técnica nacido del proyecto *Corpus Resources and Terminology Extraction* (CRATER).
4. Corpus multipropósito y no demasiado extensos. Es el caso, por ejemplo, del *Corpus Oral de Referencia de la Lengua Española Contemporánea* (CORLEC) o del *Corpus Léxico informatizado del español* (LEXESP), a partir del cual se conformó el *Diccionario de frecuencias de las unidades lingüísticas del castellano*.
5. Corpus de enfoque diacrónico. La colección más destacada de este grupo sería el *Archivo Digital de Manuscritos y Textos Españoles* (ADMYTE), de la Universidad Autónoma de Madrid, que se centra en documentos cuyo contenido está redactado en español medieval.

En cualquier caso, desde que Guillermo Rojo Sánchez propusiera esta clasificación "tipológica" de los corpus existentes en español han pasado dos décadas en las que esta metodología no ha dejado de crecer, perfeccionarse, reinventarse y actualizarse, lo cual ha fomentado su difusión científica, ya que diversas ramas de la lingüística han descubierto el inmenso potencial que este método de análisis puede aportar a sus investigaciones. Entre las áreas que más se han interesado por la aplicación y utilidad de los corpus se encuentra la didáctica de idiomas, que en los últimos años se ha dedicado a incentivar la compilación de corpus de aprendices, colecciones de textos producidos por alumnos que se encuentran en proceso de aprendizaje de una lengua extranjera. Estos recursos pueden revelarse una de las claves fundamentales en el progreso de la enseñanza de español o de cualquier otro idioma como segunda lengua. Así lo aseguran Inmaculada Mas Álvarez y Adelaida Gil Martínez, de la Universidad de Santiago de Compostela y del Instituto Cervantes de Burdeos, respectivamente, quienes

argumentan que “tener acceso a la interlengua de quienes están en el proceso de aprendizaje del español es una herramienta de gran interés para el profesorado, pues permite contrastar las dificultades de aprendizaje y el vocabulario empleado en la producción” (Mas Álvarez & Gil Martínez, 2018, p. 36).

Dado el gran potencial de explotación científica que muestran los corpus de aprendices y el gran paso que constituyen en la evolución tanto de la lingüística de corpus como de la enseñanza de idiomas, procederemos ahora a describir este tipo concreto de herramientas lingüísticas. Nos centraremos especialmente en sus propiedades más importantes, en los requisitos que deben cumplir para resultar eficientes y útiles, y en los distintos rasgos respecto a los que se pueden clasificar y definir hoy en día. A continuación, realizaremos una revisión de los corpus de aprendices de español más destacables en la actualidad (características, objetivos, producción científica asociada…) y comentaremos y analizaremos las ventajas y desventajas de esta clase de recursos, así como sus utilidades en el ámbito de la didáctica y adquisición del español (y de cualquier otra lengua).

3. Corpus de aprendices de español: la interlengua como motor de la didáctica

3.1. Definición, historia y desarrollo de los corpus de aprendices

Para determinar lo que entendemos en el presente trabajo por corpus de aprendices o aprendientes, tomaremos como base la definición de referencia aportada por Sylviane Granger a partir de la definición de corpus que hace Sinclair en 1996:

> Los corpus de aprendices son colecciones de datos textuales y auténticos procedentes de estudiantes de lenguas extranjeras o segundas lenguas y que se compilan de acuerdo con unos criterios concretos con fines didácticos o analíticos en cuanto a la adquisición de un idioma. Estos datos están codificados siguiendo un procedimiento estandarizado y homogéneo, y van acompañados de documentación acerca de su origen y proveniencia[2] (Granger, 2002 citado en Mas Álvarez & Gil Martínez, 2018, p. 36).

[2] Traducción propia.

Por tanto, estas colecciones lingüísticas se basan en muestras reales (textos orales o escritos) que, una vez que son recopiladas, se someten a un proceso de informatización y codificación (tanto las muestras propiamente dichas como los metadatos adicionales de cada una de ellas) para que su contenido pueda ser filtrado y recuperado de manera selectiva a través de una plataforma (normalmente online) y en función de los criterios de búsqueda que se introduzcan, para así lograr nuestros intereses. Algunos de esos metadatos son, por ejemplo, el sexo, la edad, la lengua materna, el nivel de competencia en la lengua extranjera, el contacto que se haya tenido con ella, el nivel de competencia en otros idiomas, etc. Estos aspectos identifican a los informantes según sus atributos sociales, lo cual puede revelarse interesante a la hora de hacer inferencias o conexiones a partir de los resultados de determinados análisis (Calero Fernández et al., 2020). En cualquier caso, dejando a un lado la información adicional que acompaña al contenido, lo fundamental, como bien indican Palacios Martínez y Sampedro Mella (2018), es que las muestras compiladas en un corpus de aprendices son "tratadas con el fin de que puedan ser representativas de la interlengua que se quiere estudiar" (p. 5). De este modo, los corpus de aprendices, al igual que cualquier corpus, permiten llevar a cabo estudios lingüísticos y/o pragmáticos de índole experimental, observadora o introspectiva (Ferreira Cabrera, 2018).

Los corpus de aprendices constituyen un recurso provechoso y eficaz para la investigación en el ámbito de la adquisición y la enseñanza de segundas lenguas, en esta última tanto en la fase de planificación como en la de aplicación al aula. Probablemente debido a este rango de empleo tan amplio, el análisis y el desarrollo de estudios a partir de corpus de aprendices requiere de un considerable conjunto de conocimientos y nociones que no resultan necesarios cuando trabajamos con corpus de nativos: un dominio notable de la metodología y procedimientos propios de la lingüística de corpus para la recopilación de muestras, su análisis y su filtrado; una base sólida de conocimientos teóricos de naturaleza lingüística en general, pero también vinculados a la adquisición de un idioma; una experiencia y una competencia relevantes en el ámbito de la enseñanza de lenguas extranjeras; la consideración del contexto y de aspectos sociales y cognitivos como elementos influyente en las producciones de los aprendices… (Granger, 2009). Esta preparación teórica previa resulta fundamental porque será lo que nos permitirá interpretar los datos lingüísticos incluidos en un corpus correctamente y así podremos aplicar todo el potencial de los resultados y conclusiones que extraigamos a diversos ámbitos de la lingüística, ya que la información que proporcionan se puede estudiar desde

ángulos muy diversos: pedagogía, adquisición de una lengua extranjera, lingüística de corpus… De entre estos enfoques cabe destacar una de las metodologías que más se ha desarrollado a partir de las ventajas que ofrecen los corpus de aprendices: el análisis de errores en el contexto de la enseñanza de idiomas, ya que el porcentaje de errores que contienen estas herramientas es más elevado y también se aprecian fenómenos de sobreuso, infrauso y usos incorrectos. A través de los corpus de aprendices se puede obtener información acerca de cómo los estudiantes aprenden la lengua extranjera que sea en cada caso y de qué cuestiones pueden resultarles más problemáticas, por lo que estos recursos han incentivado la actualización del análisis de errores y han facilitado enormemente su labor de detección, corrección y estudio de los fallos que cometen los aprendices de un idioma (Granger, 2009). No obstante, es importante recordar que los corpus de aprendices no contienen solo errores e incorrecciones, sino muestras esencialmente de la interlengua de hablantes no nativos, por lo que se trata de textos completos que presentarán usos tanto gramaticales y correctos como agramaticales. Es cierto que tradicionalmente el análisis de los escritos redactados por aprendices de un idioma se ha enfocado exclusivamente a la detección de errores, pero gracias a la evolución de los corpus constituidos a partir de las producciones de este tipo concreto de informantes se han desarrollado perspectivas más innovadoras que han diversificado de manera considerable la metodología aplicada a las investigaciones en este ámbito. Cabría concluir, pues, que la utilidad atribuida a los corpus de aprendices ha progresado notablemente desde el origen de estos recursos hasta la actualidad.

Aunque pueda parecernos sorprendente hoy en día, en un principio, se utilizaban corpus de nativos para planificar, programar y elaborar materiales para la enseñanza de idiomas, así como como fuente de inspiración para mejorar la práctica didáctica (Granger, 2002 citado en Núñez Nogueroles, 2019). No obstante, como la propia Granger también afirma, aunque el *feedback* de los informantes nativos puede ser un apoyo para la enseñanza de su lengua materna, no debería ser el único recurso de consulta, ya que no aporta información relevante acerca de aquellas cuestiones más problemáticas del idioma que pueden suponer un mayor desafío en el aprendizaje desde el punto de vista de un estudiante extranjero (Granger, 2002 citado en Núñez Nogueroles, 2019). Por tanto, siguiendo esta línea metodológica carecíamos de un acceso adecuado y provechoso a la interlengua de los aprendices que nos permitiera conocer realmente cómo la utilizan en cada caso. Para evitar esta situación desfavorable para el análisis y el progreso lingüísticos, lo ideal sería complementar los beneficios que los dos recursos aportan: los

corpus de nativos como base de datos "de control" y verificación respecto a cuestiones de gramaticalidad/agramaticalidad, frecuencia, etc., y los corpus de aprendices como fuente de información sobre la interlengua, errores y problemas de adquisición de los estudiantes extranjeros. De este modo, el planteamiento didáctico de una lengua podrá orientarse a una metodología mucho más eficiente.

Una vez que los lingüistas se hubieron dado cuenta de esto, comenzaron a desarrollar corpus de aprendices propiamente dichos a partir de los años ochenta. El primero en aparecer fue el *International Corpus of Learner English* (ICLE), dirigido por Sylviane Granger en la Universidad Católica de Lovaina. La primera versión (2002) contaba con dos millones y medio de palabras que se vieron casi duplicadas en la segunda (2009), con cuatro millones y medio, siempre extraídas de redacciones de carácter ensayístico-argumentativo (Valverde Ibáñez, 2020). Como se puede deducir del nombre del corpus, los informantes eran aprendices de inglés de diversas nacionalidades: españoles, italianos, franceses, rusos… lo cual implicaba un número considerable de lenguas maternas diferentes y fomentaba la mayor aplicabilidad o interés de este corpus.

De ahí en adelante, la elaboración de este tipo de recursos ha vivido una época de florecimiento y difusión creciente hasta llegar a nuestros días, donde se ha erigido como una herramienta de relieve en el ámbito de la enseñanza de idiomas y el estudio de su adquisición. Sin embargo, llegados a este punto nos encontramos con un problema: el desarrollo realmente productivo de este tipo de corpus se ha producido casi de manera exclusiva en el mundo angloparlante, como era de esperar dada la preponderancia del inglés en el panorama internacional de las últimas décadas. No obstante, la lengua que constituye nuestro foco de estudio es el español. ¿Qué ha sucedido con la evolución de estas herramientas en el ámbito hispano? ¿Contamos con el mismo volumen de recursos? ¿Son estos equiparables a los corpus de aprendices de inglés también en términos de calidad? Procederemos a continuación a analizar el *statu quo* de este tipo de bases de datos en el marco de la lengua española.

3.2. Corpus de aprendices de español: caracterización y clasificación

Entre el pequeño porcentaje de corpus de aprendices que no tienen como lengua meta el inglés, el español es uno de los idiomas en los que más se ha avanzado en el desarrollo de este tipo de recursos, pero, si consideramos todos los corpus de aprendices

que existen, aquellos destinados al estudio del español tan solo abarcan el 8% del total, lo cual no se corresponde con la posición que ocupa este idioma a nivel internacional como uno de los más hablados en el mundo (Alonso-Ramos, 2016). Es posible que precisamente su extensión haya dificultado el desarrollo de los corpus de aprendices en el mundo hispanohablante, dada la diversidad de normas asociadas a una lengua tanta extendida, pero esto no excluye el hecho de que el español sea la segunda lengua más hablada en el mundo a nivel nativo y una de las elecciones más habituales y crecientes a la hora de aprender una lengua extranjera. Precisamente por ello, como indica Margarita Alonso-Ramos (2016), "últimamente […] se han realizado grandes esfuerzos para llevar a cabo proyectos lingüísticos a gran escala desde una perspectiva representativa del español hablado a ambos lados del Atlántico"[3] (p. 5). En este ambiente internacional y más panhispánico, diversos corpus de aprendices de español de tamaño considerable han visto la luz, como son el *Corpus Escrito del Español L2* (CEDEL2) o el *Corpus de Aprendices de Español* (CAES), de los que hablaremos más en profundidad más adelante.

A raíz de este cambio de enfoque y del crecimiento y desarrollo tanto de la lingüística de corpus como de la aplicabilidad de los corpus de aprendices a la didáctica de lenguas, y aunque nuevamente de forma más tardía, la creación de corpus de aprendices de español se ha incrementado de manera notable en la última década, sobre todo como parte de proyectos a gran escala fomentados por diferentes universidades del mundo hispanohablante y con objetivos de la más diversa índole (Mas Álvarez & Gil Martínez, 2018). Tanto es así que, en la actualidad, los corpus de aprendices de español no sirven solo como apoyo a investigaciones de fenómenos lingüísticos y/o pragmáticos, sino también a estudios centrados en los corpus propiamente dichos. El fin de estos últimos consistiría en analizar el proceso de recolección de estas bases de datos, sus principales características, la idoneidad de sus diseños, los criterios que han seguido para erigirse como "buenos corpus" y las posibilidades de explotación científica que ofrecen.

No obstante, pese a que esta metodología de análisis de la interlengua de los aprendices esté en auge actualmente y se están fomentando su estudio y su aplicación a cada vez más ramas de la lingüística, el conjunto total de corpus de aprendices de español existentes hoy en día continúa siendo escaso. De hecho, numerosos lingüistas han insistido en la necesidad de compilar corpus longitudinales en el ámbito de la enseñanza del español y en distintas situaciones de aprendizaje que puedan servir como punto de

[3] Traducción propia.

partida para mejorar o adecuar los métodos actuales de enseñanza de este idioma a la evolución de la interlengua de un grupo meta determinado a lo largo de un cierto tiempo (Calero Fernández et al., 2020). Por ello, durante la última década, los corpus de aprendices de español no solo se han multiplicado, sino que además comparten generalmente una serie de características muy concretas, con el objetivo de ofrecer un abanico muy amplio de aplicaciones acorde a la diversidad de intereses de los diferentes enfoques o estudios lingüísticos contemporáneos. En cualquier caso, de acuerdo con la síntesis elaborada por Margarita Alonso-Ramos (2016 citado en Calero Fernández et al., 2020, p. 208), la mayor parte de corpus de aprendices de español actuales son "escritos, transversales, con textos que desarrollan temas prefijados, especialmente narraciones y textos argumentativos, redactados por informantes habitualmente anglófonos adultos que están aprendiendo español y que tienen distintos niveles de competencia". Asimismo, tienden a ser monolingües, multipropósito y representativos de la interlengua del/de los grupo(s) de informantes que se trate en cada caso. Por otro lado, si aplicamos la clasificación propuesta por Tono (2003 citado en Alonso-Ramos, 2016) en lo relativo al diseño de corpus de aprendices, las características principales de los corpus de aprendices de español, más allá de las que ya hemos mencionado, son el predominio de géneros como cartas y ensayos, de una temática general vinculada al entorno del aprendiz, de limitaciones de tiempo laxas o inexistentes, de una gran motivación y una actitud positivas hacia el aprendizaje, de contextos formales/académicos de enseñanza, y de informantes con inglés como lengua materna.

Es posible que el hecho de que la mayoría de corpus de aprendices sea transversal/sincrónicos y no longitudinal resulte sorprendente teniendo en cuenta la insistencia de los lingüistas en la imperiosa necesidad de crear corpus diacrónicos en el ámbito de las segundas lenguas, con el objetivo de evaluar el desarrollo de la interlengua en sus diferentes estadios, una metodología que puede revelarse muy útil para estudios en áreas como la didáctica, la adquisición, la psicolingüística... Sin embargo, lo cierto es que la compilación de corpus longitudinales constituye un proceso no exento de obstáculos que requiere de mucho tiempo para su elaboración. En este sentido, una solución por la que se está optando en los últimos años es la creación de corpus "quasilongitudinales" (Núñez Nogueroles, 2019, p. 179), que consisten en colecciones de textos producidos por aprendices de diferentes edades y/o niveles que permiten desarrollar un corpus con un diseño parecido al de los longitudinales.

Sea como fuere, como podemos deducir a partir de la caracterización elaborada en las últimas líneas, se pueden y se suelen aplicar los criterios de clasificación tipológica de los corpus en general a los corpus de aprendices también. Por otra parte, del mismo modo que en términos de categorización responden a los requisitos establecidos para todo tipo de corpus, su diseño también debe respetar los principios de creación de corpus propuestos por Sinclair que hemos mencionado en el apartado anterior. De hecho, según lo establecido por Inmaculada Mas Álvarez y Adelaida Gil Martínez (2018), los tres aspectos esenciales en la elaboración de un corpus de aprendices son, al igual que sucede con todos los tipos de corpus, los objetivos que se pretendan alcanzar, las variables que se tomen en consideración para la consecución de esos objetivos y la metodología seguida en la anotación. Sin embargo, como bien indican estas lingüistas, dado que los corpus de aprendices constituyen un tipo muy concreto de colecciones textuales, existen una serie de criterios o rasgos peculiares y característicos de esta modalidad de bases de datos que ellas resumen perfectamente del siguiente modo (Mas Álvarez & Gil Martínez, 2018):

1. **Variables relativas a los informantes**. Más allá de los metadatos típicos como la edad, el sexo, el nivel de lengua, la lengua materna, etc., se añaden otras que resultan de interés en el ámbito de la enseñanza y aprendizaje de idiomas. Algunas de ellas son la motivación, la situación de aprendizaje, conocimiento de otras lenguas, bilingüismo… Asimismo, el establecimiento del nivel de competencia en la lengua meta debe medirse de manera minuciosa y estandarizada (prueba de nivel) para garantizar la calidad y potencial científicos del contenido del corpus.
2. **Variables relativas al diseño de las pruebas**. Se deben tener claros ciertos aspectos relativos al proceso de compilación, análisis y tratamiento de los datos para que estos puedan ser igualmente aplicables a estudios en adquisición o didáctica de lenguas: sincronía-diacronía, espontaneidad-producción guiada experimental, duración del proceso de redacción, empleo o no de materiales de consulta, creación o no de un corpus de control de hablantes nativos para posibilitar análisis contrastivos… En función de los intereses de las diversas disciplinas que vayan a recurrir al corpus que estamos creando, priorizaremos unos factores u otros en su diseño.
3. **Variables relativas a las muestras de lengua**. Nuestro corpus de aprendices podrá ser oral (de carácter normalmente expositivo o interactivo) o escrito (diversidad de géneros: cartas, ensayos, relatos…). Asimismo, habremos de establecer la tipología textual a la que deseamos que respondan las muestras de

los estudiantes y la temática o temáticas en torno a la cual girarán sus producciones. A este respecto, es fundamental que desarrollemos un conocimiento y una conciencia críticos acerca de las ventajas e inconvenientes que presenta cada tipo de corpus para, de este modo, seleccionar adecuadamente la modalidad que más se ajuste a nuestros intereses y necesidades, y a los del público meta de nuestra colección de textos. En este sentido, Mas Álvarez y Gil Martínez apuntan lo siguiente:

> Se suele señalar que en los corpus escritos hay menos errores en la producción, los aprendices tienden a emplear estructuras más complejas —el lenguaje se presenta simplificado, se dice, en la producción oral de los no nativos— y se considera un medio particularmente apto para aprendices avanzados, especialmente en comparación con corpus de hablantes nativos. Los proyectos de corpus orales destacan, por su parte, la espontaneidad de la interacción cara a cara como premisa para obtener pruebas más directas de la interlengua, a la vez que minimizan los efectos de la autocorrección y monitoreo (Mas Álvarez & Gil Martínez, 2018, p. 39).

Dependiendo de los criterios que decidamos seguir en cuanto a estas tres dimensiones elementales del proceso de creación de un corpus, el método de compilación y diseño se realizará de una manera u otra, lo cual tendrá una influencia determinante en sus posibilidades de explotación. Por otra parte, una variable que Mas Álvarez y Gil Martínez no mencionan directamente, pero que también puede resultar importante, es el enfoque cuantitativo o cualitativo que se aplique al análisis de los datos de un corpus de aprendices. El primero nos permitirá conocer la frecuencia y/o recurrencia estadísticas de un fenómeno particular en la interlengua de los estudiantes, pero lo ideal sería que se viera complementado por un segundo análisis más "selectivo" que estudie la aptitud y el potencial de dichos datos si queremos obtener unos resultados realmente satisfactorios y útiles en nuestras investigaciones.

En esta línea, es importante que tengamos en cuenta que el desarrollo de los corpus de aprendices en términos de objetivos, diseño, contenido, enfoque, modalidad, lenguas, temática, estilo, etc. se verá siempre muy influido por el ámbito de estudio al que pertenezcan sus creadores o sus potenciales usuarios, así como de la aplicabilidad que se le quiera conceder. Esto provoca una divergencia considerable con respecto a los corpus elaborados a partir de datos aportados por hablantes nativos y nos lleva a establecer una diferencia entre tres tipos de corpus de aprendices en función de su público meta (Valverde Ibáñez, 2020), a saber, los destinados a profesionales del área de la didáctica (profesores, creadores de materiales, evaluadores…), los orientados a la investigación en

el campo de la adquisición de idiomas y los enfocados a estudios sobre el procesamiento del lenguaje o la lingüística computacional. Al primer grupo, el más amplio y común, pertenecen, por ejemplo, el *Corpus para el análisis de errores de aprendices de E/LE* (CORANE), el corpus *Aprender a Escribir en Lovaina* (Aprescrilov), y el ya mencionado CAES. Del segundo tipo, cabe destacar el CEDEL2 y el *Spanish Learner Language Oral Corpus* (SPLLOC) y en el último podría incluirse, entre otros, WordReference. En el presente trabajo, nos centraremos especialmente en los dos primeros tipos, pero, en cualquier caso, cabe destacar que estos tres grupos de corpus presentan características diferentes en algunos de los elementos que conforman su proceso de recopilación y/o diseño debido a los distintos intereses que tienen los destinatarios. Los investigadores están más centrados en describir "aprendices modelos" para cada lengua materna o contexto que se trate en cada caso, con el objetivo de comprender las dificultades que debe afrontar cada grupo cuando adquiere otra lengua. Por su parte, los profesores se enfocan más a la realidad del aula y a la aplicación práctica que sus alumnos hacen de los conocimientos que van estudiando y "aprendiendo". No obstante, en los últimos años se están buscando soluciones para reducir la brecha entre la investigación teórica y los estudios prácticos en lo relativo a las lenguas, de manera que los planteamientos y análisis de unos se puedan aprovechar de los resultados y conclusiones de los otros para mayor beneficio y mejora de ambas metodologías.

Llegados a este punto, una vez que hemos explicado en qué consisten los corpus de aprendices, cómo se ha desarrollado su elaboración y difusión en el mundo hispanohablante y cuáles son los criterios que hay que tener en cuenta a la hora de compilarlos, así como sus características principales y sus principales tipos, cabe preguntarse lo siguiente: ¿con qué corpus de aprendices contamos en español? ¿Cumplen con todos los rasgos y requisitos de diseño, contenido y tipología expuestos hasta ahora? Nos remitimos ahora a nuestro *Anexo I: Tabla de clasificación de corpus de aprendices de español*, en el que se incluye una relación de algunos de los corpus de este tipo y de mayor relieve que existen en la actualidad, ya sea por su impacto en el progreso metodológico de la lingüística de corpus, por la cantidad de producción científica que ha nacido de ellos, por su carácter innovador en cuanto al análisis lingüístico de determinados fenómenos, etc.

De cada corpus hemos tratado de aportar (no siempre ha sido posible dada la escasa información accesible sobre ellos) información sobre su nombre, la institución que ha promovido su creación y los compiladores principales (directores de proyecto), la

modalidad de corpus, algunas características importantes de los informantes (lengua materna, sexo, edad), los niveles de competencia que abarca el corpus, el tamaño, las fechas de compilación, los tipos de muestras que contiene, los objetivos principales que persigue, el enfoque adoptado con respecto a la temporalidad (transversal/longitudinal) y sobre si están disponibles/abiertos al público o no. Asimismo, aportaremos algún metadato más para caracterizar a los informantes y las muestras, y otros datos que pueden resultar de interés donde consideremos que resulta necesario. Por último, indicamos también si estos corpus están anotados o no, y qué tipo de anotación se les ha aplicado (metadatos sobre informantes y contenido, etiquetas lingüísticas o ambos)[4].

Como se puede deducir del listado de corpus de aprendices que presentamos en nuestro anexo, más allá de precisiones contextuales o relacionadas con una línea de investigación concreta, todos comparten un denominador común, un objetivo principal que se puede aplicar de manera generalizada a todos ellos y que va en línea con las características y requisitos de este tipo de corpus que venimos exponiendo hasta ahora. Este no es otro que el estudio de la interlengua de los aprendientes y el análisis de sus errores con vistas a fomentar y facilitar la investigación en el ámbito de la adquisición y de la enseñanza del español, así como la elaboración de materiales didácticos a partir de los resultados y conclusiones obtenidos. Así lo corroboran Inmaculada Mas Álvarez y Adelaida Gil Martínez (2018) al afirmar que "los datos observables que resultan en las producciones de las personas que aprenden una nueva lengua constituyen una vía esencial para conocer la interlengua" (p. 39).

Nótese que en la descripción de la finalidad de estos recursos hemos establecido una diferencia entre analizar la interlengua y analizar los errores de los aprendices. Esta distinción no es casual; a menudo en el ámbito de la investigación lingüística en esta rama se ha enfocado el estudio de la interlengua a un proceso de detección, categorización, aclaración, análisis y valoración de frecuencia/gravedad de errores exclusivamente con el objetivo de mejorar la metodología de enseñanza a través de las correcciones *a posteriori* necesarias para fomentar el progreso en la expresión de los aprendices. Es cierto que tradicionalmente este enfoque ha sido el más aplicado a los corpus de aprendices, tanto orales como escritos, en el marco de la lingüística contrastiva, que es el campo de la lingüística que se encarga, según lo describe Núñez Nogueroles, de lo siguiente:

> […] contrastar, desde un punto de vista descriptivo y sincrónico, dos o más lenguas (normalmente la lengua materna de unos estudiantes y la lengua meta que estos se

[4] Véase *Anexo I: Tabla de clasificación de corpus de aprendices de español* (páginas 113-115).

> encuentran aprendiendo) para, de este modo, determinar las similitudes y diferencias estructurales que caracterizan a esos sistemas lingüísticos objeto de estudio (Núñez Nogueroles, 2019, p. 171).

Sin embargo, las investigaciones actuales van un paso más allá y, aunque continúen utilizando el error como fuente de estudios contrastivos y análisis interlingüísticos, lo contemplan desde una postura más favorable, dada la utilidad y el impacto positivos que las incorrecciones tienen en el proceso de aprendizaje de un idioma. De hecho, el procesamiento, la clasificación y el etiquetado de errores en los corpus de aprendices ya no se hace únicamente con vistas a una corrección y/o explicación, sino con otros fines vinculados al proceso de desarrollo de la interlengua de un aprendiz y a lo valiosos que pueden resultar los resultados "erróneos" de un corpus de aprendices en investigaciones sobre adquisición de idiomas, sobre metodologías de enseñanza, sobre creación de materiales didácticos, sobre formación del profesorado… Esta reevaluación del concepto de *error* y su utilidad ha fomentado no solo que se desarrollen diversos modelos de anotación lingüística de errores, sino también que se puedan utilizar incluso como parámetros en los motores de búsqueda online a través de los que se puede filtrar el contenido de los corpus.

No obstante, en términos de potencial científico e interés de los errores, para que pueda ser realmente explotado en investigaciones sobre aprendizaje y didáctica de una lengua, el error debe ser representativo, significativo y característico. Esto implica que debe ser considerablemente frecuente en las producciones de un volumen notable de informantes que tengan algún rasgo particular y fundamental en común: lengua materna, nivel de competencia, contexto formativo… (Ferreira Cabrera, 2018). Solo de este modo las muestras de interlengua a las se puede acceder a través de los corpus de aprendices permitirán obtener resultados provechosos e interesantes que podrán servir como base de investigaciones innovadoras con conclusiones reveladoras y muy enriquecedoras tanto en el campo de la adquisición de segundas lenguas como en el de la didáctica de idiomas.

3.3. Utilidades y aplicaciones de los corpus de aprendices a la adquisición y enseñanza de español (y otras lenguas)

A lo largo de los últimos párrafos, hemos destacado en numerosas ocasiones que los corpus de aprendices de español, pese a perseguir un objetivo común de descripción

de la interlengua de los estudiantes para facilitar el análisis de sus errores, características y evolución, pueden tener aplicaciones enfocadas a diferentes tipos de investigaciones y profesionales. Para el presente trabajo, nos hemos limitado a estudiar sus utilidades fundamentalmente en dos ámbitos dentro de la disciplina de la lingüística: la adquisición y la enseñanza-aprendizaje de lenguas extranjeras. Por tanto, aunque ambas ramas puedan servirse del análisis de la interlengua desde diferentes ángulos para desarrollar sus estudios, cabe esperar que exploten las utilidades de los corpus de aprendices de manera más concreta también, siguiendo líneas de investigación precisas y enfoques particulares y más directamente vinculadas a su ámbito de estudio. Por otro lado, es fundamental tener en cuenta que las características de los corpus de aprendices influyen en la aplicabilidad de su contenido a las investigaciones en diversas ramas de la lingüística y, dada la inestabilidad de estos recursos en español a causa de su novedad, sus posibilidades de uso tienden a restringirse disciplinaria y metodológicamente de manera considerable. En este sentido, como la mayoría de corpus de aprendices se enfocaron en un principio a realizar aportaciones científicas a la didáctica del español y al análisis de errores, el terreno de la adquisición ha quedado considerablemente desatendido.

Lo cierto es que la relación del campo de la adquisición de segundas lenguas con los corpus de aprendices y con la lingüística de corpus en general ha sido cuando menos conflictiva. Esta disciplina se centra fundamentalmente en el estudio y descripción de las diversas fases en el proceso de aprendizaje de un idioma, ya sea materno o extranjero, por lo que parece evidente el beneficio que podría obtener del empleo de corpus de aprendices en sus investigaciones. Sin embargo, hasta hace relativamente poco tiempo, las bases de datos que se utilizaban en este ámbito eran de naturaleza experimental, observacional y fundamentalmente introspectiva (Ferreira Cabrera & Elejalde Gómez, 2017), por lo que no se basaba en datos auténticos y realmente objetivos del uso de la lengua para extraer conclusiones acerca de la evolución en el proceso de adquisición de un idioma. El argumento esgrimido por los profesionales de este campo se basaba en la dificultad de conseguir que algunos de los fenómenos que se pretendían estudiar se produjeran de manera natural en el habla/escritura espontánea de los aprendices, debido al carácter impredecible de las variables extralingüísticas que podían intervenir; de ahí que siguieran procedimientos más ortodoxos o rígidos que incentivaran directamente la aparición de dichos fenómenos, aunque fuera en detrimento de la representatividad y autenticidad de los datos obtenidos, lo cual dificultaba la completa aceptación y recepción de sus

conclusiones en el mundo de la lingüística, ya que eran difícilmente generalizables y probados.

Sin embargo, en los últimos años, los lingüistas de esta rama han admitido la ventaja que constituye para sus estudios el acceso a un volumen considerable de datos reales sobre el uso de la lengua como complemento necesario a sus fuentes de información experimentales e introspectivas previas (de las que no deberían y no tienen por qué prescindir), ya que les servirán para verificar los resultados extraídos de sus análisis y llegar así a conclusiones más interesantes, convincentes y "definitivas". Es posible que a esta necesidad de comprobación y constatación se deba el hecho de que las investigaciones en adquisición de segundas lenguas suelan evaluar el contenido y los resultados extraídos de los corpus de aprendices siempre en comparación con un subcorpus de control de informantes nativos (Mas Álvarez & Gil Martínez, 2018), de modo que las conclusiones sean válidas y eficientes para estudiar las teorías que nacen en el seno de su disciplina.

Así, algunos de los objetivos más específicos que persigue la investigación en adquisición de segundas lenguas a partir de corpus de aprendices son los siguientes:

- Analizar el nivel de desarrollo de la lengua estudiada en cada uno de los estadios de aprendizaje de estudiantes agrupados por diferentes características para describir de manera empírica de los niveles de referencia y ver en qué medida las diferencias en diversos factores lingüísticos y extralingüísticos son determinantes en la adquisición de un segundo idioma (lengua materna, edad, contexto académico…)
- Detectar en qué nivel de dominio de la lengua meta se cometen determinado tipo de errores habituales o recurrentes y con qué grado de frecuencia.
- Examinar la fosilización de ciertas estructuras fonológicas, morfológicas, sintácticas, léxicas y pragmáticas, tratando de buscar los motivos de esa fijación errónea y planteando posibles soluciones a través, en muchas ocasiones, de estudios comparativos entre diversas lenguas maternas.
- Investigar aspectos paralingüísticos, a menudo relegados a un segundo plano en los estudios sobre adquisición de segundas lenguas por falta de materiales o fuentes de autoridad sobre la que basar los argumentos, hipótesis y teorías formulados.

- Explicar el proceso de adquisición de distintos elementos del lenguaje, que pueden resultar especialmente problemáticos o interesantes desde un punto de vista investigador y/o didáctico, por parte de distintas poblaciones de aprendices de español dividas atendiendo a diversos criterios: nivel de competencia, lengua nativa, contacto con el español, contacto con otras lenguas extranjeras…
- Determinar de manera objetiva qué errores en el proceso de aprendizaje de una lengua se deben a interferencias con la lengua materna y en qué medida a través de análisis del tipo de error y su gravedad, y teniendo en cuenta el tiempo durante el que se ha estudiado el idioma, el nivel de dominio, las características de la estructura en la que se cometen esas incorrecciones…
- Identificar de manera empírica y sobre una base sólida de datos auténticos cuáles son los aspectos gramaticales, léxicos, morfosintácticos, pragmáticos, sociolingüísticos, etc. que entrañan una mayor dificultad de aprendizaje para los estudiantes de español (Palacios Martínez & Sampedro Mella, 2018), tanto en la modalidad oral como en la escrita, en términos generales y también en función de las características concretas de cada grupo (esto permite dirigir más acertadamente hacia esas áreas los materiales de español como lengua extranjera).
- Desarrollar estudios comparativos entre producciones escritas/orales de informantes nativos y extranjeros para confrontar la interlengua de los aprendices con la expresión nativa, o varias interlenguas entre sí, así como para analizar de manera contrastiva y en paralelo la adquisición de elementos gramaticales en la lengua materna y en la extranjera (Palacios Martínez & Sampedro Mella, 2018).

Podríamos asegurar, por consiguiente, que en la última década hemos asistido a un acercamiento entre la adquisición de lenguas extranjeras y los corpus de aprendices de español, ya que la primera reconoce a los segundos como un recurso científico representativo, fiable y de calidad que puede conceder autoridad a sus estudios e hipótesis, y en consecuencia ha diversificado sus métodos de trabajo para incorporar estas bases de datos a varias de sus líneas de estudio. No obstante, aún queda mucho camino por recorrer, puesto que las teorías e investigaciones que se han desarrollado en el campo de la adquisición de lenguas extranjeras a partir de corpus de aprendices no han progresado notablemente aún. Cabe concluir, por tanto, que todavía es difícil determinar el verdadero potencial de estas herramientas en el análisis de la adquisición de lenguas extranjeras, así

como en la aplicación didáctica de algunas de las líneas de investigación de esta disciplina.

En lo relativo a la didáctica de lenguas extranjeras, constituye el ámbito que más ha explotado el potencial y las utilidades de los corpus de aprendices por razones obvias, ya que su principal interés u objetivo es instruir a los estudiantes en el correcto empleo de una lengua diversa a la materna. En este sentido, es evidente que un recurso en el que se analicen las características de la interlengua de los hablantes, sus errores y los principales problemas que el aprendizaje del idioma extranjero les supone en función de determinados factores extralingüísticos (lengua materna, nivel de competencia, situación formativa…) será una pieza clave en el proceso de enseñanza y adquisición, ya que permitirá adaptar la metodología docente a las necesidades del público meta. Si analizamos cómo los aprendices utilizan el idioma de manera natural y relativamente espontánea en contextos comunicativos que incentiven la aplicación de las estructuras y léxico aprendidos en el aula, podremos comprobar los aspectos que constituyen un mayor obstáculo en el aprendizaje, cuáles se han fosilizado, cuáles son recurrentes en varios niveles y grupos, cuáles nacen de interferencias, etc. (Ferreira Cabrera & Elejalde Gómez, 2017). Y si existe una herramienta que permita estudiar todos estos factores de manera rápida, semiautomática, eficiente y significativa, por tratarse de bases de datos digitalizadas, accesibles a través de Internet y representativas, son los corpus de aprendices. La aplicación de estos recursos en el campo de la enseñanza de lenguas es multidimensional, en el sentido de que pueden ser explotados por tres tipos de usuarios distintos: los propios alumnos, los profesores y los investigadores en la enseñanza del español como lengua extranjera (ELE).

Desde este enfoque pedagógico, que es el que a nosotros más nos concierne, podríamos decir que se han adoptado dos métodos distintos para la aplicación de estos recursos a la realidad del aula: un uso pedagógico desplazado y un uso pedagógico inmediato (Granger, 2009). El primero es más común y se produce en un plano secundario con respecto a su aplicación, ya que se basa en la recopilación de datos con fines investigadores orientados a describir la interlengua de los informantes y poder desarrollar a partir de ello una metodología de enseñanza y unos currículos realistas y adecuados a los estudiantes a los que enseñamos, así como a crear materiales didácticos más adaptados a las necesidades de grupos concretos de estudiantes. La segunda metodología se ha empezado a aplicar recientemente y consiste en emplear el propio corpus de aprendices, elaborado por el profesor, como material de enseñanza-aprendizaje en el aula en forma

de diversas actividades en las que los alumnos sean los que interactúan directamente con la base de datos de manera autónoma o a través de la mediación del docente. La ventaja que ofrece este tipo de exposición más directa a datos lingüísticos reales es que son los estudiantes los que trabajan sobre los textos que ellos mismos han redactado, lo cual incentiva un proceso de reflexión e introspección lingüística sobre las diferencias estructurales entre la lengua materna y la extranjera que puede revelarse muy provechoso, al permitir al aprendiz sacar sus propias conclusiones. En una línea paralela, Leech (1997 citado en Palacios Martínez & Sampedro Mella, 2018) propone otra clasificación bipartita para las aplicaciones de los corpus de aprendices a la pedagogía, a saber:

- Usos directos que influyen directamente en la metodología y técnicas didácticas: estudios sobre el aprendizaje de ELE, análisis de la interlengua y de errores para descubrir obstáculos o áreas conflictivas, desarrollo de materiales didácticos (libros de texto, gramáticas, glosarios, diccionarios, etc.), de actividades y de prácticas concretas destinados a su implementación en el aula y adaptados a los factores más problemáticas para grupos específicos…
- Usos indirectos en los que el empleo de los corpus de aprendices en términos pedagógicos de produce en un segundo plano y con un objetivo más bien formativo u organizativo. Por ejemplo, en la instrucción de profesores para evitar lagunas en el aprendizaje y conocimiento de los estudiantes, para abordar el tratamiento de aspectos especialmente conflictivos o para determinar cómo se deben aproximar a la corrección de errores y qué comentarios deben hacer al respecto para extraer el máximo rendimiento de ellas. También se pueden aplicar las utilidades de los corpus de aprendices al diseño curricular y al análisis de la correspondencia entre el nivel de competencia real de los estudiantes y los contenidos establecidos por las programaciones para cada nivel de referencia. En este sentido, progresan la precisión y la calidad de las planificaciones y currículos de los distintos grados de dominio a partir de datos auténticos, objetivos, representativos y divididos por niveles que reflejan el dominio del idioma de los estudiantes (Palacios Martínez & Sampedro Mella, 2018). Otro ejemplo de uso indirecto de los corpus de aprendices se centra en la confección de pruebas de idiomas sólidas y realistas, acordes al nivel real de competencia de los estudiantes.

Sin embargo, cabe destacar que, pese al inmenso abanico de posibilidades de aplicación didáctica que ofrecen los corpus de aprendices y que los docentes reconocen,

su empleo se basa fundamentalmente en usos pedagógicos desplazados o en usos indirectos, dependiendo de la terminología que adoptemos. De hecho, en la mayoría de los casos los profesores solo recurren a estas herramientas para realizar consultas esporádicas de dudas lingüísticas, concordancias, frecuencias y contextos. En un segundo plano, y en una frecuencia mucho menor, los corpus de aprendices también se emplean para la confección de materiales, la formación docente, la consulta de ejemplos se interlengua y, en último lugar, la investigación en ELE (Abad Castelló, 2019). Algunas de las utilidades específicas que se le ha dado a los corpus de aprendices desde este enfoque más "diagonal" son las siguientes:

- Detectar deficiencias, frecuencias de uso, contextos de aparición, variedad, diferencias pragmáticas específicas, sobreuso e infrauso de estructuras, etc. en la interlengua de los hablantes con vistas a su corrección, práctica o trabajo (marcadores del discurso, anáforas, colocaciones, diferencias entre ser y estar, preposiciones, verbos de cambio...).
- Realizar estudios comparativos/contrastivos entre corpus de hablantes nativos y corpus de aprendices para analizar similitudes y diferencias entre formas y funciones y conocer así los factores más complejos en el aprendizaje y adquisición de una lengua extranjera, así como para detectar deficiencias y clasificar tipos de errores con tipos de soluciones más específicos.
- Estudiar la validez de los diferentes métodos de enseñanza de español en relación con el tipo de aprendientes al que van dirigidos
- Determinar la sistematicidad de un error y resolver si un estudiante o conjunto de estudiantes domina o no una regla en base a dicha sistematicidad.
- Examinar la correspondencia uso real-nivel de dominio en diferentes fenómenos lingüísticos para comprobar si la realidad se ajusta a lo establecido por la "norma".
- Analizar la riqueza diatópica, diastrática y diacrónica de una lengua, así como las distintas normas que rigen cada variedad para evitar catalogar un uso válido en alguna de ellas como un error.
- Observar e investigar el desarrollo de la variedad y la riqueza léxicas y/o gramaticales en textos orales y escritos.

La aplicación directa de los corpus de aprendices a tareas o actividades en el aula está muy relegada con respecto a sus usos indirectos, tanto si se trata de una interacción

directa del alumnado con el corpus como si hablamos de ejercicios que incluyen la intervención y guía del profesor. De hecho, la práctica de uso inmediato más habitual consiste en ejemplificar y explicar, a través de consultas espontáneas, un fenómeno lingüístico concreto mediante fragmentos de las producciones reales y contextualizadas que estos corpus contienen, con el objetivo de huir de descripciones simples e improvisadas que no muestran adecuadamente su utilización.

¿A qué se debe el desaprovechamiento de muchas de las posibilidades que estos recursos tan útiles ofrecen? A grandes rasgos, podríamos decir que la causa son las carencias formativas del profesorado en este tipo de recursos. No obstante, en los últimos años se está apreciando una corriente innovadora en el ámbito de la didáctica paralela al desarrollo de estos instrumentos renovadores que fomenta su aplicación a la realidad del aula desde diversos enfoques, más directos o más indirectos y, en consecuencia, proporciona una instrucción precisa sobre su uso a los docentes. De ahí que, en la actualidad, el ámbito de la didáctica de lenguas sea no solo la rama de la lingüística que más y mejor aplica las utilidades de los corpus de aprendices, sino también la que más ha diversificado su metodología investigadora y docente recientemente para incorporar el mayor número posible de posibilidades de aplicación que proporcionan estos recursos. En definitiva, el enfoque basado en análisis de corpus de aprendices nos ha aproximado al proceso de enseñanza-aprendizaje del español como lengua extranjera y ha impulsado el progreso metodológico, investigador y de formación docente en la didáctica de este idioma.

Como se puede deducir a partir de las utilidades específicas de los corpus de aprendices de español que aprovechan cada una de las dos ramas de la lingüística expuestas, los temas o cuestiones investigados a través de estos recursos giran fundamentalmente en torno a fenómenos gramaticales y léxicos: concordancias, flexión verbal, modo verbal, conjunciones, uso de tiempos verbales, colocaciones, falsos amigos, expresiones idiomáticas, sinónimos y antónimos, variedad y frecuencia del léxico… Así, Magdalena Abad Castelló (2019), profesora del Instituto Cervantes, recalca que los corpus de aprendices y las diversas herramientas que se les asocian, como las listas de frecuencia, las líneas de concordancia, etc. permiten contextualizar el término en concreto para verificar su forma y su contenido semántico, así como la asiduidad con que se usa y las limitaciones sintácticas que se aplican a la hora de utilizarlo, lo cual facilita la corrección y elaboración de tareas más originales, entretenidas y provechosas para el aprendizaje. Asimismo, otro factor analizado y trabajado habitualmente mediante las

producciones contenidas en estos corpus es la ortografía (estas herramientas permiten elaborar prácticas deductivas e inductivas para fomentar su correcto desarrollo). Por otro lado, existen aún áreas del aprendizaje y adquisición de una lengua extranjera en las que apenas se han explotado las utilidades de los corpus de aprendices, como la fonología, la sociolingüística, la dialectología, la variación lingüística y la pragmática. En cualquier caso, la metodología de este tipo de corpus es una práctica cada vez más extendida entre los profesionales de la lengua, sobre todo en el ámbito de la didáctica de idiomas, por lo que presagiamos un futuro prometedor para estos recursos también en estos campos de la lingüística y la enseñanza/adquisición de lenguas extranjeras.

Siguiendo la división establecida entre las utilidades de los corpus de aprendices para la investigación y aplicación en didáctica de lenguas y/o en adquisición de idiomas, y tomando en cuenta los aspectos más estudiados y analizados a través de estos recursos que acabamos de describir, procederemos ahora a comentar algunos ejemplos concretos de producción científica en estos dos ámbitos fundamentales a partir de determinados corpus de aprendices para destacar e insistir de nuevo en la versatilidad y las múltiples posibilidades que ofrecen estos recursos desde un punto de vista investigador y pedagógico, pero esta vez quizá de manera más concreta.

En el campo de la adquisición de lenguas, los dos corpus de aprendices de español más destacables son el CEDEL2 en la modalidad escrita y SPLLOC en la modalidad oral. Con respecto al primero, "constituye un proyecto sólido que cuenta con numerosos estudios publicados e investigaciones en curso" (Mas Álvarez & Gil Martínez, 2018, pp. 49-50). Los estudios publicados a partir del análisis de sus datos, muy variados y representativos[5], están enfocados sobre todo al análisis de la adquisición de la competencia léxica y colocacional de los estudiantes desde diversos ángulos (análisis contrastivos, análisis de errores, etc.) Algunos de ellos son *Frecuencia y corrección colocacional en la producción escrita de aprendices de español* (García Salido, 2017b), *Comparing learner's and native speakers' use of collocations in written Spanish* (García salido, 2017a), *Análisis de errores colocacionales en un corpus de aprendices de ELE* (Pérez Serrano, 2014), *Rasgos de la competencia léxica del verbo* (Sánchez Rufat, 2014), o *Colocaciones, diccionario y corpus de aprendices* (Alonso-Ramos, 2013) … Por otra parte, también se ha analizado la adquisición de cuestiones pragmáticas y gramaticales: *Pragmatic principles in anaphora resolution at the syntax-discourse interface: advanced*

[5] Véase *Anexo I: Tabla de clasificación de corpus de aprendices de español* (páginas 113-115).

English learners of Spanish in the CEDEL2 corpus (Lozano, 2016), *El uso transitivo y ditransitivo de dar en un corpus escrito contrastivo* (Sánchez Rufat, 2016), *Discourse markers in CEDEL2 and SPLLOC corpora of learner Spanish: Analysis of some lexical-pragmatic failures* (Vázquez Veiga, 2016), etc.[6]Con esto, los lingüistas han podido examinar y ratificar o refutar diversas teorías e hipótesis generales y básicas de la adquisición de segundas lenguas. Asimismo, han podido rechazar falsas, pero comúnmente aceptadas, suposiciones y han insistido en la necesidad de elaborar materiales didácticos a partir del contenido de este corpus de aprendices para facilitar la correcta adquisición y fomentar el tratamiento y ejercicio de fenómenos lingüísticos a menudo desatendidos en la enseñanza del español, como las colocaciones. Sin embargo, como indican Inmaculada Mas Álvarez y Adelaida Gil Martínez (2018), resulta complicado trasladar los resultados y conclusiones de estas investigaciones al ámbito de la didáctica de segundas lenguas de manera eficiente y realmente práctica.

En cuanto a SPLLOC, "cuenta con numerosas publicaciones derivadas y constituye un hito en la investigación de la adquisición del español como segunda lengua" (Mendikoetxea, 2014 citado en Mas Álvarez & Gil Martínez, 2018, p.43). Como se puede observar en la Anexo I, este corpus está divido en dos subcorpus o proyectos, SPLLOC I y SPLLOC II, cuyos principales objetos de estudio son, entre otros, la adquisición de pronombres clíticos, del orden de palabras, de la competencia léxica (estudio contrastivo) y del sistema tiempo-aspecto. En esta línea, algunas de las publicaciones procedentes del análisis de datos de SPLLOC son *Optionality in L2 Grammars: the Acquisition of SV/VS Contrast in Spanish* (Domínguez & Arche, 2008), y *Vocabulary use during conversation: a cross-sectional study of development from year 9 to year 13 amongst learners of Spanish and French* (Mardsen & David, 2008).[7]

Por otro lado, en lo relativo al ámbito de la enseñanza de idiomas, nos encontramos con un mayor número de corpus de aprendices cuyo contenido persigue

[6] Para más información sobre estas investigaciones y otras, se puede consultar el siguiente enlace: http://cedel2.learnercorpora.com/about/studies. En la sección bibliográfica final únicamente se incluirá la referencia a la página web de la que se han extraído los datos de estos estudios y artículos de investigación, no las publicaciones en sí (por tratarse estas solo de ejemplos ilustrativos), a menos que aparezcan citadas o mencionadas en alguna de las obras que forman parte de la bibliografía consultada para redactar este trabajo.

[7] Para más información sobre estas investigaciones y otras, se puede consultar el siguiente enlace: http://www.splloc.soton.ac.uk/publication.html. En la sección bibliográfica final únicamente se incluirá la referencia a la página web de la que se han extraído los datos de estos estudios y artículos de investigación, no las publicaciones en sí (por tratarse estas solo de ejemplos ilustrativos), a menos que aparezcan citadas o mencionadas en alguna de las obras que forman parte de la bibliografía consultada para redactar este trabajo.

fines pedagógicos, por lo que nos centraremos solo en algunos de ellos, aquellos que consideramos más importantes o de los que han nacido más investigaciones y/o publicaciones. Dentro de la modalidad escrita, quizá el corpus de aprendices que más ha destacado en los últimos años sea el del Instituto Cervantes, el CAES. La propia institución, en su apartado introductorio a este recurso, establece claramente su función y finalidad didácticas al señalar lo siguiente:

> Se trata de una herramienta que permite a los profesionales del campo de ELE (profesores, investigadores, evaluadores, autores de materiales didácticos, responsables y equipos de centros e instituciones lingüísticas, etc.) llevar a cabo investigaciones aplicadas sobre la base de datos sólidos y objetivos, ya que puede proporcionar información sobre dificultades de aprendizaje, errores más comunes, vocabulario más o menos empleado, etc., que se podrá aplicar con facilidad en las aulas o integrar en los manuales o materiales (Instituto Cervantes, 2020).

Lo cierto es que este corpus constituye un proyecto muy ambicioso y con muchas posibilidades, no solo por la calidad de su contenido, de su plataforma de consulta y de su proceso de compilación, selección, anotación, sino también porque incluye datos de informantes con diversas lenguas maternas, algunas de las cuales apenas han recibido atención en los estudios realizados a partir de o sobre corpus de aprendices de español. Hasta la fecha, el CAES no ha dado lugar a publicaciones o investigaciones concretas notorias, aunque ya se han presentado varias propuestas o ejemplos de aplicación. A modo de ejemplo, Ignacio Palacios Martínez (2018) sugiere un estudio sobre los fasos amigos en español para hablantes anglófonos, francófonos y lusófonos, así como un análisis exhaustivo del uso de las preposiciones más utilizadas en español por parte de estudiantes con diversas lenguas maternas para determinar cuáles suponen un mayor obstáculo. Por su parte, Claudia Parodi (2015 citado en Mas Álvarez & Gil Martínez, 2018) ha sugerido una investigación sobre el empleo del pronombre neutro *ello* por parte de aprendices de español. El CAES fue puesto a disposición del público a través de Internet en 2014, por lo que aún se encuentra en proceso de creación, difusión y crecimiento, pero podemos concluir desde ya que su potencial es innegable e inconmensurable; todavía no somos realmente conscientes de su verdadera y completa capacidad de aplicación. De hecho, la ambición de este proyecto no conoce límites, puesto que se pretenden aumentar las lenguas maternas y niveles de competencia que abarca, así como incluir también muestras orales. En cualquier caso, por el momento podemos asegurar ya que "[…] el CAES constituye una aportación que marcará un antes y un después en el terreno de la investigación de corpus de aprendices y la enseñanza de español como lengua extranjera" (Mas Álvarez & Gil Martínez, 2018, p. 51).

Aprescrilov y CAELE también son dos de los corpus de aprendices de español escritos que han aportado una gran cantidad de material y datos de calidad para el desarrollo de diversos estudios e investigaciones de corte lingüístico y/o pragmático. La producción científica generada a partir del contenido de estos corpus se enfoca al estudio de las diferencias entre *ser* y *estar*, los verbos de cambio y el uso de preposiciones en el caso de Aprescrilov (Buyse & González Melón, 2013; Buyse, Fernández Pereda & Verveckken, 2015 citados en Mas Álvarez & Gil Martínez, 2018), y, en lo referente al CAELE, al análisis de errores más frecuentes en las producciones escritas de los aprendices (a saber, los fallos de atribución de género y las faltas de ortografía por omisión de tilde) para elaborar una metodología de corrección apropiada y eficaz, adaptada a los diferentes niveles de dominio de la lengua extranjera (Ferreira Cabrera & Elejalde Gómez, 2017).

Otros corpus de aprendices de español de enfoque didáctico, en este caso compilados a partir de muestras orales, son el CORELE y el CORINÉI. El primero va acompañado incluso de una propuesta didáctica de explicación de errores a diversos grupos de aprendices y de formación en la corrección de fallos que pueden resultar de gran ayuda a docentes y profesionales de la enseñanza de ELE (Campillos Llanos, 2014 citado en Mas Álvarez & Gil Martínez, 2018). Algunos de los artículos o publicaciones más importantes relacionados con la investigación a través del CORELE son *Designing a search interface for a Spanish learner spoken corpus: the end-user's evaluation* (Campillos Llanos, 2012) y *A XML-tagged Spanish learner oral corpus for learner language research* (Campillos Llanos, 2011), así como la tesis doctoral del propio Leonardo Campillos Llanos (2012), titulada *La expresión oral en español lengua extranjera: interlengua y análisis de errores basado en corpus*[8]. En cuanto a CORINÉI, las investigaciones que ha fomentado se basan en el análisis de cuestiones y estructuras pragmáticas, fraseológicas, relacionadas con la cortesía, con los turnos de habla y con la dinámica de una conversación natural y relativamente fluida en general (Martín Sánchez, Pascual Escagedo & Puigdevall Bafaluy, 2017), con el objetivo de mejorar la práctica docente y proporcionar una base sólida y fiable para el desarrollo de materiales didácticos.

[8] Para más información sobre estas investigaciones, se puede consultar el siguiente enlace: http://www.lllf.uam.es/ESP/CORELE.html. En la sección bibliográfica final únicamente se incluirá la referencia a la página web de la que se han extraído los datos de estos estudios y artículos de investigación, no las publicaciones en sí (por tratarse estas solo de ejemplos ilustrativos), a menos que aparezcan citadas o mencionadas en alguna de las obras que forman parte de la bibliografía consultada para redactar este trabajo.

Entre los estudios y publicaciones más relevantes llevados a cabo a partir del CORINÉI, destacan *La problemática de la afinidad entre el español y el italiano en la enseñanza/aprendizaje desde la fraseología: el corpus de interlengua oral CORINEI* (González Royo, 2011), *La evaluación de la interacción oral: la conversación didáctica natico/no-nativo (aprendizaje colaborativo a distancia)* (Chiapello et al., 2012), *Skype y la interacción oral nativo/no-nativo: funciones y rutinas conversacionales en Corinei, un corpus de interlengua español e italiano* (González Royo, 2012), *Estrategias para la toma y cesión de los turnos de habla en la conversación de estudiantes italiano de E/Le y españoles de I/LE* (Pascual Escagedo, 2012), o *Prácticas docentes para el desarrollo de la competencia conversacional en ELE: teletándem* (Martín Sánchez & Pascual Escagedo, 2016)[9].

3.4. Ventajas y desventajas de la aplicación de corpus de aprendices a la didáctica del español (y otras lenguas)

A partir de todo lo expuesto en este capítulo y en el anterior, resulta evidente que el empleo de corpus en general y, más particularmente, de corpus de aprendices puede reportar grandes beneficios al ámbito de la investigación y la didáctica de lenguas extranjeras, en nuestro caso el español. Las principales ventajas que ofrecen estos recursos en la recopilación en una única base de datos de consulta de múltiples muestras contextualizadas, seleccionadas, analizadas y anotadas (muchas más y más variadas que en métodos de corte similar que se aplicaban antes de la aparición de los corpus) producidas en situaciones comunicativas reales, lo cual garantiza su objetividad, empirismo y autenticidad, y permite estudiar fenómenos lingüísticos y/o pragmáticos de manera eficiente y, sobre todo, imparcial. Asimismo, la posibilidad de acceder a los datos de forma rápida y eficaz en función de diversos parámetros de búsqueda y filtrado (gracias a la anotación y lematización de la información) acelera la obtención de dichos datos, lo cual economiza el tiempo invertido en su estudio. El hecho de que los corpus de

[9] Para más información sobre estas investigaciones y otras, se puede consultar el siguiente enlace: https://dti.ua.es/es/teletandem-corinei/publicaciones/publicaciones.html. En la sección bibliográfica final únicamente se incluirá la referencia a la página web de la que se han extraído los datos de estos estudios y artículos de investigación, no las publicaciones en sí (por tratarse estas solo de ejemplos ilustrativos), a menos que aparezcan citadas o mencionadas en alguna de las obras que forman parte de la bibliografía consultada para redactar este trabajo.

aprendices incluyan una mayor cantidad de contenido lingüístico de calidad, además, los convierte en recursos significativos y representativos de la lengua o variedad que se trate desde un punto de vista investigador y analítico. Por consiguiente, son una herramienta didáctica de gran utilidad que permite abordar aspectos que no siempre aparecen en los materiales de consulta e instrucción habituales, con lo cual se obtienen respuestas más claras, fiables y sólidas a muchas preguntas y obstáculos que surgen a lo largo del proceso de enseñanza y aprendizaje de una lengua extranjera (Hincapié, 2018). Por tanto, constituyen el complemento ideal para corroborar aquellos conocimientos procedentes de la intuición, la introspección, la experiencia e incluso la improvisación, sobre todo en el caso de profesores no nativos. En definitiva, el inmenso abanico de posibilidades que ofrecen los corpus de aprendices favorece "[…] la investigación y el análisis de producciones escritas reales de ELE y sus problemáticas lingüísticas, enriqueciendo así también el ámbito de la adquisición y enseñanza del español como lengua extranjera" (Ferreira Cabrera & Elejalde Gómez, 2017, p. 535).

Por otro lado, desde un punto de vista pedagógico más directo, la aplicación de los corpus de aprendices a la realidad del aula puede revelarse muy beneficiosa para los alumnos y fomentar su nivel de precisión, corrección y fluidez lingüísticas. La exposición a muestras genuinas del uso de la lengua meta constituyen una fuente de ejemplos e información muy valiosa en términos ilustrativos, creativos y de elaboración de materiales didácticos. En este sentido, según Vyatkina y Boulton (2017 citado en Abad Castelló, 2019, p. 149), los corpus "[…] proporcionan a los aprendices ejemplos de uso evidenciados, les ayudan a desarrollar destrezas analíticas y de resolución de problemas y promueven la autonomía en el aprendizaje". Esto permite adoptar y aplicar un enfoque más comunicativo a la enseñanza de un idioma, pues fomenta el aprendizaje autónomo e inductivo a través de una metodología contrastiva que prepare e instruya al alumno para identificar aquellos fenómenos de la segunda lengua que suelen ocasionar problemas recurrentes y errores frecuentes. Esta práctica, ya se produzca a través de la observación directa de los datos por parte de los estudiantes o a través de actividades confeccionadas por el docente, reduce considerablemente el nivel de interferencia procedente de la lengua materna (Buyse & González Melón, 2013), dado que son los propios aprendices los que examinan el uso de dichas estructuras problemáticas y extraen sus propias conclusiones al respecto, lo cual agiliza el proceso de aprendizaje y promueve el desarrollo de estrategias fundamentales de análisis y resolución de problemas lingüísticos relacionados

con la morfología, la sintaxis, el léxico, el orden de las palabras, las variedades diatópicas/diastráticas/diafásica del idioma, las connotaciones, etc.

Sin embargo, aunque resulta evidente que los corpus de aprendices pueden generar grandes beneficios y avances positivos en los ámbitos de la adquisición y de la enseñanza de segundas lenguas para la investigación y el desarrollo de materiales y metodologías didácticas útiles, lo cierto es que muchos docentes y diversos profesionales del campo de estudio de los idiomas extranjeros se muestran un tanto reticentes a la hora de implementar su uso en el aula o en la planificación didáctica. El argumento principal que esgrimen para defender su postura es que, como sucede con cualquier herramienta de apoyo a la investigación y/o enseñanza, los corpus de aprendices no son "la panacea". Tienen defectos e inconvenientes que es necesario considerar para hacer un uso realista, razonable y realmente eficiente de las posibilidades que ofrecen (Granger, 2009 citado en Núñez Nogueroles, 2019). Asimismo, existen otras razones por las que se relega el uso de corpus de aprendices a meras necesidades de consulta rápida, sobre todo relacionadas con la ausencia de familiaridad con estos recursos. Los programas de instrucción de profesionales dentro del mundo de la enseñanza de idiomas rara vez contemplan una formación técnica, lingüística y metodológica específica con respecto a los corpus de aprendices, su funcionamiento, la información que contienen y las herramientas de las que disponen, lo cual deriva en una carencia de conocimientos sobre sus posibilidades de explotación y su utilidad pedagógica. La escasez de materiales disponibles para desarrollar esa instrucción y habilidades tecnológicas por cuenta propia tampoco ayudan a que la situación formativa cambie. Por otra parte, el diseño de los corpus no siempre es todo lo intuitivo que debiera o pudiera ser, lo cual no hace sino acrecentar la necesidad de poseer una instrucción técnica concreta para utilizarlos y, por ende, fomentar ese desuso por parte de los docentes no familiarizados con estos recursos, que son la mayoría. Si sumamos esto a la dedicación y el tiempo que se ha de invertir para confeccionar materiales, actividades, pruebas, etc. a partir del contenido ofrecido por un corpus, podemos intuir a qué se debe esa reticencia a su aplicación en la realidad del aula. Por supuesto, estos docentes y profesionales de la enseñanza y adquisición de lenguas extranjeras no cuestionan la utilidad que pueden tener los corpus de aprendices, pero independientemente de que reconozcan su potencial y valor didácticos, "su utilización no parece haberse hecho realidad en la práctica diaria entre los docentes de español como lengua extranjera. Solo muy recientemente los manuales de ELE incluyen actividades concretas de uso de corpus y existen muy pocos materiales para usar en el aula" (Abad

Castelló, 2019, p. 152). Resulta evidente que esto constituye una realidad que hemos de modificar para fomentar el progreso tanto de la lingüística de corpus como de los corpus de aprendices en concreto y también de todos aquellos ámbitos relacionados con el estudio de las lenguas que pueden servirse positivamente de estos recursos.

3.5. Futuros pasos para los corpus de aprendices: ampliación de horizontes

Una vez considerados todos los beneficios, utilidades, ventajas e inconvenientes de los corpus de aprendices, y a modo de conclusión, podríamos decir que se trata de una metodología y de un recurso en auge actualmente, pero aún queda mucho camino que recorrer. Es cierto que las aportaciones realizadas hasta ahora por estas herramientas han sido muy valiosas e interesantes para el desarrollo de investigaciones, estudios, materiales, metodologías docentes, actividades, etc. Este es el motivo por el que su compilación no ha hecho sino aumentar en la última década, pero lo cierto es que actualmente seguimos contando con un número exiguo de corpus de aprendices de español, debido fundamentalmente a la dificultad que implica recolectar, filtrar y anotar la información que se desea que contengan. En cualquier caso, la versatilidad y las posibilidades de aplicación transversal o diagonal e los corpus de aprendices los convierten en una fuente de información y consulta provechosa para múltiples disciplinas, siempre y cuando se respeten cuidadosamente los requisitos metodológicos de recolección, selección, anotación y diseño en los que hemos hecho hincapié a lo largo de estas páginas. Siguiendo esta línea argumentativa, Mas Álvarez y Gil Martínez (2018) destacan que "la disponibilidad de corpus que reúnan estas características supondrá un fuerte espaldarazo para explotar de manera eficiente, y en múltiples direcciones, las colecciones de datos" (p. 51).

El futuro de los corpus de aprendices parece favorable y halagüeño, especialmente en los ámbitos de la adquisición y enseñanza de lenguas extranjeras, donde se están desarrollando y difundiendo en mayor cantidad y con mayor interés investigador debido a las posibilidades que ofrece para estudiar fenómenos lingüísticos y/o pragmáticos novedosos o ya examinados, pero esta vez aplicando nuevas metodologías y desde otras perspectivas. Sin embargo, para que el potencial de explotación de estos recursos alcance su punto álgido, es necesario que se produzca un proceso de perfeccionamiento para

mejorar aspectos relacionados con su compilación, diseño, funcionamiento, aplicación o difusión. Entre otros, destacamos los siguientes:

- Compilación de más corpus de carácter longitudinal. Como se puede comprobar en la tabla de clasificación de corpus de aprendices de español, la gran mayoría son transversales, lo cual dificulta un verdadero análisis significativo de la interlengua de los estudiantes y de su progreso a lo largo de diferentes niveles de instrucción. Es evidente que la recopilación de colecciones textuales de naturaleza longitudinal resulta mucho más compleja y requiere de una inversión de tiempo mucho mayor, pero consideramos que la investigación y la didáctica de lenguas extranjeras se verían ampliamente beneficiadas en términos de progreso. Además, gracias a los avances tecnológicos, existen cada vez más herramientas de anotación semiautomática que podrían facilitar y agilizar considerablemente el proceso.
- Creación de más corpus multimodales como LANGSNAP y de más corpus orales. Contar con muestras orales y escritas en una misma base de datos puede ahorrar mucho tiempo a la investigación y abrir nuevas puertas para análisis lingüísticos desde diversos ángulos quizá nunca antes explorados, por lo que serán más completos, objetivos, realistas e incluso revolucionarios. Por otro lado, el porcentaje de corpus de aprendices orales es mucho menor que el de base escrita, por lo que lo ideal sería desarrollar más recursos que se centraran en la producción e interacción oral de los aprendices de cara a su enseñanza, práctica y confección de materiales adecuados al respecto.
- Ampliación de las lenguas maternas de los aprendices y de las tareas que se les pide realizar. Cuanto mayor sea el rango lingüístico que abarquemos, mayor será el potencial de explotación investigadora y didáctica de estos recursos, y más se facilitará el acceso a la interlengua de determinados grupos de hablantes para trabajar sus dificultades concretas con materiales adaptados adecuadamente a sus necesidades.
- Ampliación de los tipos/géneros textuales de las tareas incluidas en los corpus. Particularmente, el foco de atención está comenzando a centrarse ahora en interacciones cibernéticas, dado que nos encontramos en la era de la informática y la comunicación virtual: correos electrónicos, mensajes instantáneos, videollamadas, publicaciones en redes sociales, blogs… (Mas Álvarez & Gil

Martínez, 2018). De este modo, se podrán analizar las narraciones, descripciones, argumentaciones y exposiciones que se realizan en este tipo de medios tan habitualmente utilizados en la actualidad, así como las características específicas que presenta la comunicación a través de estos canales. Esto puede dar lugar a un aprendizaje significativo del estudiante, ya que es muy probable que estos contenidos despierten mucho su interés y lo motiven a aprender.

- Desarrollo de corpus de aprendices para la enseñanza con fines específicos. Hay estudiantes que aprenden español con un objetivo muy concreto, normalmente laboral. Por tanto, la recolección de muestras en estos contextos formativos y su análisis podría resultar muy útil para ajustar los materiales y metodología docente a las necesidades muy concretas de estos grupos, así como a la presión de tiempo para aprender que normalmente suelen sufrir. De este modo, el profesor se asegura de ofrecer contenidos realmente útiles y no dedica sesiones a fenómenos o cuestiones que pueden ser innecesarios.
- Adición de información sociolingüística. Dada la multiplicidad de variedades diatópicas del español que existen, así como diastráticas, consideramos que sería interesante marcar o anotar determinados rasgos de alguna manera para facilitar el análisis de los datos y facilitar el proceso para determinar si resultan necesarios y/o válidos para nuestras investigaciones y estudios o no (Rodríguez Muñoz & Ruiz Domínguez, 2017). Algunos corpus ya aplican este tipo de etiquetas a su metodología, pero creemos conveniente que se extienda a todo el conjunto.
- Formación del profesorado en el empleo de corpus de aprendices. Es necesario que la instrucción de docentes de idiomas incluya un módulo en el que se aborden este tipo de recursos tecnológicos aplicables al ámbito de la investigación y creación de materiales en ELE. Si los profesionales de la didáctica no están familiarizados con estas valiosas fuentes de información y ejemplos, con su funcionamiento y con sus posibilidades de explotación, resultará imposible aprovechar todo el potencial de los corpus de aprendices y su progreso y difusión de estancarán, lo cual supondría una gran pérdida no solo para la enseñanza de idiomas, sino para la lingüística en general.

Si aplicamos estas buenas prácticas y las añadimos a las que ya hemos presentado en secciones anteriores, conseguiremos construir corpus de aprendices realmente representativos, no solo porque serán un reflejo fiel del uso de una lengua o variedad, sino

también en el sentido de que resultarán significativos para los profesionales de la investigación y la docencia en ELE. De hecho, la lingüista Margarita Alonso-Ramos (2016), de la Universidad de A Coruña, establece una serie de criterios para determinar cómo debe ser una investigación acerca de los corpus de aprendices de español para resultar significativa y representativa. Como se puede observar, sus conclusiones sobre la variedad que se debe abarcar para que un estudio realizado sobre o a partir de este tipo de recursos sea real u fielmente descriptivo y característico se ajustan y corresponden metodológicamente a las que nosotros hemos alcanzado en el presente trabajo con respecto a la creación de estos materiales de consulta.

1. La muestra debería cubrir diferentes modalidades de corpus: de lengua escrita y oral, así como corpus multimodales
2. Las muestras deberían abordar las producciones lingüísticas de aprendices de diferente procedencia, con diferentes lenguas maternas y con diferentes niveles de dominio de la segunda lengua
3. Los corpus descritos deberían estar disponibles en formato electrónico
4. La investigación debería explotar los datos contenidos en el corpus relativos a aspectos lingüísticos diversos: fonología, morfología, sintaxis, léxico y discurso
5. La investigación debería analizar los diversos enfoques multinivel que se aplican a los corpus de aprendices: anotación de partes de la oración, de errores y de fenómenos específicos, como colocaciones y anáforas
6. La investigación debería incluir estudios tanto transversales como longitudinales.
7. La investigación debería cubrir tres direcciones disciplinarias o metodológicas principales: la adquisición de segundas lenguas, la enseñanza-aprendizaje del español como lengua extranjera o como segunda lengua y la lingüística de corpus.[10] (Alonso-Ramos, 2016, pp. 18-19)

Solo construyendo corpus de aprendices que permitan llevar a cabo investigaciones de corte similar al expuesto por Alonso-Ramos podremos convencer a aquellos que aún se muestran reticentes a utilizarlos como herramienta de aprendizaje y conseguiremos que el progreso metodológico y el brillante futuro de aplicación pedagógica y científica de que estos recursos pueden gozar se conviertan en una realidad, abriendo así un inmenso horizonte de posibilidades a la investigación y la didáctica. Por el momento, ya se está produciendo un proceso de acercamiento e integración de la praxis basada en corpus en otros ámbitos distintos de la enseñanza de idiomas, como lo es el de la adquisición de una segunda lengua. En este sentido, Mas Álvarez y Gil Martínez (2018) insisten en la necesidad de que se produzca una fusión o, en todo caso, un "trasvase fructífero de las aportaciones de la investigación psicolingüística —en el ámbito de la adquisición de segundas lenguas— y las de la investigación en enseñanza de español

[10] Traducción propia.

como lengua extranjera" (p. 52). Son conscientes de los desafíos y obstáculos que ello conlleva, pero aseguran que tender puentes metodológicos para salvar las distancias entre el campo de la enseñanza de un idioma y el del análisis los mecanismos de adquisición y desarrollo psicológico de dicha lengua puede constituir la clave para progresar definitivamente hacia un enfoque pedagógico, programador, formativo e investigador eficiente, prometedor y con un gran potencial de explotación en el desarrollo de los corpus de aprendices de español.

4. Un ejemplo de aplicación práctica de los corpus de aprendices a partir de COWS-L2H: los pronombres personales sujeto en español

Con el fin de demostrar la utilidad investigadora y didáctica que pueden tener los corpus de aprendices que español, ambas complementarias y no excluyentes, procedemos ahora a mostrar los resultados y conclusiones obtenidas a partir de un estudio realizado con estudiantes angloparlantes y checoparlantes de la Universidad de California Davis y la Universidad de Bohemia del Sur, respectivamente. El objetivo fundamental de esta pequeña investigación es comparar los obstáculos y dificultades que aprendices de ambas lenguas maternas afrontan a la hora de utilizar (o no utilizar, más bien) el pronombre personal sujeto en español. Para ello, vamos a basarnos en los resultados extraídos a partir de un proyecto de colaboración en el que hemos participado a lo largo del año académico 2020/2021. La meta de dicho proyecto consiste en crear una base de datos de textos escritos corregidos y etiquetados contenidos en el *Corpus of Written Spanish of L2 and Heritage Speakers* (COWS-L2H), desarrollado por la Universidad de California Davis. Las producciones escritas contenidas en este corpus han sido redactadas por estudiantes angloparlantes que estudian español en la Universidad de California Davis y por hablantes de herencia. Nuestro cometido era corregir dichos textos de manera que fueran comprensibles y adecuados en español, no necesariamente "perfectos", y etiquetar los errores que contuvieran atendiendo a una serie de pautas. El fin que se persigue con este proceso es "entrenar" a un ordenador a partir de estos datos "editados" y anotados contenidos en el COWS-L2H para que él mismo pueda corregir producciones escritas y detectar errores. Ni qué decir tiene el avance metodológico y pedagógico que este

proyecto supone en términos de análisis de errores e identificación de problemas concretos de los alumnos angloparlantes a la hora de aprender español, así como las ventajas para el docente, pues le permite economizar el tiempo invertido en la corrección.

Dados el potencial investigador y la utilidad didáctica que consideramos que supone el enfoque de este proyecto, decidimos extrapolarlo a otro contexto formativo, con diferentes aprendices y una lengua materna distinta, con el propósito de comprobar el grado de explotación pedagógica que podría alcanzarse con este tipo de planteamientos científico-analíticos. Por ello, hemos llevado a cabo un modesto estudio metodológicamente paralelo al desarrollado por la Universidad de California Davies con el COWS-L2H en la Universidad de Bohemia del Sur (República Checa), en este caso con estudiantes checoparlantes de diferentes niveles de dominio del idioma. Cabe destacar que hemos tratado de emular en todo lo posible, dadas las circunstancias, las condiciones procedimentales de recopilación y análisis de muestras del planteamiento estadounidense, con el fin de obtener resultados comparables que nos permitan extraer conclusiones relevantes e interesantes desde un punto de vista contrastivo. No obstante, la situación pandémica ha afectado considerablemente al desarrollo de nuestra actividad investigadora y nos hemos visto obligados a diversificar y adaptar nuestro planteamiento inicial en determinados aspectos que se irán concretando a medida que se introduzcan y expliquen en las próximas páginas. En cualquier caso, consideramos que el resultado final de nuestro modesto estudio ha sido satisfactorio y nos ha permitido alcanzar una serie de conclusiones contrastadas muy enriquecedoras y prometedoras en lo relativo al ámbito de la didáctica de idiomas.

Presentaremos a continuación, por tanto, nuestro proyecto de investigación. En primer lugar, describiremos el corpus de aprendices COWS-L2H y nuestra participación en él para contextualizar el marco en el que se insertará después el estudio "paralelo" que hemos realizado con los alumnos checoparlantes. Posteriormente expondremos el tema gramatical que nos atañe y sobre el que gira nuestro estudio: el uso y omisión del pronombre personal sujeto en español. Será entonces cuando pasaremos a describir nuestra investigación, los participantes y la metodología aplicada (incluidas las distintas fases de las que consta y el enfoque adoptado para analizar los datos). Por último, presentaremos los resultados y expondremos las conclusiones que hemos alcanzado con este estudio comparativo.

4.1. El corpus de aprendices COWS-L2H y nuestra participación

Como explicamos al final del apartado anterior, una de las mejoras requeridas para perfeccionar el compendio de corpus de aprendices de español que sugerimos como futuro paso recomendable en el desarrollo de este tipo de recursos es la necesidad de compilar más corpus de carácter longitudinal. Es cierto que ya contamos con algunos, como Aprescrilov o LANGSNAP. Sin embargo,

> aunque Aprescrilov contiene datos longitudinales, estos se limitan a un período de tiempo muy reducido (un cuatrimestre académico), y recopila datos de hablantes que tienen el neerlandés y el francés como lenguas maternas, y aprenden el español como tercera lengua. Por consiguiente, puede que no resulte especialmente útil para aquellos interesados en aprendices de español angloparlantes. Por otro lado, el corpus LANGSNAP, compila datos longitudinales a relativamente largo plazo de estudiantes de español como segunda lengua, pero se limita a un número reducido de participantes.[11] (Yamada et al., 2020, p. 21).

Por tanto, pese a que contamos con corpus de aprendices longitudinales de español, es evidente que precisamos de más recursos de este tipo, pero de mayor envergadura y con datos extraídos de estudiantes de español como segunda lengua o primera lengua extranjera. Un posible contraargumento a esta afirmación sería el CAES, un corpus longitudinal con un volumen considerable de muestras. Sin embargo, más allá de su novedad y de que todavía no se han publicado estudios de corte didáctico o lingüístico-investigador concretos a partir de la información que contiene, el CAES constituye una base de consulta de naturaleza muy variada y diversa, debido a que se ha construido a partir de la compilación de datos extraídos de las producciones escritas de estudiantes de diversas filiales del Instituto Cervantes. Esta heterogeneidad de la información dificulta estudios que se quieran llevar a cabo dentro de un marco contextual concreto: una situación formativa específica, un grupo muy particular de hablantes, etc. En consecuencia, podemos concluir que no solo existe una necesidad de corpus longitudinales considerablemente amplio, sino que, a su vez, muestren un notable grado de homogeneidad entre sus datos e informantes (Yamada et al., 2020).

Por otra parte, los corpus de aprendices de español suelen centrarse en aquellos estudiantes que han comenzado a estudiar la lengua de cero en diversos contextos formativos, pero existe un colectivo de informantes que se ha desatendido casi sistemáticamente en la creación de este tipo de herramientas: los hablantes de herencia. De hecho, como indican los responsables del desarrollo del corpus COWS-L2H, “la

[11] Traducción propia.

mayor parte de la investigación en hablantes de herencia se ha centrado en el análisis de diferencias entre estos y los hablantes nativos de español o los aprendices de español como segunda lengua" (Yamada et al., 2020, p. 21). Por tanto, no se han realizado estudios empíricos sobre el desarrollo de la interlengua de este tipo concreto de aprendices y sus características, necesidades y problemas particulares de aprendizaje, pese a que existan asignaturas específicas para ellos en algunas universidades estadounidenses. Precisamente por ello, el corpus COW-L2H resulta tan significativo y novedoso.

El COWS-L2H (*Corpus of Written Spanish of L2 and Heritage Speakers*) es un corpus escrito de aprendices desarrollado por un equipo de la Universidad de California Davis dirigido por Claudia H. Sánchez Gutiérrez. Los principales objetivos que persigue este proyecto lingüístico están enfocados a cubrir las deficiencias y carencias del conjunto de corpus de aprendices de español desarrollados hasta ahora. En este sentido, pretende aportar novedades y beneficios a la investigación en ELE proporcionando una base de datos de aprendices de español longitudinal, extensa, homogénea y representativa de la interlengua y su evolución tanto en estudiantes de español como segunda lengua como en hablantes de herencia dentro de un contexto formativo específico (educación superior norteamericana en una única institución universitaria). Un corpus de este estilo facilitará el análisis de la competencia léxica y gramatical, y su evolución a lo largo de diferentes niveles, de aprendices que comparten un mismo programa y entorno de instrucción (Yamada et al., 2020), lo cual permitirá, a su vez, determinar los avances en el dominio de las lenguas extranjeras que se producen a partir de los currículos aplicados a la enseñanza de idiomas en esta universidad de EE.UU. De este modo, se podrán extraer conclusiones generales y útiles para desarrollar una metodología docente más efectiva y confeccionar materiales didácticos acordes a las necesidades y características de los aprendices de español que aprenden el idioma en esta situación académica concreta y siguiendo una programación pedagógica específica. Dado el creciente multilingüismo de EE.UU., recursos como el COWS-L2H pueden revelarse una pieza clave en la investigación lingüística nacional e internacional.

Para garantizar una recopilación longitudinal y representativa de datos, se fomenta la participación de múltiples informantes en diversos momentos a lo largo de sus estudios universitarios (en varios cuatrimestres consecutivos), lo cual permite aplicar tanto un enfoque transversal como longitudinal al análisis de la información compilada (Yamada et al., 2020). Hasta la fecha, se ha recogido un total de 3 498 redacciones (887 418 palabras) a lo largo de ocho cuatrimestres (desde 2017 hasta hoy) y se pretende continuar

con el proyecto al menos otros cinco años (Yamada et al., 2020). La temática de las producciones escritas gira en torno a cuatro ejes (títulos) diferentes: *una persona famosa* y *unas vacaciones perfectas* por un lado (textos compilados en la primera fase, entre 2017 y 2018), y *una persona especial en tu vida* y *una anécdota terrible* por otro (textos recopilados en la segunda fase, de 2018 hasta la actualidad). Como se puede ver, se trata de temas considerablemente generales, simples e indefinidos, lo cual fomenta la creatividad y el uso de múltiples mecanismos lingüísticos (Yamada et al. 2020).

No cabe duda, por tanto, del interés que suscita y el progreso que puede suponer un corpus de aprendices de estas características. Es evidente que el COWS-L2H ocupará un lugar preeminente como base de datos de referencia en el ámbito de la investigación en ELE a través de estos recursos lingüísticos de consulta y análisis. Cabe esperar una considerable producción científica generada a partir de este proyecto, pero aún queda mucho por descubrir respecto a su potencial de explotación y sus prometedoras aplicaciones. Por el momento, el COWS-L2H ya se encuentra disponible en línea, abierto al público en formato TXT, y su contenido se puede descargar para utilizarlo y analizarlo en diversos estudios[12].

Uno de los avances que se pretende desarrollar dentro del marco de este proyecto es crear una base de datos que contenga las producciones escritas recopiladas en el COWS-L2H con correcciones y anotaciones de los errores que presenten. El objetivo fundamental de este subproyecto es doble. En primer lugar, su intención es conseguir que un ordenador pueda realizar procesos de corrección de manera automática y lo más autónoma posible (aunque siempre necesite un cierto grado de supervisión) a partir de las correcciones realizadas por colaboradores del ámbito de la enseñanza de español como lengua extranjera, como ya se ha indicado al inicio de este capítulo. En segundo lugar, pretende facilitar el estudio de patrones de evolución en errores concretos cometidos por los aprendices en sus composiciones, para lo cual requiere también que se etiqueten algunos de los fallos detectados en dichos textos.

En esta ramificación del proyecto general es donde entra en juego nuestra participación. Junto con un equipo de correctoras de la Universidad de Salamanca, hemos formado parte del desarrollo de esta base de datos editados y anotados, que aún está en su etapa inicial de creación. Nuestra misión ha consistido en corregir varios conjuntos de textos debidamente anonimizados y de diferentes niveles (en ningún momento se nos ha

[12] Para acceder a la información contenida en este corpus, se puede visitar el siguiente enlace: https://github.com/ucdaviscl/cowsl2h.

indicado a qué nivel pertenece cada uno de ellos), cuyo contenido giraba en torno a los dos primeros temas de los cuatro principales expuestos hace algunos párrafos: *una persona famosa*, en su mayor parte, y también algunas composiciones redactadas en torno a *unas vacaciones perfectas*. Por tanto, nuestro cometido se ha centrado en las producciones escritas compiladas durante la primera fase de recopilación (2017-2018).

Como se puede deducir del doble propósito o función que se persigue con este subproyecto, hemos tenido que realizar dos tareas complementarias, pero no necesariamente interdependientes: corregir los textos, por un lado, y posteriormente etiquetar aquellos errores indicados en las pautas de anotación que habíamos de respetar y aplicar. Por tanto, no debíamos etiquetar todos y cada uno de los fallos detectados en las producciones escritas de los aprendices, solo un conjunto muy reducido y muy concreto. Con respecto a la primera fase de nuestra labor, teníamos que tener en cuenta que solo debíamos corregir fallos ortográficos, gramaticales y léxicos claros y/o graves. Nuestro cometido no consistía, por tanto, en intervenir en el registro o estilo apreciados en las composiciones escritas; bastaba con que modificáramos lo mínimo imprescindible para que el texto fuera comprensible y coherente, aunque su corrección no se ajustara meticulosamente a la(s) norma(s) del español.

En cuanto a la segunda parte del proceso de revisión de los textos, centrada en la anotación de los errores, solo teníamos que etiquetar cinco tipos de errores, a saber:

1. Errores de concordancia de género y/o número (*unos casas grande*)
2. Errores de atribución indebida de género y/o número (*el vacación perfecto*)
3. Errores de presencia o ausencia indebida de pronombres de sujeto o de artículos (*yo me llamo Pedro y yo soy el profesor*)
4. Errores por confusión entre las preposiciones *por* y *para*, o entre los verbos *ser* y *estar* (*han llegado estas flores por ti / soy preparada para el viaje*)
5. Errores en la colocación de los adjetivos (*mi preferido color*)

Por tanto, si a lo largo de la fase de corrección detectábamos más fallos de tipología diversa a estos cinco grupos, debíamos corregirlos, pero no etiquetarlos después en la segunda parte de la revisión textual, por lo que el documento anotado contendría las etiquetas pertinentes según los errores que se debían marcar, pero el resto del texto permanecería como en su versión original, sin correcciones.

Durante el curso académico 2019/2020, nuestro equipo de revisoras ha realizado cinco tareas de corrección y etiquetado de producciones escritas, con paquetes de textos de distinto tamaño distribuidos en diferentes momentos del año. Huelga decir que se ha tratado de una experiencia muy enriquecedora a nivel profesional, ya que hemos podido comprobar de primera mano la versatilidad de las posibilidades de explotación que ofrecen los corpus de aprendices de español y hemos podido analizar aspectos muy interesantes relacionados con las características de redacción y la interlengua de estudiantes angloparlantes de diferentes niveles y contexto socio-personales dentro de un contexto formativo reglado concreto. A partir de nuestro trabajo no solo se podrá "entrenar" a una máquina para que sea capaz de corregir producciones escritas de manera automática y autosuficiente, sino que además se pueden extraer una serie de datos relevantes, prometedores y muy interesantes desde un punto de vista investigador. En este sentido, es probable que del tratamiento y observación de estos datos surjan diversos estudios que amplíen la producción científica sobre la adquisición y enseñanza del español como lengua extranjera. Nosotros hemos querido hacer una pequeña demostración del potencial científico de nuestra labor, de este subproyecto relacionado con el COWS-L2H y de este corpus en general elaborando una pequeña investigación sobre uno de los errores corregidos y etiquetados a lo largo de nuestras revisiones: la presencia o ausencia indebida de pronombres de sujeto. En este sentido, hemos querido también comprobar si los resultados y conclusiones extraídos a partir del análisis de textos redactados por angloparlantes serían extrapolables y aplicables a un contexto formativo similar, pero con informantes de otra lengua materna: checoparlantes. De este modo, hemos desarrollado una suerte de estudio contrastivo español-checo (y, hasta cierto punto, español-checo-inglés, aunque el tercero tenga un papel más bien de "intermediario"), centrándonos sobre todo en este último, para verificar no solo la utilidad de los corpus de aprendices, sino también su papel como fuente de inspiración e impulso creativo para otros contextos lingüísticos distintos a aquellos en los que fueron concebidos.

4.2. Pronombres personales sujeto en español y sujetos tácitos

Los pronombres personales sujeto constituyen un fenómeno gramatical problemático a la hora de enseñar y aprender español, ya que en este idioma la práctica habitual consiste en omitirlos, aunque sigan presentes de manera tácita, reflejados en la

flexión verbal. Este hábito de elisión gramatical supone un foco de conflicto para muchos estudiantes de español como lengua extranjera, ya que en sus idiomas nativos se siguen procedimientos diferentes. Por ello, procederemos a analizar a continuación el funcionamiento y características de este fenómeno gramatical en español, con el objetivo de comprender y determinar de manera más informada, rigurosa y adecuada las dificultades que puede entrañar para los aprendices de nuestra lengua. Según establece la *Nueva gramática de la lengua española*,

> Los pronombres personales se caracterizan por designar a los participantes en el discurso, sean estos quienes fueren. Esta propiedad los desprovee en cierta medida de contenido propiamente léxico y los convierte en categorías DEÍCTICAS (§ 17.1). […] Los pronombres personales son, además, elementos DEFINIDOS, propiedad que comparten con los artículos determinados y con los nombres propios (Real Academia Española & Asociación de Academias de la Lengua Española, 2009-2011, pp. 189-190)

Estas unidades lingüísticas pueden clasificarse de manera diferente dependiendo del criterio que apliquemos en cada caso para su categorización: persona, número, tonicidad, reflexión, caso… El conjunto de estos atributos flexivos y morfológicos de cada pronombre es lo que los caracteriza e identifica como un elemento lingüístico único, distintivo y significativo desde un punto de vista semántico. Dos de estos rasgos clasificadores nos interesan especialmente en el presente trabajo, dada la función específica (de entre todas las que pueden desempeñar los pronombres personales) que pretendemos analizar: su papel como sujeto. Estas características identificativas son la persona y la tonicidad. Atendiendo a las personas del discurso, que son las que establecen los participantes implicados en el evento comunicativo (el que emite, el que recibe y aquello en torno a lo que gira el acto verbal), y de acuerdo con el contenido de la *Nueva gramática* (2009-2011, p. 192), podemos categorizar los pronombres personales de la siguiente manera:

- **Pronombres de primera persona**: *yo, mí, me, conmigo, nosotros, nosotras, nos*.
- **Pronombres de segunda persona**: *tú, vos, ti, te, contigo, vosotros, vosotras, os, usted, ustedes*.
- **Pronombres de tercera persona**: *él, ellos, ella, ellas, ello, le, les, la, las, lo, los, se, sí, consigo*

Por lo que respecta a la tonicidad (p. 212), existen dos grupos de pronombres personales, a saber:

- **Tónicos**: *yo, tú, vos, usted, él, ella, ello, nosotros, nosotras, vosotros, vosotras, ustedes, ellos, ellas, mí, ti, sí, conmigo, contigo, consigo*.
- **Átonos**: *me, te, se, le, lo, la, nos, os, les, los, las*.

A este respecto, cabe destacar la controversia que se ha generado en los últimos años en torno a los pronombres tónicos de primera y segunda persona, ya que, pese a su denominación, lo cierto es que no constituyen elementos sustitutivos de sus referentes en el acto verbal, ya que siempre identifican y determinan, respectivamente, al participante que comunica una información (emisor) y al que la recibe (receptor). Por tanto, no tienen un referente discursivo propiamente dicho, sino que constituyen unidades deícticas, y de ahí que en algunas tradiciones y teorías gramaticales se les conozca como "nombres personales" (Real Academia Española & Asociación de Academias de la Lengua Española, 2009-2011, p. 223).

En cualquier caso, volviendo a la cuestión que nos atañe en este trabajo, de todos los pronombres personales expuestos según el criterio de persona, así como el de personalidad, los que pueden funcionar sintáctica y semánticamente como sujeto son los pronombres de primera, segunda y tercera persona que, además, son tónicos, a excepción de *mí, ti, sí, conmigo, contigo, consigo*. Por tanto, serían *yo, tú, él, ella, ello, nosotros, nosotras, vosotros, vosotras, ellos, ellas, usted, ustedes y vos* (este último solo en aquellas zonas hispanoamericanas en las que se aplica el fenómeno del voseo). ¿Cómo sabemos si estos pronombres actúan como sujetos o como otro tipo de argumento verbal? En función de tres marcas formales-gramaticales básicas: la concordancia con el verbo, el caso (solo en los pronombres *yo* y *tú*, que se corresponden al caso nominativo o recto) y la posición sintáctica (Real Academia Española & Asociación de Academias de la Lengua Española, 2009-2011, p. 2528). No obstante, este último no es definitivo ni concluyente, ya que únicamente resulta útil en el caso de que existan varios sintagmas nominales que puedan desempeñar la función de sujeto porque ambos concuerdan con el verbo, mas ni siquiera entonces constituye un criterio totalmente taxativo y fiable en el que basarse, ya que el sujeto puede ubicarse en diferentes posiciones sintácticas, como consecuencia de la riqueza flexiva del español y en función de la intención estilística que tenga el emisor, del efecto que pretenda causar (normalmente la posición del sujeto puede relacionarse con un propósito enfático).

Existen distintas ópticas desde las que catalogar los sujetos en español, de las cuales nos vamos a centrar en dos: la clasificación atendiendo a las categorías

gramaticales y la clasificación atendiendo a su presencia o ausencia en la oración (Real Academia Española & Asociación de Academias de la Lengua Española, 2009-2011, p. 2529). De acuerdo con lo establecido en la primera, los tipos de palabras o grupos de palabras que pueden funcionar como sujetos son los sustantivos, los pronombres, los sintagmas nominales y pronominales, y finalmente las oraciones subordinadas sustantivas. Según la segunda clasificación que nos concierne, los sujetos pueden ser explícitos, que aparecen de manera expresa en la oración, o implícitos, ausentes, en cuyo caso se conocen como sujetos tácitos y se tienden a entender como variantes de los pronombres personales (Real Academia Española & Asociación de Academias de la Lengua Española, 2009-2011, p. 2530). La posibilidad de omitir los sujetos en español, como hemos explicado, está directamente relacionada con la riqueza en la flexión verbal romance, pues la concordancia con el verbo y los rasgos de número y persona que este predicado muestra lo vinculan siempre a un sujeto, incluso cuando este último no queda patente en el discurso porque está elidido o porque, directamente, no existe (como en las oraciones impersonales). La razón que permite que se produzca este fenómeno incluso con sujetos tácitos son las propiedades morfológicas de carácter mayoritariamente pronominal que presentan esos sujetos nulos, lo cual implica la activación del proceso de concordancia en persona y número con el verbo.

En cualquier caso, para que la comunicación sea efectiva, no resulta apropiado elidir un sujeto desde el primer momento. Cabría preguntarse, entonces, por qué resulta necesario introducir siempre un sujeto explícito antes de recurrir a la variante tácita. La explicación es muy sencilla. Los sujetos nulos suelen cumplir una función temática en la oración, es decir, designan una realidad sobre la que se aporta una información nueva y relevante (rema), y sobre la que se supone un conocimiento previo. Por tanto, requieren de un vínculo con una entidad introducida anteriormente en el discurso para poder construir un hilo comunicativo coherente y eficiente: un sujeto explícito[13]. Una vez que se haya establecido un elemento expreso como antecedente, se establecen relaciones lógicas de correferencia, reflejada en los afijos verbales, que permiten, además, que los sujetos nulos se sucedan, ya que el primero puede convertirse en referente de otros elementos anafóricos o catafóricos (no solo sujetos elididos, sino también pronombres

[13] Esto no se aplica en el caso de que el sujeto elidido haga referencia a elementos extraídos del contexto deíctico del discurso, es decir, al emisor (primera persona) y al receptor (segunda persona), en cuyo caso el sujeto tácito encontrará su referente en la realidad extralingüística que rodea el acto comunicativo y no se tratará de un elemento presente directamente en el discurso.

relativos a él) debido a su coindexación con el antecedente explícito inicial. Ese antecedente que actúa como referente de los sujetos tácitos puede ser de naturaleza múltiple, pero las categorías más frecuentes que desempeñan esta función son los sustantivos y los pronombres. De ahí que la sustitución más "lógica" y habitual de un sujeto no expreso se base en recurrir a un nombre o a un pronombre, sobre todo a este último (de hecho, las lenguas que no tienen equivalentes para los sujetos tácitos recurren prácticamente siempre a los pronombres personales).

Por consiguiente, siguiendo esta línea argumentativa, cabe destacar y hacer especial hincapié en el hecho de que, aunque no tengan una manifestación fonética o gráfica, los sujetos tácitos forman parte de la oración, la condicionan y se ven condicionados por ella y por el contexto extralingüístico que la rodea. Esto se aprecia en que tienden a constituir elementos deícticos, en el caso de que "equivalgan" a pronombres de primera y segunda persona, o anafóricos, en el caso de que "equivalgan" a pronombres de la tercera. Con respecto a este último caso, la lingüista Marta Luján (1999) indica que "[…] un argumento nominal tácito no sólo [sic] se entiende como una forma del pronombre, funciona también como tal pues toma la referencia de un antecedente estructural" (p. 1292). Esto corrobora que el sujeto elidido, en caso de poseer rasgos de tercera persona, suele remitir a una entidad con función temática que ha aparecido anteriormente o que aparecerá posteriormente en el discurso, y con la que establece una correferencia textual. Por tanto, la interpretación más razonable y definida de los sujetos nulos en estos casos es la referencial: vincular el elemento omitido al sintagma nominal introducido previamente/ulteriormente en el hilo discursivo con el que concuerde en número y persona. Su omisión constituye el proceso más apropiado en términos de economía del lenguaje, puesto que el tema discursivo ya ha sido introducido y lo que nos interesa a partir de ese momento es la información remática o novedosa que se pueda ofrecer sobre él, lo que convierte la explicitud del sujeto en una cuestión innecesaria y superflua.

Sin embargo, es importante tener en cuenta que el concepto de sujeto tácito elidido genera cierta controversia, ya que resulta complejo 1) determinar cómo un elemento ausente puede desempeñar una función tan fundamental como la de sujeto y equipararse a otras formas explícitas, como sustantivos y pronombres, y 2) justificar por qué existen ciertos tipos de oraciones en las que ese sujeto omitido no puede sustituirse de manera natural por un pronombre, como sucede en la mayoría; esto suele ocurrir en aquellos enunciados en los que el pronombre tiene rasgos de tercera persona y se refiere a

realidades "no humanas" (*Tu mensaje lo hizo reflexionar, aunque era corto = ¿Tu mensaje lo hizo reflexionar, aunque* ***él*** *era corto?*) (Real Academia Española & Asociación de Academias de la Lengua Española, 2009-2011, p. 2549). Para zanjar este tema de manera breve, hemos de asumir que los pronombres de tercera persona, pese a tratarse de realizaciones fonéticas plenas, tangibles y expresas, también tienen sus restricciones sintácticas de aplicación o manifestación que los limitan a referentes "animados", de ahí que la sustitución de su versión omitida por su uso explícito pueda resultar forzada en determinados contextos, pero eso no implica que no exista una relación referente-anáfora o catáfora-referente entre ellos. Por otra parte, del mismo modo que un sujeto elidido asume las mismas "barreras sintácticas" que los pronombres, sustantivos, etc. con los que establece una correferencia, también recibe su carga semántica, lo cual le permite actuar como sujeto independientemente de su realización fonética expresa.

La conclusión que extraemos de los últimos párrafos es que en español los sujetos tácitos constituyen elementos deícticos o anafóricos que suelen corresponderse en la mayoría de los casos con pronombres de primera, segunda o tercera persona (sus equivalentes más naturales en las lenguas que no pueden omitir sus sujetos), por lo que pueden sustituirse por ellos. Pero, ¿su interpretación es siempre igual de específica? ¿Siempre aluden con precisión a los participantes en el acto comunicativo o a aquello de lo que hablan? Lo cierto es que los sujetos nulos siempre muestran rasgos deícticos o anafóricos, pero su apreciación puede no ser siempre concreta. Existen oraciones en las que no se identifica con exactitud al referente de los sujetos elididos, en las que su interpretación suele ser más indefinida o genérica, como es el caso de las impersonales reflejas o impersonales con verbos en tercera persona del plural.

> *Se vive bien en España* (= "Todo el mundo vive bien en España" → interpretación general).
>
> *Se llamó a la policía para que interviniera* (= "Alguien [una persona en concreto] llamó a la policía para que interviniera" → interpretación inespecífica, pero restringida).
>
> *En España están siempre de fiesta* (= "Todo el mundo en España está siempre de fiesta" → interpretación genérica).
>
> *Han llamado esta mañana* (= "Alguien [una persona en concreto] ha llamado esta mañana" → interpretación indefinida, pero restringida).

Asimismo, por otro lado, existen verbos en los que ese sujeto tácito no tiene una interpretación ni definida ni indefinida o genérica, puesto que no constituye un argumento

seleccionado por el verbo por el mero hecho de que hay verbos que, aunque parece que muestran una concordancia morfológica con un potencial sujeto, no admiten sujetos, como los de fenómenos atmosféricos. Los propios lingüistas no se ponen de acuerdo en este tema: unos dicen que, en estos casos, existiría un sujeto tácito no argumental, como sucede con los pronombres pleonásticos del francés y el inglés, por ejemplo, mientras que otros rechazan esta teoría y consideran que estos verbos presentan rasgos de tercera persona del singular simplemente porque se trata de la forma de concordancia no marcada en español (Real Academia Española & Asociación de Academias de la Lengua Española, 2009-2011, p. 2554). Esta dicotomía interpretativa no resulta fundamental para los objetivos que persigue el presente trabajo, pero consideramos importante recalcarla y comentarla como prueba de la inestabilidad del concepto de "sujeto tácito" en español y de la controversia que ya hemos indicado que suscita.

En cualquier caso, lo relevante para nuestro estudio no son las interpretaciones indefinidas, genéricas o polémicas de los sujetos tácitos, pese a que haya que tenerlas en cuenta para comprender la naturaleza de este fenómeno en español, sino precisamente su lectura como equivalentes definidos de los pronombres personales que pueden ser sustituidos por ellos, ya que en la mayoría de los casos se corresponden con esta categoría gramatical. No obstante, aparte de la definición, existe otro rasgo propio de estos sujetos tácitos nacidos de la omisión de un pronombre personal que tenemos que tener en cuenta: su carácter no contrastivo, derivado de su no explicitud, de su ausencia de realización fonética. Por el contrario, si el pronombre al que equivale ese sujeto nulo se materializa de manera expresa, adquiere normalmente propiedades contrastivas. ¿Por qué sucede esto? Vamos a tratar de explicarlo a continuación a partir de las diversas razones que pueden motivar el uso explícito de un pronombre personal en lugar se su versión elidida.

En realidad, el uso expreso del pronombre personal sujeto, según parece haber acordado la mayoría de la comunidad de lingüistas españoles, nace de dos necesidades expresivas interconectadas: la focalización contrastiva y el énfasis. En aquellos contextos en los que resulte necesario establecer una comparación o una contraposición, es necesario utilizar un pronombre expreso como término de dicho contraste con un propósito distintivo y enfático. Por ejemplo, ante la pregunta *¿Lucía ha fregado los platos?*, en caso de que no sea Lucía la que lo haya hecho, sino el receptor de dicha pregunta, la respuesta que seguiría de manera más natural sería *No, los he fregado YO /*

YO los he fregado[14], mientras que resultaría poco idiomático y confuso contestar simplemente *No, los he fregado*. Esto se debe a que se produce una situación de contraste entre dos términos que necesitan materializarse para que dicha contraposición se enfatice, se haga así realmente patente y no quede incompleta, dificultando la comunicación, ya que "si el pronombre tónico tiene función contrastiva, no debe reemplazarse por la forma tácita o inacentuada, que no tiene tal función" (Luján, 1999, p. 1301).

Asimismo, hay ocasiones en las que la presencia expresa del pronombre personal con valor contrastivo puede utilizarse, además de para enfatizar una confrontación, para tratar de establecer una diferencia con un posible referente con el que, en caso de que utilizáramos el sujeto tácito, podría producirse una correferencia de manera más evidente o directa. Para ilustrar este uso, veamos el siguiente ejemplo:

*Eso sí, si **él** bebe en sus ratos libre, Pedro no dice nada.*
Eso sí, si bebe en sus ratos libres, Pedro no dice nada.

Como se puede apreciar, la materialización de pronombre personal parece contener un matiz diferenciador, derivado de su naturaleza contrastiva, que implica que la persona que bebe y Pedro no son la misma, mientras que, si lo omitimos, es posible entender tanto lo anterior como que el que bebe es Pedro. De hecho, esta segunda lectura de la oración sería lo más habitual y natural, dado el carácter no contrastivo de los sujetos nulos. No obstante, todo depende de la interpretación que se haga del enunciado y del contexto en que se enmarque. Del mismo modo, el uso enfático del pronombre personal sujeto puede no tener un valor estrictamente contrastivo y, en tal caso, cabría la posibilidad de interpretar que el pronombre *él* alude a Pedro y se emplea simplemente por razones enfáticas. Es decir, puede responder a motivos de mero realce, sin pretensiones de confrontación. Para mayor ilustración y comprensión de este fenómeno, fijémonos en los siguientes ejemplos:

Mi hermana me juró que ELLA vendría a la boda.
Pedro dice que ÉL no ha sido.[15]

La presencia del pronombre tiene un claro valor enfático en estos casos, pero no tiene por qué interpretarse de manera contrastiva. Es posible que se contraponga el sujeto

[14] Mayúscula para destacar el elemento enfático.

[15] Ibidem.

a un grupo de potenciales sujetos que no han actuado del mismo modo, en el primer caso, o que sí han hecho aquello de lo que se acusa al sujeto en cuestión, en el segundo caso. Sin embargo, también puede interpretarse que el antecedente de esos pronombres son los sintagmas nominales precedentes y, por tanto, se establecería una correferencia similar a la que se da entre un referente y un sujeto tácito, con un valor enfático pero no contrastivo en esta ocasión.

En cualquier caso, resulta evidente que, aunque sea lo más habitual, la versión elidida de un pronombre no siempre constituye la opción más idiomática y correcta en español; depende del contexto en que dicha omisión se produzca. En este sentido, a modo de conclusión, podemos decir que

> Los contextos que requieren el uso de una forma léxica contrastiva (o acentuada) del pronombre atestiguan su función complementaria en relación con la forma nula (o inacentuada). En esos contextos la omisión del pronombre es claramente anómala o disonante, y los afijos de concordancia que legitiman la forma nula no son suficientes para denotar un contraste con un término dado (Luján, 1999, p. 1303)

No obstante, esos contextos que justifican el uso del pronombre expreso no siempre responden a razones enfáticas o antitéticas. Más allá de la función focalizadora y de realce en situaciones contrastivas o de identidad inexacta, existen otros argumentos y pretextos que motivan la expresión explícita del pronombre personal sujeto, aunque no resulte necesario para facilitar la interpretación correcta de una oración. Según la *Nueva gramática de la lengua española* (2009-2011, p. 2557), algunas de estas razones pueden ser de tipo morfológico, léxico o estilístico. Como ejemplo de las primeras, el imperfecto de indicativo tiene la misma forma para la primera y la tercera persona del singular, por lo que puede ser necesario aclarar el sujeto. Respecto a las segundas, hay expresiones cristalizadas en español con pronombres personales sujeto que han perdido su valor contrastivo: *que yo sepa, lo que tú digas, yo diría…* (Luján, 1999, p. 1282). Finalmente, en cuanto a las terceras, el uso explícito del pronombre puede derivar del objetivo de crear determinados efectos en una obra literaria (por ejemplo, construir y aplicar figuras retóricas como paralelismos, polisíndeton, hipérboles, etc.).

Tanto en todos estos contextos de uso del pronombre personal sujeto de manera expresa como en aquellos en los que se usa por motivos enfáticos y/o contrastivos, la omisión de dicho pronombre y el uso de su equivalente tácito, aunque en ningún momento resulte agramatical ni contradiga las leyes de la sintaxis española, desde un punto de vista semántico dificulta la correcta interpretación del contenido discursivo y obstaculiza la comunicación. Sin embargo, a excepción de este conjunto minoritario de situaciones de

interacción, el recurso a los sujetos nulos constituye la práctica más habitual y más natural en el español hablado y escrito, fundamentalmente por motivos de economía del lenguaje. Dado que la flexión verbal y sus afijos nos conceden la posibilidad de agilizar y sintetizar nuestra expresión, la presencia de los pronombres personales sujeto ha de estar muy justificada (por razones y contextos como los que hemos explicado) para que un nativo haga uso de ellos, pues de lo contrario se expone a resultar redundante sin motivo. Esta característica de la gramática española puede resultar conflictiva para los estudiantes cuyas lenguas maternas no tengan una flexión verbal tan rica y, por tanto, deban explicitar siempre los sujetos, como es el caso del inglés. Pero, ¿qué sucede en el caso de los checoparlantes, el otro tipo de aprendices en los que se centra este pequeño estudio? El verbo en checo también presenta afijos de persona y número, por lo que no resulta necesario utilizar un sujeto explícito. ¿Constituirá este fenómeno aparentemente compartido un caso de transferencia positiva entre la lengua materna y la extranjera? ¿O, por el contrario, los sujetos elididos supondrán un obstáculo en el aprendizaje del español, como sucede en el inglés por motivos de interferencia? ¿Dependerá esto del nivel de dominio del idioma? Estas son las preguntas a las que pretendemos dar respuesta con nuestra investigación.

4.3. Metodología de la investigación

4.3.1. Participantes

Los aprendices del corpus COWS-L2H que redactaron las producciones escritas que hubimos de revisar, corregir y etiquetar durante el año académico 2020/2021 son estudiantes universitarios, tanto hombres como mujeres, que estudian español como segunda lengua o hablantes de herencia, todos ellos matriculados en asignaturas de español en la Universidad de California Davis, que se trata de una institución de enseñanza pública (Yamada et al., 2020). Cabe destacar, a este respecto, que desconocemos los metadatos de los estudiantes en lo relativo a su nivel de competencia exacto; únicamente sabemos que el COWS-L2H abarca varios estadios de dominio del idioma.

El número total de participantes voluntarios en este proyecto es de casi 2 000 informantes a lo largo de diferentes niveles de instrucción y dominio del idioma (por tratarse el COWS-L2H de un corpus de carácter longitudinal), por lo que algunos de ellos realizaron varias contribuciones escritas al corpus y entran dentro del cómputo total tantas veces como composiciones hayan redactado. Si no tenemos en cuenta las diversas aportaciones de estos aprendices de la Universidad de California Davis y los contamos de manera individual, independientemente del número de redacciones que proporcionaron al COWS-L2H, el total de informantes sería de 1 370 hasta la fecha: 850 angloparlantes y 117 sinohablantes (Yamada et al., 2020). Asimismo, como indican los responsables del proyecto, de todos estos aprendices, 420 han redactado y presentado textos durante al menos dos cuatrimestres, 150 durante al menos tres, y 38 durante al menos cuatro, lo cual permite el estudio tanto transversal como longitudinal de los datos, uno de los principales objetivos que se perseguía con la creación del COWS-L2H. En definitiva, podemos decir que el volumen de participantes en este proyecto es considerablemente amplio y, además, está en continuo crecimiento, ya que el proyecto sigue en pie y cada vez son más los aprendices que se ofrecen a colaborar con sus composiciones escritas. Esto no hace sino facilitar la consecución de las metas que se plantearon al inicio de la construcción de este corpus, así como fomentar e incrementar sus múltiples utilidades y aplicaciones.

Por lo que respecta a los informantes checos a los que solicitamos que redactaran producciones escritas breves para poder desarrollar nuestro pequeño estudio contrastivo, se trata de alumnos universitarios checoparlantes, tanto hombres como mujeres, que estudian español como segunda lengua o lengua extranjera, con vistas a su enseñanza, en la Universidad de Bohemia del Sur (České Budějovice, República Checa), una institución educativa también pública. En este sector de la investigación fue donde nos encontramos con el primer problema metodológico que nos llevó a tener que adaptar nuestra investigación a las circunstancias actuales. Como es lógico, dada la situación pandémica y el impacto que ha tenido en la vida universitaria en términos de carga de trabajo, tanto para estudiantes como para profesores, no aspirábamos a compilar un corpus tan extenso como el de COWS-L2H en un período tan breve de tiempo. De haber sido así, habríamos requerido de muchos meses, lo cual era imposible, y, además, nos habríamos tenido que enfrentar a numerosos problemas a la hora de analizar las redacciones, ya que la cantidad de textos sería inabarcable y, por otra parte, no resultaría pertinente para las necesidades, exigencias y limitaciones metodológicas y de extensión del presente trabajo. Por

consiguiente, el número de informantes que precisábamos era considerablemente menor al del proyecto estadounidense.

Sin embargo, en un lapso tan reducido de tiempo y teniendo en cuenta las fechas en las que nos encontrábamos (a punto de finalizar el segundo cuatrimestre del año académico 2020/2021), tan solo hemos podido contar con una participación muy escasa y limitada de aprendices, los cuáles, debido a los encargos, exámenes y demás responsabilidades y cometidos que tienen, han aportado un único texto cada uno. Por ello, resulta evidente que los resultados obtenidos del estudio a partir del corpus creado con las producciones escritas de los alumnos checoparlantes no podrán ser nunca del todo comparables a los extraídos del proyecto COWS-L2H en términos cuantitativos ni permiten desarrollar estudios longitudinales de los datos y la interlengua de los aprendices. No obstante, desde una perspectiva más cualitativa, no consideramos que por ello no se pueda establecer un estudio transversal contrastivo inglés-checo-español interesante y enriquecedor para la investigación en ELE a partir de los datos obtenidos en el análisis de ambos corpus. El argumento que esgrimimos a favor de esta postura es que el conjunto de participantes de ambos estudios comparte muchos otros requisitos y propiedades. En nuestra opinión, estos denominadores comunes confieren a la investigación sobre los textos de checoparlantes, y a la metodología que hemos aplicado, un estatus suficientemente justificado, apropiado y válido como para poder paragonarla al proceso de análisis, corrección y etiquetado de redacciones que hemos llevado a cabo como parte del proyecto COWS-L2H. En este sentido, consideramos que, si bien las condiciones de esta investigación contrastiva no son las ideales y nos habría gustado que fueran más paralelas y equivalentes, podremos alcanzar aun así unas conclusiones pertinentes y provechosas respecto al tema que nos ocupa: el pronombre personal sujeto en español y los beneficios que puede tener para su enseñanza la aplicación de una metodología basada en corpus de aprendices.

Entre los rasgos comunes relacionados con los informantes de ambos corpus, destacan el desconocimiento de su nivel de dominio real del idioma (sabemos que contamos con producciones escritas de diversos niveles en ambos proyectos, pero no conocemos el estadio de competencia de cada informante con precisión), la variedad de sexos en ambos casos (hombres y mujeres), la franja de edad (estudiantes universitarios), el contexto de instrucción (enseñanza reglada en una institución pública) y la relación de los estudiantes con la lengua meta estudiada (español como segunda lengua o lengua extranjera). Existen otras características compartidas entre los dos proyectos, pero están

más bien vinculadas al proceso de recopilación de las producciones escritas, su temática, etc., por lo que profundizaremos en ellas en los próximos apartados. En cualquier caso, consideramos que todos estos puntos en común justifican el carácter "comparable" de los análisis que hemos llevado a cabo de manera independiente, pero paralela, así como de los resultados obtenidos.

Por otro lado, respecto a la cantidad de textos analizados en uno y otro proyecto, en el proceso de corrección y etiquetado de los textos del COWS-L2H no tuvimos que etiquetar todas las producciones escritas que contiene el corpus, solo un conjunto limitado de ellas (195 cada colaboradora). Por tanto, aunque este número es considerablemente más amplio que el de las redacciones obtenidas para nuestro estudio contrastivo con checoparlantes, no lo excede de manera tan desmesurada y reduce la investigación a una cantidad de textos abarcable en ambos casos que nos permitirán, pese a su volumen, extraer unas conclusiones comparativas suficientemente contrastadas, muy interesantes y útiles desde un punto de vista didáctico.

4.3.2. Recopilación de textos

La recolección de textos para este estudio contrastivo se podría dividir en dos secciones claramente diferenciadas: una "impuesta", procedente del COWS-L2H, y otra creada *ad hoc*, con producciones escritas de checoparlantes, para que se correspondiera en todo lo posible con la primera y conseguir así elaborar dos subcorpus equiparables que permitieran extraer resultados hasta cierto punto análogos que se pudieran contraponer para alcanzar conclusiones realmente provechosas (dentro de un marco razonable teniendo en cuenta las condiciones del estudio, lógicamente). Como ya hemos indicado, lo idóneo hubiera sido compilar dos corpus de tamaño semejante, pero dado que no ha sido posible (el de angloparlantes consta de 195 producciones y el de checoparlantes, de 15), hemos tratado de cubrir esa clara carencia cuantitativa asegurándonos de que cualitativamente las dos bases de datos fueran comparables en la medida de lo posible.

La sección constituida por las redacciones del COWS-L2H, que se nos adjudicó por parte de la Universidad de Salamanca y la Universidad de California Davis para su corrección y etiquetado de errores con carácter no permutable, se utilizó como "modelo" para crear después el pequeño corpus de aprendices checoparlantes, que era el que más interés tenía para nuestro estudio. Este conjunto textual originario está formado por 195

composiciones redactadas, como ya sabemos, por diferentes aprendices voluntarios, con distintos niveles de dominio del idioma y matriculados en diversas asignaturas de español, a los que se les encargaron breves tareas de producción escrita de temática sencilla y amplia, poco específica. La mayoría fueron compiladas en el primer semestre del año académico 2017-2018 (que empieza en septiembre de 2017) y aproximadamente un tercio se recopilaron en el segundo semestre del curso 2016-2017 (que comienza en febrero de 2017). Por tanto, los temas en torno a los que giran estas composiciones son dos: *una persona famosa*, por un lado (dos tercios del total) y *mis vacaciones perfectas* (el tercio restante).

Como se ha explicado previamente, nuestra labor en relación con esos textos no servía específicamente a los objetivos del presente trabajo, sino que consistía en la corrección y anotación de un conjunto de fallos muy concretos. La metodología que hemos seguido con respecto al corpus de checoparlantes se ha basado en extrapolar uno de esos errores que, como revisoras, habíamos de corregir y etiquetar (la presencia o ausencia indebida de pronombres personales sujetos) a nuestra pequeña investigación para poder estudiarlo desde un punto de vista contrastivo y de manera abarcable con respecto a otra lengua además del español: el checo. Con esta aclaración pretendemos incidir en el hecho de que el proceso de análisis de ambos corpus ha sido considerablemente diferente porque cada uno servía a unas necesidades específicas distintas y nuestra tarea no seguía las mismas directrices: en el caso del corpus de angloparlantes, teníamos que respetar unas pautas de corrección y etiquetado, pero en el caso del corpus de checoparlantes nuestro cometido giraba más bien en torno a la compilación de datos concretos para su posterior investigación con respecto a otros datos ya recopilados a los que habíamos tenido acceso de manera más "indirecta", como usuarias del corpus (no participamos en el proceso de recolección del COWS-L2H).

En este trabajo se presentarán específicamente los resultados obtenidos del análisis de los pronombres personales sujeto en el corpus de checoparlantes para su contraste sobre todo con el español, puesto que este es el interés central de nuestra investigación, aunque también estableceremos comparaciones con el inglés en determinados momentos. No obstante, incluiremos solo un número reducido de datos concretos extraídos del análisis de los textos anglófonos y no en todos los aspectos examinados, ya que la metodología seguida para el estudio de estas composiciones y las herramientas que nos exigieron emplear fueron totalmente diferentes. Por tanto, la equiparación o contraste que se pueda establecer entre los datos checos y los datos

ingleses se basará en una cifra limitada de ejemplos, en el conocimiento que hemos adquirido gracias a nuestra experiencia directa analizando las redacciones de los aprendices angloparlantes, así como de nuestra competencia como hablantes no nativos de esta lengua, pero no podremos incorporar gráficos o representaciones visuales de los resultados obtenidos del análisis de dichos textos (como sí haremos con las producciones de los informantes checos) debido al distinto enfoque y procedimientos aplicados a su estudio.

Una vez aclarado este punto acerca de la metodología aplicada en nuestro trabajo, procederemos a explicar cómo se compiló el corpus de aprendices checoparlantes a partir del "modelo" establecido por los textos que hubimos de revisar para el proyecto del COWS-L2H. En primer lugar, dadas las limitaciones temporales y de extensión de nuestro trabajo, tuvimos que decidir cuál de los errores corregidos y anotados en el proceso de revisión de los textos de los angloparlantes resultaría más interesante desde un punto de vista contrastivo español-checo-inglés, ya que no podíamos analizarlos todos. Había dos en concreto que podían ser muy útiles y reveladores para nuestra investigación: los errores de presencia o ausencia indebida de pronombres personales sujeto, por un lado, y de artículos, por otro. Tras un largo discurrir, decidimos trabajar sobre el primer tipo de error gramatical, ya que se trataba de una cuestión que habíamos estudiado en menor profundidad a lo largo de nuestra formación como docentes de español como lengua extranjera, además de que considerábamos que los resultados obtenidos a partir de nuestra investigación podrían resultar más enriquecedores no solo desde un punto de vista personal, sino también contrastivo: idealmente, conseguiríamos ampliar la escasa bibliografía existente destinada a estudiar el uso de los pronombres personales sujeto de manera comparada entre el español, el checo y el inglés, aunque de manera modesta.

Decidido el fenómeno gramatical sobre el que trabajaríamos, la siguiente fase de nuestro proceso de compilación textual se centró en la elaboración de la tarea de redacción que solicitaríamos a los aprendices checos para obtener composiciones en las que analizar después el uso del pronombre personal sujeto. Para confeccionar las instrucciones de dicha tarea, seguimos un procedimiento basado en la temática establecida por las producciones escritas del COWS-L2H que habíamos tenido que corregir y etiquetar, de nuevo, con vistas a elaborar un corpus que siguiera una línea paralela y coherente, desde una perspectiva cualitativa, respecto al proyecto estadounidense. Teniendo en cuenta que la mayoría de las producciones se trataba de descripciones sobre un personaje famoso que el aprendiz admirara, consideramos que la aplicación de este tema a la tarea de redacción

de los estudiantes checos constituía una decisión apropiada y pertinente si pretendíamos aspirar a esa "comparabilidad cualitativa" que ya hemos mencionado. Por consiguiente, elaboramos el siguiente enunciado:

> **¿Cuál es tu persona famosa favorita? ¿Por qué? Describe a un personaje famoso que admires y explica por qué. Puedes hablar de su edad, su nacionalidad, su aspecto físico, su personalidad, su trabajo, su vida privada, sus aficiones, cosas que le gustan, cosas que no le gustan, sus logros o éxitos…**

> **¿Has acabado ya tu redacción? ¡Gracias por ayudarme! Antes de enviármela, ¿podrías hacerme otro favor? Escribe a continuación los materiales de apoyo y consulta que has utilizado (si has utilizado alguno) para escribir tu texto: diccionarios online, gramáticas de la lengua española, traductores, libros de texto…**

Como se puede comprobar, las pautas establecidas favorecían la obtención de unos resultados equiparables a la mayoría de las composiciones del proyecto estadounidense en términos de contenido, extensión y variedad de nivel de dominio. Asimismo, con esta propuesta de tarea podemos constatar de manera más directa y visual que, efectivamente, este eje temático inspirado en el COWS-L2H, al igual que los otros tres, es lo suficientemente simple y general como para dejar margen a la creatividad de los estudiantes y, al mismo tiempo, fomentar el potencial uso de múltiples elementos lingüísticos como los pronombres personales sujeto (ya sea de forma correcta o incorrecta), que son aquellos en los que se centra el interés de nuestra investigación. Por otro lado, cabe destacar la última parte de la tarea, en la que solicitamos a los aprendices que indiquen los materiales que han consultado a lo largo del proceso de redacción. El objetivo de requerir esta información era tratar de averiguar 1) si los alumnos utilizaban corpus de aprendices o corpus en general como fuente de datos lingüísticos fiable a la hora de escribir un texto en una lengua extranjera, y 2) si alguno de los materiales utilizados por los aprendices podía tener algún tipo de influencia positiva evidente respecto al correcto uso/omisión de los pronombres personales sujeto.

Cuando tuvimos elaborada la tarea de manera, a nuestro juicio, satisfactoria, la distribuimos entre los estudiantes de diversas asignaturas de español como segunda lengua, y con distintos niveles de competencia, en la Universidad de Bohemia del Sur. En este sentido, cabe destacar que adaptamos las exigencias de extensión de la redacción a cada nivel (80-100 palabras para el nivel inicial, 100-120 para el intermedio y 120-150 para el nivel avanzado). Para la difusión contamos con la inestimable ayuda de los docentes y asistentes lingüísticos de la Facultad de Letras, concretamente del departamento de español. Aplicando esta práctica pretendíamos emular el proceso de

recolección textual del COWS-L2H en la medida de lo posible, para conseguir así un número notable pero abarcable de composiciones de aprendices distintos en diferentes estadios de dominio del idioma extranjero, dadas las características y metas de nuestro estudio. Sin embargo, como hemos indicado en el apartado anterior, pese a que aspirábamos a contar con una participación considerable, tan solo obtuvimos cuatro respuestas. Desconocemos el número exacto de alumnos que recibieron nuestra tarea de redacción, pero sabemos que eran muchos más de los que se prestaron a colaborar. Resulta evidente que la situación pandémica y el aumento en el volumen de carga de trabajo que toda persona vinculada al mundo académico ha experimentado como consecuencia dificultan el desarrollo de investigaciones como la que procurábamos realizar en el marco del presente trabajo. Esta realidad ha influido negativamente en nuestro proceso de recopilación de redacciones, ya que no era pertinente ni provechoso llevar a cabo un estudio sobre el uso que hacen los aprendices del pronombre personal sujeto en español basado en un número tan reducido de textos.

Ante esta situación, tuvimos que tomar una decisión para poder sacar nuestro pequeño proyecto de investigación adelante. Por ello, resolvimos recurrir a redacciones empleadas en el encargo final de una asignatura del Máster en Lingüística Español de la Universidad de Bohemia del Sur que hemos cursado durante el año académico 2020/2021. Estas composiciones procedían también de alumnos universitarios checos que habían estudiado español como lengua extranjera en la Universidad de Bohemia del Sur en años anteriores y que mostraban diversos niveles de competencia en el idioma. Teniendo en cuenta sus características personales (sexo, edad, condición de estudiantes…), su contexto de instrucción (estudios superiores, enseñanza reglada…) y los diversos estadios de dominio de la lengua en los que se encontraban, podían equipararse a los informantes del COWS-L2H del mismo modo que los aprendices de español a los que habíamos enviado directamente nuestra tarea de producción escrita. Por consiguiente, el grado de coincidencia en los metadatos de los participantes anglófonos y checos continuaba siendo suficiente como para justificar la incorporación y análisis de estas redacciones en nuestro estudio. No obstante, antes de introducirlas de manera definitiva en la investigación que pretendíamos desarrollar sobre el uso/omisión de pronombres personales sujeto por parte aprendices checoparlantes, hubimos de resolver un problema adicional vinculado a la temática de estas composiciones.

Como cabía esperar, dado que las producciones escritas que pretendíamos integrar en nuestro pequeño corpus de checos no se habían generado como respuesta a la tarea de

redacción concreta que habíamos elaborado para nuestro estudio, sino como un ejercicio de clase realizado en diferentes niveles de instrucción y dominio del idioma, los temas que abordaban eran muy diversos, así como la extensión. Esto implicaba un claro problema metodológico para nuestra investigación, ya que afectaría negativamente a la calidad del contenido de este subcorpus, la característica fundamental en la que procurábamos centrar nuestra atención e interés para poder confeccionar un estudio en cierto modo coherente, enriquecedor y útil. Por ello, decidimos seleccionar solo aquellos textos cuya temática se asemejara de alguna manera y lo máximo posible a alguno de los dos ejes argumentativos de las producciones escritas del COWS-L2H que habíamos tenido que analizar: *una persona famosa* y *mis vacaciones perfectas*. Como resultado de este proceso de filtrado y criba, obtuvimos once redacciones más: dos descripciones de un lugar desde un punto de vista turístico, seis relatos sobre viajes y tres anécdotas sobre un recuerdo bonito de la infancia. La última sección de composiciones se añadió por analogía con otro de los temas que engloban los textos del COWS-L2H, *una anécdota terrible*, con el objetivo de mantener una línea lógica de recolección y análisis textuales, aunque no trabajáramos directamente con redacciones que giraran en torno a este argumento.

De este modo, a través de todo este proceso de recopilación de textos, no exento de obstáculos metodológicos y procedimentales, obtuvimos un total de quince composiciones: cuatro procedentes de nuestra tarea de producción escrita y once incorporadas *a posteriori* por necesidades investigadoras. En términos de contenido, temática, y edad, sexo nivel de dominio y contexto formativo de los participantes, podríamos considerarlas equiparables a las 195 redacciones analizadas, corregidas y etiquetadas como parte del proyecto relativo al COWS-L2H en el que colaboramos. Como es lógico, nos hubiera gustado obtener más composiciones y respuestas a nuestra tarea, con el fin de incentivar el interés y aplicación práctica de nuestra investigación contrastiva, y de alcanzar unos resultados y conclusiones más sólidos, reveladores, y beneficiosos para el progreso en el ámbito de la enseñanza de ELE, pero, dadas las circunstancias, esto era lo máximo a lo que podíamos aspirar en un lapso tan reducido de tiempo. En cualquier caso, consideramos que el análisis de las composiciones de los checoparlantes y el estudio concreto del fenómeno de la presencia/ausencia del pronombre personal sujeto en español que hemos llevado a cabo tienen mucho potencial y pueden resultar significativos e inspiradores desde un punto de vista didáctico y/o investigador.

A fin de poder realizar este análisis de las composiciones de los checoparlantes, hemos trabajado con la herramienta Analec, para lo cual ha sido necesario convertir estas redacciones, escritas en formato .docx, a documentos de texto sin formato (.txt). Aquellas producciones que habían sido elaboradas a mano tuvieron que ser sometidas a un proceso de transcripción previo a su conversión a archivos de texto sin formato.

4.3.3. Identificación, clasificación y análisis de errores y aciertos en Analec

La detección y el estudio de los errores relativos al uso y omisión indebidos de los pronombres personales sujeto en las redacciones de los aprendices checos se han llevado a cabo a través de la plataforma Analec, un software de anotación y análisis de corpus escritos desarrollado por el Laboratorio Lattice: Langues, Textes, Traitements informatiques, Cognition[16], perteneciente a la Universidad de la Sorbona, y enfocado a la enseñanza y al estudio lingüístico, así como al tratamiento automático de los idiomas. Hemos elegido trabajar con esta herramienta de etiquetado porque su funcionamiento es muy sencillo e intuitivo, pese a que la interfaz esté en francés, y porque permite la gestión, agrupación y visualización global y particular de los diferentes elementos anotados en un corpus escrito. Esto es posible debido a que incluye recursos complementarios que permiten buscar correlaciones en los textos o crear listas de frecuencias y diversos tipos de representaciones estadísticas (estructuras jerárquicas en forma de árbol, esquemas, cadenas, gráficos…). Aún está en proceso de desarrollo, pero consideramos que, ya en su estadio actual, esta plataforma ofrece oportunidades de aplicación notablemente provechosas para el análisis lingüístico y/o pragmático de corpus escritos. Una de las utilidades más ventajosas y prácticas que nos brinda, a nuestro juicio, es la posibilidad de diseñar nuestra propia estructura de anotación lingüística en función de las necesidades concretas del estudio que pretendamos llevar a cabo.

En nuestro caso, tomando en consideración el fenómeno gramatical que pretend3íamos estudiar (los fallos en el uso/omisión del pronombre personal sujeto en español), hemos confeccionado una estructura de anotación con la siguiente disposición para aplicar al análisis de las composiciones de los checoparlantes:

[16] Para más información sobre esta institución y su labor, se puede consultar el siguiente enlace: https://www.lattice.cnrs.fr/.

1. *****Ausencia indebida pron. pers. SUJETO**
2. *****Presencia indebida pron. pers. SUJETO**
3. *****Uso correcto pron. pers. SUJETO**
4. **Ausencia indebida pron. pers. SUJETO**
5. **Presencia indebida pron. pers. SUJETO**
6. **Uso correcto pron. pers. SUJETO**
    a. **Contraste**
    b. **Contraste + Énfasis**
    c. **Necesidades morfológicas/léxicas**
        - **Necesidades léxicas**
        - **Necesidades morfológicas**
    d. **Énfasis**

Como se puede comprobar, hemos basado la estructura de análisis y anotación, por un lado, en los conocimientos teóricos sobre el correcto uso y omisión del pronombre personal sujeto en español, expuestos en el apartado 4.2. de este trabajo, y, por otro lado, en las directrices que recibimos para guiar la corrección y etiquetado de los textos del COWS-L2H que se nos asignaron. La aplicación de este esquema nos ha permitido identificar los errores cometidos por los aprendices checoparlantes en el uso u omisión indebidos de los pronombres personales sujeto de manera sencilla, ágil y efectiva, así como detectar también las prácticas correctas y establecer el valor que cada empleo adecuado tiene o aporta a la oración en cuestión. La apreciación de estos matices de significado tiende a ser subjetiva, pero trataremos de justificar nuestras decisiones en todo momento, según lo establecido por la *Nueva gramática de la lengua española* (2009-2011) y por Marta Luján en su capítulo "Expresión y omisión del pronombre personal", que forma parte de la *Gramática descriptiva de la lengua española* (Bosque Muñoz & Demonte Barreto, 1999). Cabe destacar que, a lo largo del proceso de análisis de errores y aciertos, nos hemos encontrado con algunos casos de interpretación semántica ambigua, por lo que nos ha resultado complicado y conflictivo determinar si se trataba de usos correctos o incorrectos. Analizaremos estos ejemplos concretos con detenimiento en la siguiente sección y trataremos de defender y explicar de manera clara nuestra postura respecto a cada uno de ellos, desde un punto de vista contrastivo, a la hora de decidir si constituyen errores o aciertos.

Por otro lado, una característica fundamental relativa al modelo de estructura elaborado y aplicado que queremos comentar y justificar es la adición de tres categorías marcadas con tres asteriscos iniciales. Estos niveles coinciden con los tres ejes

fundamentales del proceso de identificación y análisis de errores que pretendemos desarrollar para nuestro estudio: ausencia o presencia indebida de pronombres personales sujeto, y su empleo correcto. La inclusión de estas variables se debe a que, en ocasiones, nos hemos encontrado con usos de pronombres personales sujeto pertenecientes a estas categorías que contenían errores, ya fueran ortográficos, de concordancia, de posición en la oración… En consecuencia, decidimos crear nuevas ramas en nuestra estructura de anotación lingüística que nos permitieran clasificarlos dentro del tipo de fallos que les corresponde, pero al mismo tiempo indicando que contienen errores de alguna clase.

Una vez confeccionada e insertada en Analec la estructura de etiquetado de usos correctos e incorrectos de los pronombres personales sujeto, examinamos una a una y en profundidad las quince composiciones de los estudiantes checoparlantes para identificar los errores y aciertos relativos al fenómeno gramatical que nos ocupa y clasificarlos debidamente. Terminado este proceso de análisis, fusionamos (concatenamos) todas las redacciones en un único documento (en el formato de Analec, .ec) para poder obtener estadísticas y gráficos relevantes, conjuntos y representativos de los datos de interés recopilados, lo cual nos permitiría establecer comparaciones con el español y el inglés más enriquecedoras, fundamentadas y prácticas desde un punto de vista contrastivo.

4.4. Resultados

Presentaremos a continuación los resultados cosechados en nuestro estudio, abordando en orden las distintas ramas de las que se compone nuestra estructura de anotación lingüística (las categorías marcadas con asteriscos se tratarán al final). En esta muestra de los datos recolectados en el marco de nuestra investigación, aplicaremos en todo momento un enfoque contrastivo, basado en la exposición y explicación de ejemplos significativos y característicos que hayamos detectado de cada fenómeno concreto, y en la equiparación de usos españoles, checos y, en ocasiones, ingleses. Consideramos que esta metodología resulta pertinente y útil para alcanzar los objetivos que perseguimos en cuanto a servir de fuente de inspiración y enriquecimiento investigadores para el ámbito de la enseñanza de ELE.

4.4.1. Ausencia indebida de pronombres personales sujeto

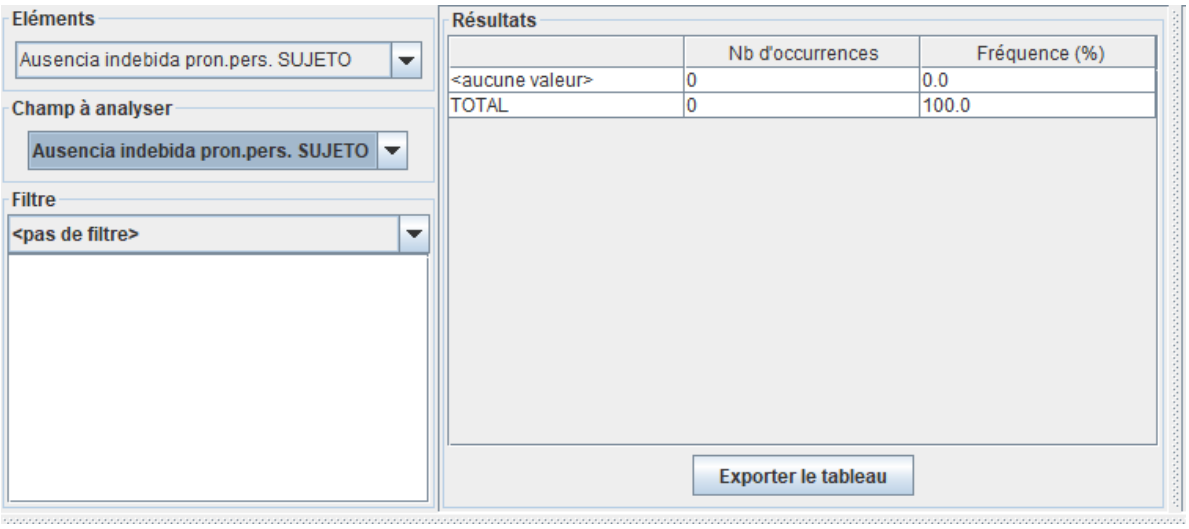

Ilustración 1: Gráfico de ausencia indebida de pronombres personales sujeto extraído de Analec

A partir de este gráfico extraído de Analec observamos que en las quince composiciones de los checoparlantes analizadas no encontramos ningún caso en el que consideraríamos apropiado o necesario incluir un pronombre personal sujeto que no aparece en el texto original para evitar conflictos comunicativos. Por consiguiente, constituye el tipo de error menos frecuente en las redacciones de los aprendices. Esto nos lleva a plantearnos una pregunta interesante: ¿les han explicado en alguna ocasión a los estudiantes checos en qué contextos, aunque sintácticamente resulte innecesario, se utiliza el pronombre personal sujeto de manera explícita por motivos semánticos de contraste o énfasis? ¿Son conscientes de que en determinados casos la omisión del pronombre personal sujeto no es lo más correcto ni idiomático? Como podemos observar, la tendencia clara, incluso en niveles iniciales de competencia en el idioma (las producciones escritas recopiladas para el análisis abarcan diversos estadios de dominio de la lengua), refleja una cierta conciencia por parte de los aprendices checos en lo relativo a la habitual omisión del pronombre personal sujeto en español: les han enseñado que constituye la práctica más habitual que se debe aplicar para utilizar el español de manera natural. Sin embargo, no siempre es el caso. Existen situaciones comunicativas en las que, como ya sabemos, resulta imperativa la aparición explícita del pronombre personal sujeto para garantizar una interacción efectiva. Es evidente que estos contextos, dada su especificidad, constituyen conocimientos de nivel intermedio-avanzado en el dominio del español. Pero, ¿han sido abordados explícitamente en las clases de lengua de nuestros informantes checos? Esto es una realidad que no podemos conocer a ciencia cierta, pero,

a juzgar por el análisis de los usos correctos del pronombre personal sujeto que expondremos en las próximas páginas, los aprendices checoparlantes habrían sido expuestos, según parece, al empleo y valores contrastivos y/o enfáticos típicamente asignados a los pronombres sujeto explícitos en algún momento a lo largo de su proceso de aprendizaje (hecho apreciable sobre todo en aquellas redacciones que denotan un nivel intermedio o avanzado de competencia en el idioma). No obstante, este uso correcto que evidencian algunos de los informantes podría deberse también al conocimiento previo de lenguas afines al español en las que el empleo explícito de los pronombres personales presenta los mismos valores semánticos contrastivos y/o enfáticos, como el italiano. También podría deberse a la casualidad. En cualquier caso, en nuestra opinión, resultaría muy interesante y enriquecedor estudiar la metodología de inclusión de este tipo de cuestiones en el aula de idiomas en un contexto formativo universitario checo, si es que su explicación se inserta de alguna manera en los planes curriculares de las asignaturas sobre gramática española. De no ser así, resultaría igualmente positivo y provechoso estudiar cómo introducir el uso explícito de los pronombres personales sujeto y en qué nivel de dominio. Proponemos esta reflexión como potencial ampliación o foco de interés para futuras investigaciones, pero no podemos desarrollarla en profundidad en el presente trabajo por carencia de tiempo y espacio.

Por otro lado, es importante tener en cuenta que la gramática checa cuenta con una flexión verbal fusional muy elaborada, mucho más rica y compleja que la española. Esto implica que los afijos verbales permiten, al igual que en español, establecer conexiones con el sujeto de la oración, aunque este no sea explícito. Por tanto, podríamos aventurarnos a deducir que el empleo expreso de pronombre personales sujeto podría atenerse a las mismas reglas tener los mismos valores, o similares, que en el caso del español: contraste y énfasis. Sea como fuere, lo cierto es que esta práctica de omisión habitual de los pronombres sujeto no resulta del todo desconocida y ajena para los aprendices checos, por lo que su asimilación no supondría un obstáculo demasiado grande para los checoparlantes ni se producirían serios problemas de interferencia con su lengua materna. No así en el caso de los informantes anglófonos, puesto que la flexión verbal inglesa es mucho más analítica y, en consecuencia, requiere de la presencia explícita de los pronombres personales sujeto para garantizar la eficiencia comunicativa. De ahí que en las redacciones de los estudiantes estadounidenses se observe una mayor tendencia al sobreuso innecesario de estos elementos, como se puede apreciar en el siguiente ejemplo

extraído del COWS-L2H (breve, puesto que se abordará este tema en mayor profundidad en la próxima sección):

> Brad Pitt es un actor famoso. El apellido es Pitt. **El** es alto y muy guapo, pero **el** es viejo. **El** no es feo, gordo, calvo, y bajo. **El** tiene cincuenta y tres anos.

A modo de conclusión de este apartado, podríamos inferir que, aparentemente, los aprendices anglófonos presentarán mayores dificultades en la adquisición de este fenómeno básico y característico de la gramática española que los checoparlantes, debido a la transferencia positiva que se produce entre el checo y el español, frente a la interferencia desventajosa inglés-español, así como a la relativa analogía entre los sistemas verbales de estos dos idiomas, en contraposición a la morfología verbal inglesa. No obstante, antes de terminar con el análisis de a ausencia indebida de pronombres personales sujeto, nos gustaría comentar en cierto detalla un fenómeno observado en las composiciones de los informantes checos que captó considerablemente nuestra atención. Para ello, exponemos los siguientes ejemplos:

> No tengo muchas memorias de mi infancia pero los que tengo son memorias de pura felicidad y satisfacción. Casi cada fin de semana **íbamos** con mis abuelos a nuestro chalet a un pueblito que se llama Slapy.
> También **salíamos** y **viajábamos** mucho con mis abuelos, por ejemplo a zoológicos, a esquiar a Austria, Alemania, Suiza etc.
> Tengo muchas memorias desde cuando era pequeño. Mis mejores son las que cada domingo **íbamos** a casa de mis abuelos.
> Semana pasada nuestro profesor nos dijo que escribieramos una redacción con una storia de nuestra infancia. Primero que me vino a la mente fue un recuerdo del tiempo cuando tenía siete años y cuando **queríamos** aprender a montar a caballo.

Los cuatro ejemplos proceden de textos de carácter anecdótico. En los dos primeros casos, procedentes de la misma composición, podría deducirse que el informante está hablando de sí mismo y su familia inmediata (padres y hermanos/as), pero el contexto es demasiado ambiguo como para estar totalmente seguros de que esa es su intención expresiva. Esta inferencia tendría cabida también en la interpretación del tercer ejemplo. En lo que respecta al cuarto, no resulta suficientemente claro si el emisor se refiere a sí mismo y a sus compañeros de clase (que también tienen que realizar la tarea solicitada por el profesor) o a otro grupo de personas desconocido. Por tanto, como interlocutores nativos, habríamos agradecido la explicitación de un sujeto para agilizar y garantizar una comunicación efectiva, pero, en este caso, lo más conveniente sería que no se tratara de un pronominal. Nuestro argumento se sustenta en el hecho de que el

contenido semántico de los pronombres personales no habría ayudado a desambiguar la expresión, ya que este se establece por correferencia con un elemento que se ha introducido en el discurso previamente (anáforas) o poco después (catáfora). Dado que este no es el caso en ninguna de las muestras, de poco habría servido recurrir al uso expreso de los pronombres sujeto. En definitiva, la ausencia de estos elementos no es indebida en estos casos, pero la ausencia de un sujeto sí lo es.

4.4.2. Presencia indebida de pronombres personales sujeto

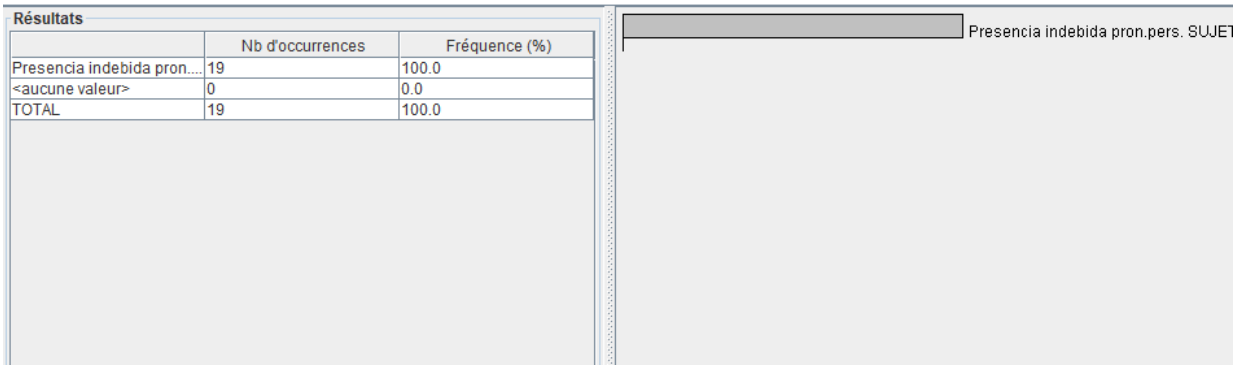

Ilustración 2: Gráfico de presencia indebida de pronombres personales sujeto extraído de Analec

El uso explícito de pronombres personales sujeto en contextos en los que resulta innecesario constituye el error más frecuente de las composiciones de los checoparlantes, con una diferencia considerable respecto a la omisión indebida. El hecho de que en quince redacciones de extensión breve se produzcan diecinueve fallos por presencia indebida de pronombres sujeto resulta muy significativo y, al mismo tiempo, sorprendente teniendo en cuenta la "analogía" que existe entre el español y el checo en términos de flexión verbal. ¿A qué se debe entonces el uso redundante, prescindible y poco natural de estos elementos? Procederemos ahora a analizar algunos ejemplos concretos para tratar de determinar cuál es el motivo que se halla detrás de esta incorrección.

> Al primer vistazo parece una mujer fuerte, independiente, un poco salvaje, pero al mismo tiempo hermosa y atractiva con sus grandes ojos fijos y pelo indomable. Si alguien dijera 'fuera de serie' su imagen es lo que me vendría a la mente. Al mirarla uno se da cuenta de que **ella** es excepcional inmediatamente.
> En esta ciudad hay más de 90 mil habitantes y abunda de monumentos históricos y museos. Pero **yo** no quiero hablar sobre al Plaza de Přemysl Otakar II.
> Cuando ella fui en el hospital, nosotros visitamos el Stonehenge con amigo Pavel. Allí fui muchos gentes , Pavel estuvó enojado. Despues **nosotros** cogemos el bús y fui a la restaurante pequeña, cerca de nuestros hotel.

> Una media hora después noticimos que **ella** fue detrás de nosotros. **Ella** tenía miedo de la gente y de los autobuses públicos. Todo pasé solamente porque **ella** quería hacer muchos fotos para usar en el proyecto de alguna clase en la secundaría. Ultimamente **nosotros** visitamos todos los monumentos durante nuestros viajes.
> Hace 7 años cuando **yo** visité un país extranjero por la primera vez.
> Recuerdo este día muy bien. Fuimos a un paseo en naturaleza. **Yo** tuve una yegua muy tranquila y por eso no tuve miedo. Estabamos pasando alrededor de un árbol cuando de repente apareció un ciervo. Mi yegua se asustó mucho, saltó y dio una vuelta de 180 grados. Claro que **yo** caí y por suerte no me ocurrió nada. **Yo** empezé llorar y mientras intentaba coger mi caballo, el ciervo escapó.

En los dos primeros casos parece existir una clara intención enfática. El nivel de dominio que demuestran estos dos informantes en su expresión a lo largo de su redacción se ubicaría en torno al intermedio alto-avanzado, por lo que cabría intuir que conocen perfectamente las normas gramaticales que rigen el uso/omisión del pronombre personal sujeto y es probable que sean conscientes de las razones y contextos que motivan la explicitación de estos elementos en la lengua española. No obstante, en estos casos, consideramos que el énfasis resulta innecesario o que queda patente a través de otros mecanismos. Así, en el primer ejemplo, dado que ya se han introducido conceptos como *indomable*, *fuera de serie* y *excepcional* para referirse con una intensidad acentuada a las virtudes de la persona de la que se está hablando, la adición del pronombre explícito resultaría redundante y repetitiva, lo cual deriva en un exceso de recursos expresivos para enfatizar una idea. En el segundo caso, no encontramos ningún motivo en el contexto de la oración por el que resulte imperativo destacar expresamente el sujeto, por lo que su presencia sería del todo indebida e innecesaria.

Por lo que respecta a los otros cuatro fragmentos extraídos de las redacciones de los aprendices checos, se puede apreciar por el marco discursivo en que se contextualiza el uso explícito de los pronombres personales sujeto que no hay ningún tipo de propósito contrastivo, enfático o morfológico detrás que justifique su aparición. Se trata de composiciones que denotan un menor nivel de competencia en el idioma por parte de los estudiantes, por lo que se podría deducir que este empleo expreso de los pronombres se debe a un menor dominio de la gramática española. No obstante, cabría preguntarse por qué se produce este fenómeno, si la flexión verbal checa también permite la omisión del pronombre personal sujeto debido a su riqueza en afijos y, de hecho, constituye la práctica habitual, como en el español. En este sentido, existen dos razones que podrían explicar este hecho cuando menos sorprendente.

La primera de ellas es la inseguridad lingüística. Cuando aún no se domina una lengua extranjera, o desconfiamos de la estabilidad y solidez de nuestro conocimiento lingüístico, los aprendices tendemos a pecar por exceso en lugar de por defecto en lo que a fenómenos lingüísticos se refiere. Por tanto, existe una clara predisposición o inclinación extendida entre los estudiantes hacia el sobreuso frente al infrauso, ya que suelen preferir errar por haber utilizado elementos innecesarios que por haber omitido aquellos que sí resultaban indispensables o fundamentales para garantizar la eficiencia y el éxito comunicativos. Como aprendices de lenguas extranjeras, conocemos bien este temor y somos conscientes de que puede ser la causa por la que los estudiantes checos en niveles iniciales o incluso intermedios recurran al uso explícito del pronombre personal sujeto en contexto indebidos, pese a estar familiarizados con la omisión de estos elementos a raíz de su lengua materna, por carencias en su "confianza" y "autoestima" lingüísticas.

El segundo motivo que podría justificar la presencia indebida de pronombres personales sujeto en las composiciones de los checoparlantes es la interferencia con una lengua extranjera aprendida previamente. Es muy probable que estos aprendices hayan estudiado inglés en algún momento anterior a su contacto con el español, sobre todo teniendo en cuenta las exigencias lingüísticas de la enseñanza reglada primaria y secundaria en Europa. Como sabemos, el inglés constituye una lengua analítica que carece de la riqueza flexiva del checo o del español, por lo que requiere del uso explícito de más mecanismos lingüísticos que estas dos lenguas en términos discursivos para que no se produzcan malentendidos o problemas comunicativos en actos de interacción de cualquier índole. Esta es la razón por la que los aprendices anglófonos de español recurren con excesiva frecuencia a los pronombres personales sujeto, lo cual fomenta la redundancia, desnaturaliza la expresión y da muestras de un bajo nivel de dominio del idioma por parte del alumno. Sirva el siguiente fragmento como ejemplo ilustrativo de esta realidad:

> Un persona famoso es Patrick Stump. **Él** es un cantante en la banda de rock. **El** nombre de su banda es Fall Out Boy. **Él** es muy comico y inteligente. **Él** es muy talentoso. Puede tocar muchos instrumentos. Puede tocar la trompeta, los tambores y la guitarra.Su voz es muy hermosa. Su cumpleaños es el veintisiete de abril.Patrick tiene treinta y tres años.**Él** escribe muchas de las canciones de la banda. Le gusta leer y aprender nuevos idiomas. **Él** es de una pequeña ciudad cerca de Chicago, Illinois. Sus músicos favoritos que crecen incluyen Michael Jackson y Elvis Costello.

En definitiva, para concluir con el análisis de la presencia indebida de pronombres personales sujeto, cabe destacar que, pese a la familiaridad de los estudiantes checos con la omisión de estos elementos, ya que constituye un hábito lingüístico-gramatical también en su idioma materno, su uso explícito es el error considerablemente más frecuente en las producciones escritas compiladas. A partir del análisis de los contextos de expresión escrita en los que se producen este tipo de fallos, hemos podido deducir que los cometen presumiblemente por motivos de carencias instructivas, de ausencia de confianza en su dominio del idioma a lo largo de su desarrollo lingüístico, o de transferencia negativa entre lenguas extranjeras.

4.4.3. Uso correcto de pronombres personales sujeto

4.4.3.1. Contraste

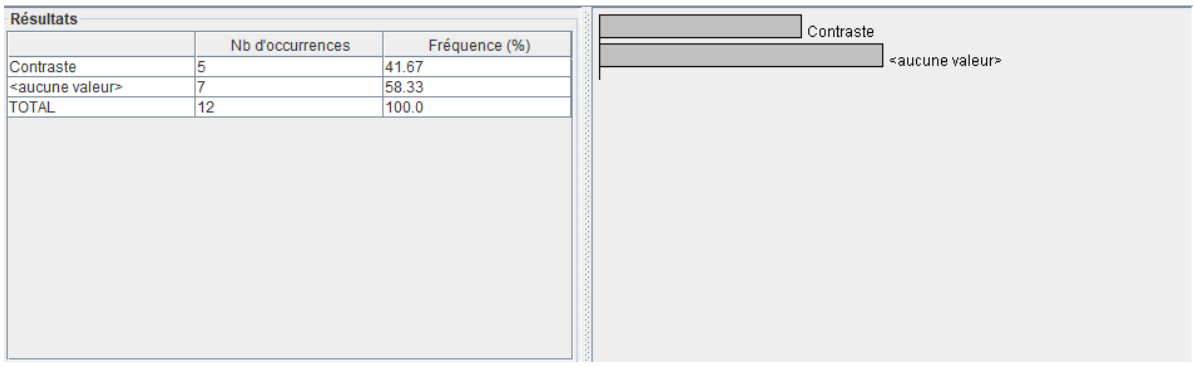

Ilustración 3: Gráfico de usos contrastivos correctos de pronombres personales sujeto extraído de Analec

De los catorce casos de usos correctos de pronombres personales sujeto explícitos que hemos detectado en total en las composiciones de los informantes checos, su empleo contrastivo es el más frecuente. Cabe destacar, llegados a este punto y antes de continuar, que resulta difícil diferenciar cuándo un pronombre sujeto expreso se utiliza con fines enfáticos, contrastivos o ambos. Por tanto, queremos puntualizar que nuestra clasificación de usos en unas subcategorías u otras dentro de esta rama de la estructura de anotación lingüística se ha basado en la interpretación "subjetiva" de los valores de cada pronombre concreto en cada caso; ahora bien, teniendo en cuenta en todo momento tanto el contexto de uso como los conocimientos teóricos expuestos en el apartado 4.2. acerca de la presencia justificada de los pronombres personales sujeto en español, con el fin de

desarrollar un análisis lo más objetivo posible dentro de la subjetividad inherente de este tipo de clasificaciones.

Una vez aclarada esta cuestión, procedemos a exponer y explicar aquellos segmentos en los que se ha utilizado correctamente el pronombre personal sujeto expreso por enmarcarse en contextos de contraste. A tal efecto, véanse los siguientes ejemplos:

> Yo y mis dos amigos fuimos a la Inglaterra el año pasado. Pasé mucho tiempo en el hospital, porque mi amiga Tereza craó de la ventana. Esta ventana está en el tercero piso. Creo que Tereza puede ser feliz que es vivo. Cuando **ella** fui en el hospital, **nosotros** visitamos el Stonehenge con amigo Pavel.
> Durante el día en la ciudad mi hermana mayora quería ver muchos monumentos famosos, y tenía un argumento con mis padres, porque **yo** no podría caminar tan prontor como los adultos.
> Allí **ella** trabajas cómo la periodista y tienes un blog en el internet… y **yo** camino más pronto, porque soy un adulto y también quiero ver todos los monumentos y sitios históricos.

En el primer fragmento, podría fácilmente interpretarse que se trata de un uso incorrecto por presencia indebida de pronombres personales sujeto. De hecho, esa fue nuestra clasificación inicial, dado el bajo nivel de dominio que demostraba el informante. No obstante, teniendo en cuenta el contexto en el que se encuadra esta oración, podría aplicársele un valor contrastivo, ya que contrapone dos términos claramente diferenciados: la persona que se queda en el hospital y los otros dos miembros del grupo, que se van a hacer turismo (obsérvese el fragmento subrayado). Por tanto, aunque en un primer momento pueda parecer redundante, lo cierto es que el uso explícito de los pronombres personales sujeto en este caso concreto no resultaría del todo inadecuado según la postura o enfoque que se le adopten.

La segunda muestra constituye un claro ejemplo de contraposición entre dos realidades confrontadas: por un lado, los adultos (padres y hermana del informante, en este caso) y, por otro, una persona en edad infantil (el aprendiz). Lo mismo sucede en el tercer fragmento, en el que se contraponen las distintas acciones, destacables por su contraste, de dos personas diferentes, con la intención de reflejar una evolución en el comportamiento y/o características de cada una de ellas. En este último caso resulta muy evidente que se desnaturalizaría considerablemente la expresión y se obstaculizaría la comunicación si se omitieran los pronombres personales sujeto. Cabe destacar, asimismo, que la lectura del valor del pronombre en el segundo ejemplo podría tener un valor morfológico también, ya que el verbo *podría* se corresponde tanto con la primera persona del singular (yo) como con la tercera persona del singular (él/ella/ello) y previamente se

ha hablado de una tercera persona (la hermana del informante), por lo que el uso expreso del pronombre sujeto podría responder también a necesidades gramaticales que, de no cubrirse, perjudicarían tanto al sentido de la oración como, por ende, a la efectividad de la comunicación.

Al respecto de este tipo de presencia apropiada de pronombres personales sujeto, hemos de dejar constancia de que, por sorprendente que parezca, en las 195 redacciones de anglófonos analizadas para el subproyecto desarrollado a partir del COWS-L2H en el que hemos colaborado, solo hemos detectado un único uso contrastivo. De hecho, prácticamente todos los casos de empleo explícito de este tipo de pronombres constituían errores por presencia indebida. Incluimos aquí ese único ejemplo contrastivo para analizar su sentido y motivación.

> Ellos tengieron muchas victorias el año pasado. **Yo** juge futbol de americano pero yo no como bueno como esa gente ese juegan para la Universidad de Florida o Florida estado Universidad

En este fragmento, el emisor contrasta sus aptitudes en un ejercicio físico concreto con las de los miembros de un equipo muy habilidoso en esa disciplina deportiva, lo cual justifica adecuadamente la explicitación del pronombre sujeto. El hecho de que solo hayamos encontrado una muestra de este tipo de uso en las composiciones de los anglófonos es una realidad totalmente inesperada y un tanto inverosímil, dada la considerable diferencia de volumen entre los dos subcorpus con los que hemos trabajado (cabría esperar encontrar más muestras de este fenómeno en la base de datos estadounidense). Los motivos que pueden hallarse detrás de este insólito fenómeno irían desde el nivel de dominio hasta el desconocimiento de estas estructuras contrastivas, pasando por el relativo o insuficiente potencial que presentaría la temática de las redacciones para favorecer la aparición de este tipo de contextos de uso tan concretos. Cabe destacar, por otro lado, que no hemos registrado tampoco ningún uso explícito correcto de los pronombres personales sujeto ni por motivos enfáticos ni por motivos enfáticos y contrastivos a su vez, por lo que huelga comentar de nuevo esta cuestión en dichos subapartados.

4.4.3.2. Contraste + énfasis

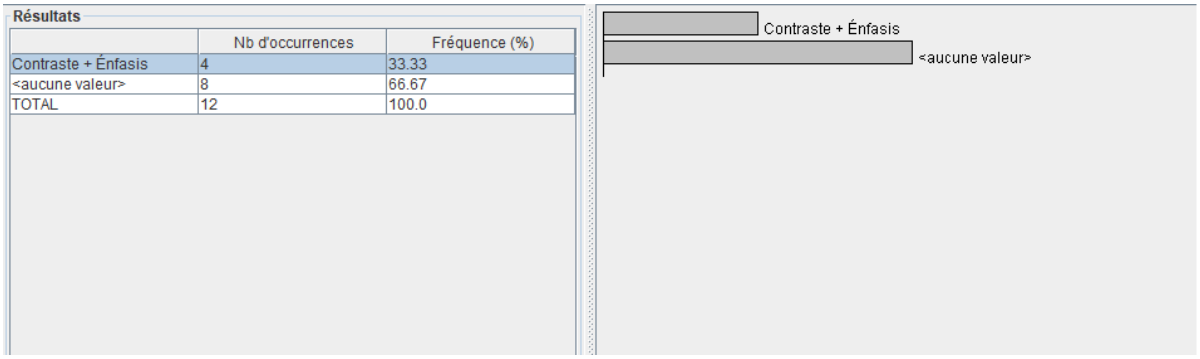

Ilustración 4: Gráfico de usos contrastivos y enfáticos correctos de pronombres personales sujetos extraído de Analec

Los ejemplos en los que se aprecia una combinación entre valores contrastivos y enfáticos son los segundos más numerosos dentro del correcto empleo de los pronombres personales sujeto. Examinaremos ahora estos usos a partir de las muestras ilustrativas encontradas en el proceso de análisis textual:

> Quiero presentaros sitios y monumentos situados en el centro histórico que aunque son bastante interesantes no son tan conocidos. **Yo** recomendaría empezar por la mañana con el paseo por el parque Na Sadech y así se puede llegar a la plaza Piaristické náměstí.
> Y por eso quiero decir algo más sobre el viaje a Fuerteventura que pasó el mes pasado. Yo con mi novio queríamos visitar el mar. **Yo** siempre decía que quería ir a España – practicar la lengua, conocer más sobre la cultura y otro. Fue difícil, pero al final é dijó: "Pues, vamos!"
> Para prestar el coche fue necesario depositar el dinero (mucho dinero) y nosotros en jueves y viernes también olvidaremos que lo habíamos hacido. En sábado **yo** estaba preocupada y decía: "¡Qué vámos a hacer?" Graciadamente la mujer no olvidó y volvió nuestro dinero.
> Allí ella trabajas cómo la periodista y tienes un blog en el internet… y **yo** camino más pronto, porque soy un adulto y también quiero ver todos los monumentos y sitios históricos. En todos los países del mundo.

En el primer ejemplo, el informante utiliza expresamente el pronombre personal sujeto para diferenciar y destacar su opinión con respecto a la de otras posibles personas que puedan oponerse a ella, por lo que contrasta dos realidades, enfatizando la que él defiende como más importante o válida en el contexto de su discurso. En el segundo y el tercer caso, la informante se individualiza frente a su pareja y recalca sus intenciones, pareceres y emociones, concediéndoles un mayor relieve dentro de su relato; de ahí que al contraste establecido por la contraposición entre dos términos (la aprendiz y su novio) se añada un matiz enfático.

En cuanto al último fragmento, lo hemos incluido también en el apartado anterior porque consideramos que, si bien la interpretación de *ella* sería más evidentemente contrastiva, el valor del pronombre *yo* resulta más controvertido. Como ya hemos explicado, nos hemos encontrado con casos ambiguos que podían presentar una doble lectura y, a modo ilustrativo, hemos considerado oportuno incluir este ejemplo. No cabe duda de que ese pronombre de primera persona del singular cumple una función contrastiva en contraposición al *ella* que lo precede. No obstante, si tenemos en cuenta el conjunto textual (narración sobre un viaje), el informante alude previamente al hecho de que su hermana (*ella*) discutía con los padres de ambos porque quería visitar más lugares de lo que sus progenitores consideraban conveniente teniendo en cuenta la corta edad del informante. Por consiguiente, sería plausible interpretar que el aprendiz pretende no solo individualizarse con respecto a su hermana, sino también incidir en una nueva realidad que considera importante: es adulto y ya puede seguir un ritmo de visita turística más exigente. En suma, consideramos que este pronombre podría tener dos sentidos o, más bien, dos matices semánticos igualmente válidos, ya que ambos son contrastivos; simplemente en uno de ellos se incluye un matiz enfático en la expresión que el otro no contempla.

4.4.3.3. Necesidades morfológicas/léxicas

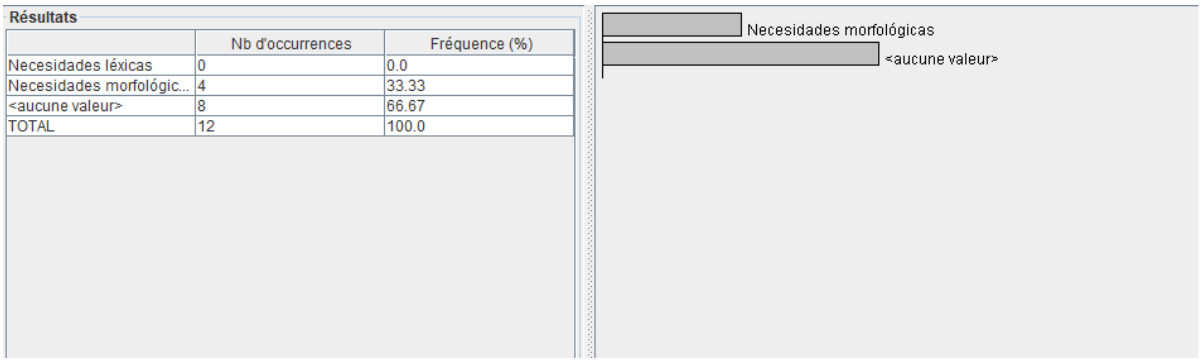

Ilustración 5: Gráfico de usos morfológicos/léxicos correctos de pronombres personales sujeto extraído de Analec

Esta subcategoría engloba el empleo necesario de los pronombres personales sujeto por motivos léxicos o morfológicos. No hemos encontrado ningún ejemplo en las quince redacciones de los checoparlantes ni en las 195 de los anglófonos correspondiente al primer tipo, que comprende aquellas estructuras o expresiones cristalizadas en las que se requiere del uso explícito de estos pronombres, sin valor contrastivo (*que yo sepa*, *yo diría…)*. Es posible que esto se deba a que este lenguaje formulaico en español se suele

enmarcar dentro de la enseñanza en niveles intermedio-alto avanzado y se utiliza en contextos muy concretos, quizá no propiciados por la temática aplicada a las composiciones de estos aprendices.

Por lo que respecta a las necesidades morfológicas, hemos detectado cuatro casos en los que la morfología o la sintaxis requerían de la explicitación del sujeto a través de pronombres, a saber:

> Allí quedábamos toda la familia con mis primos y tíos. Siempre mi primo y **yo** jugábamos con mi abuelo. […] Somos una familia grande, así que mi primo y **yo** estábamos jugando con nuestro abuelo y mis primas siempre jugaban con sus muñecas en el otro lugar.
> Durante el día en la ciudad mi hermana mayora quería ver muchos monumentos famosos, y tenía un argumento con mis padres, porque **yo** no podría caminar tan prontor como los adultos. Después del argumento **ella** quería ir solamente con la cámara de mi padre.

En los dos primeros usos, el pronombre personal sujeto expreso resulta necesario para aclarar a qué "nosotros" se refiere el informante cuando alude al grupo de personas que realiza las acciones *jugábamos con mi abuelo* y *estábamos jugando con nuestro abuelo,* ya que previamente ha mencionado a "toda la familia" como un *nosotros*, pero no es toda la familia la que lleva a cabo esas actividades. Por tanto, la construcción "X y yo" constituye un recurso claro y clásico en español para explicitar de manera más específica qué "nosotros" es el agente de la acción verbal. De haber escrito el aprendiz simplemente *nosotros*, cambiaría el sentido de la oración (toda la familia jugaría con el abuelo o estaría jugando con el abuelo, ya que es el sujeto de primera persona del plural inmediatamente anterior) y la expresión en español sería poco idiomática y natural; de hecho, lo habríamos catalogado como presencia indebida de pronombre personal sujeto, pues habríamos interpretado que se referiría a "toda la familia" y resultaría, por tanto, redundante y superfluo. Por tanto, cabe destacar la necesidad claramente morfológica que justifica estos dos usos explícitos del pronombre sujeto, pero también su matiz semántico.

En cuanto a los dos últimos ejemplos, la presencia de los pronombres personales sujeto se hace totalmente necesaria debido a las características de la flexión verbal en las formas de imperfecto y condicional simple de indicativo. En estos dos tiempos verbales, las desinencias para la primera y la tercera persona del singular coinciden. Dado que ambas personas han aparecido en el discurso previo al empleo de estas formas, cabría la posibilidad de que se produjera una confusión tanto en el primer caso (se podría interpretar que es la hermana mayor la que no sería capaz de caminar tan rápido como los

adultos) como en el segundo (podríamos entender que el informante era el que quería irse a hacer turismo solo con la cámara tras la discusión, ya que se ha introducido un *yo* previamente). Es posible que el contexto discursivo ayudara a evitar los malentendidos comunicativos, pero lo cierto es que la ambigüedad expresiva seguiría presente, ya que dicho contexto no es especialmente preciso ni esclarecedor. Por tanto, el uso explícito de los pronombres personales sujeto resulta imperativo.

Cabe destacar, asimismo, que el empleo del primer pronombre personal sujeto (*ella*) tendría un cierto matiz enfático, ya que la presencia de este elemento intensifica la idea de que es la hermana mayor la que quiere disgregarse del grupo tras una disputa familiar para seguir sus propios planes. Por tanto, aunque consideramos que la razón fundamental que lleva al uso explícito del pronombre es de naturaleza morfológica, pensamos que, además, existe un valor de énfasis vinculable a este elemento y, en consecuencia, lo hemos clasificado bajo dos subcategorías: necesidades morfológicas y énfasis (constituye el único ejemplo en el que apreciamos un uso enfático no necesariamente vinculado a un contexto también de contraste, por lo que carece de sentido que desarrollemos un subapartado para tratar exclusivamente esta rama de la estructura de anotación diseñada). En cualquier caso, como ya hemos indicado, la percepción de matices contrastivos y/o enfáticos depende de la interpretación un tanto subjetiva que se haga del contexto en que aparece el pronombre.

En lo que respecta a los textos de los anglófonos se puede apreciar también el empleo de pronombres sujeto por exigencias morfológicas que, de no cumplirse, obstaculizarían la interacción y perjudicarían a la efectividad de la comunicación. No obstante, la presencia de este fenómeno en dichas composiciones es mínima: de nuevo, tan solo podemos observar un ejemplo ilustrativo de esos usos por parte de los aprendices estadounidenses que presentamos a continuación.

> Usted puede disfrutar de la comida española, música y festivals. España es un país hermoso con playas increíbles también. España está situada en el mar Mediterráneo. Por la mañana **yo** disfrutaría de una taza de café y leería el periódico. Esto es mi mañana ideal.

Se especifica explícitamente el sujeto de primera persona del singular para diferenciarlo de manera clara con el de tercera persona del singular (*usted*) que ha aparecido anteriormente en el discurso. Es cierto que podría aplicársele a ese uso un valor contrastivo e incluso enfático también, ya que individualiza la opinión del informante sobre la del resto de individuos. No obstante, consideramos que aquí las necesidades

morfológicas son la razón de mayor peso que justifica la aparición expresa del pronombre sujeto.

Si tomamos en consideración todo lo expuesto respecto al uso de pronombres personales sujeto expresos por razones morfológicas, podríamos concluir que los aprendices checos demuestran un mejor conocimiento de este tipo de estructuras y necesidades, teniendo en cuenta que el volumen de redacciones de checoparlantes analizas es muy inferior al número de composiciones de anglófonos revisada pero, sin embargo, el conjunto de usos explícitos de pronombres sujeto por motivos morfológicos es proporcionalmente mayor. En este caso, dudamos que esta realidad se deba a cuestiones de transferencia positiva entre la lengua materna y la extranjera, ya que el checo tiende a utilizar de manera mucho más reducida los pronombres personales en función de sujeto de manera expresa dada la riqueza de su flexión verbal, a diferencia del inglés, idioma en el que el empleo explícito de los pronombres es obligatorio en prácticamente todos los casos. Probablemente la necesidad de recurrir a estas estructuras está vinculada de manera directa a los temas en torno a los que giran las producciones escritas de unos y otros aprendices, ya que ciertos ejes argumentales se podrían prestar mejor a la incorporación de formas verbales cuyas desinencias pueden provocar confusiones de no explicitar los sujetos: imperfecto, pluscuamperfecto y condicional simple y compuesto de indicativo, pretérito perfecto de subjuntivo… En este sentido, los relatos anecdóticos presentes en el subcorpus de estudiantes checos, pero no en el de estadounidenses (porque no se nos asignaron textos con esta temática), favorecen el uso de tiempos verbales de pasado de indicativo, entre los cuales se encuentran algunas de esas formas "problemáticas".

4.4.3.4. Variables marcadas con asteriscos

Como se indicó en el apartado 4.3.3., en el que explicamos el diseño y los niveles de nuestra estructura de anotación lingüística creada para trabajar con Analec, decidimos insertar tres ramas adicionales encabezadas por tres asteriscos con el objetivo de clasificar los aquellos fallos o aciertos analizados en nuestro estudio (ausencia/presencia indebida de pronombres personales sujeto y uso correcto de dichos pronombres) que presentaran algún tipo de incorrección ortográfica, de posición, etc. Una vez examinadas las quince redacciones de los checoparlantes, tan solo hemos encontrado cuatro casos que se corresponderían con estas categorías: un error por empleo explícito indebido de un pronombre personal sujeto y tres usos expresos correctos de estos elementos.

> **Yo** con mi novio queríamos visitar el mar. Yo siempre decía que quería ir a España – practicar la lengua, conocer más sobre la cultura y otro. Fue difícil, pero al final **é** dijó: "Pues, vamos!"
> Cuando tenía solo cuatro años, **yo** y mi familia viajemos a Italia.
> **Yo** y mis dos amigos fuimos a la Inglaterra el año pasado.

En el primer fragmento, la redacción del pronombre de tercera persona del singular *él* como "é" lo convierte en un error por presencia indebida de pronombre personal sujeto (no resulta necesario ya que no hay motivos contrastivos, enfáticos o morfológicos que lo justifiquen) que, además, contiene un fallo ortográfico. Respecto al resto de pronombres clasificados bajo estas ramas, constituirían usos correctos de los pronombres personales sujeto dentro de la estructura típica "X y yo", que se emplea para especificar de manera clara el "nosotros" que realiza la acción verbal, como ya hemos explicado en el apartado anterior. No obstante, apreciamos en estas construcciones un error en la posición de dicho pronombre personal, ya que, según las reglas de la gramática española, tendría que ir pospuesto al resto del grupo nominal que conforma el "nosotros".

Los estudiantes anglófonos también presentan fallos ortográficos notable en lo relativo a la redacción de los pronombres personales sujeto, más allá de que los utilicen explícitamente de manera indebida. Veamos algunos ejemplos en los siguientes fragmentos extraídos del COWS-L2H:

> Brad Pitt juega baloncesto y golf en escuela. **El** estudia periodismo en la Universidad de Missouri. **El** tiene un hermano y un hermana. Vive en Los Angelos.
> **Ellos** pone en el muchos esfuerzo y cree en ellos mismos.
> **El** tienen pelo cafe y no tiene bigote.

En estos casos se puede apreciar la ausencia de la tilde en el pronombre personal de tercera persona del singular *él* o los problemas de concordancia entre el pronombre sujeto y los verbos (véase los elementos subrayados). No obstante, respecto a la posición del pronombre sujeto en estructuras de tipo "X y yo", que en el caso de los checoparlantes presentan fallos de orden, los aprendices anglófonos muestran un mejor dominio, ya que no hemos detectado ningún segmento en el que el pronombre anteceda al resto del grupo nominal. Sirva el siguiente ejemplo como muestra de esta afirmación

> Pasaría el tiempo con mi familia en vistas. Es mi actividad favorita de los padres. Se sentarían y beberían maragaritas. **Mi hermana y yo** comería patatas fritas. Durante el día, **mi hermana y yo** también tomaríamos siestas.

A nuestro juicio, el motivo que explica el acierto de los anglófonos, por un lado, y el error de los checoparlantes, por otro, está relacionado con la influencia positiva o negativa de la lengua materna, respectivamente. En inglés, la construcción más habitual de este tipo de estructuras también antepone el grupo nominal al pronombre personal sujeto (*my sister and I*), por lo que esta práctica habitual se transfiere y aplica a la lengua extranjera en el proceso de aprendizaje, en esta ocasión, el español. Por el contrario, en lo que respecta a la gramática checa, el orden de los elementos es mucho más libre que en el caso del español o el inglés, probablemente como consecuencia de su riqueza flexiva (sistema de casos y declinaciones) que permite saber con precisión la función de cada elemento en la oración independientemente del lugar que este ocupe. Por tanto, sería posible encontrar ejemplos tanto de la expresión *moje sestra a já* como de *já a moje sestra*, y ninguno se consideraría del todo incorrecto. Esta variabilidad habitual en checo no es extrapolable al español, cuyas normas gramaticales respecto al uso del pronombre personal sujeto explícito en estos casos son más rigurosas: tiene que aparecer al final del grupo nominal. Por tanto, se produciría en estas estructuras una interferencia procedente de la lengua materna checa que influiría negativamente en el desarrollo de la interlengua y el dominio y conocimiento lingüísticos del aprendiz checoparlante.

4.5. Conclusiones del estudio

En el planteamiento y explicación de la fundamentación teórica de nuestra investigación nos propusimos responder a una serie de preguntas con respecto al uso/omisión del pronombre personal sujeto por parte de los aprendices checos: dado que el verbo en esta lengua también presenta afijos de persona y número que eximen a los hablantes de la necesidad de explicitar el sujeto, ¿constituirá este fenómeno aparentemente compartido un caso de transferencia positiva entre la lengua materna y la extranjera? ¿O, por el contrario, los sujetos elididos supondrán un obstáculo en el aprendizaje del español, como sucede en el inglés por motivos de interferencia? ¿Dependerá esto del nivel de dominio del idioma? Expondremos ahora las respuestas a las que hemos llegado a través de nuestro pequeño estudio.

A modo de resumen, podemos concluir que, efectivamente, se aprecian casos de extrapolación positiva de reglas y estructuras lingüísticas que facilitan considerablemente la asimilación de la habitual elisión de los pronombres que funcionan como sujetos, a

excepción de contextos de uso muy concretos. En este sentido, es cierto que hemos detectado un mayor número de errores que de aciertos en el empleo de los pronombres sujeto en las composiciones de los alumnos checos, pero no por una diferencia demasiado significativa (veinte fallos frente a dieciocho aciertos). Asimismo, solo hemos encontrado ejemplos de uno de los dos tipos de errores que pretendíamos analizar (presencia indebida) y ninguna muestra del otro (ausencia indebida). Este hecho resulta sorprendente y muy satisfactorio, ya que, por un lado, corrobora la influencia positiva que puede tener la lengua materna de los aprendices checos en su aprendizaje del español, pese a tratarse de lenguas tipológicamente muy diversas y alejadas, y, por el otro, demuestra un cierto grado de exposición tanto a usos correctos como incorrectos de los pronombres personales sujeto que redunda en un dominio notable de aquellos contextos en los que su explicitación se hace necesaria y adecuada (se aprecian buenas prácticas incluso en textos que denotan un bajo nivel de competencia en español, pese a tratarse estos fenómenos de conocimientos, a nuestro juicio, pertenecientes a niveles intermedios y avanzados de dominio del idioma).

Sin embargo, no debemos olvidar que el checo y el español son dos lenguas distintas pertenecientes a familias lingüísticas diferentes, por lo que, aunque presenten algunas similitudes y analogías en sus sistemas y afijos verbales que les conceden la posibilidad de omitir los pronombres personales sujeto, han atravesado dos procesos de evolución histórico-lingüística muy diversos que impide considerarlos "lenguas afines". Esto se aprecia claramente en el hecho de que, pese a que pueda haber transferencias positivas del checo al español, también pueden producirse interferencias (como sucede con la posición del pronombre en grupos nominales en función de sujeto formados también por otros sustantivos y/o pronombres: "X y yo").

En cualquier caso, debido al parecido entre el checo y el español en términos de desinencias verbales, los aprendices checos presentarán menos dificultades a la hora de aplicar la omisión de los pronombres sujeto que los estudiantes anglófonos, el otro grupo de participantes de nuestro estudio, ya que las características analíticas de su lengua obligan a utilizar necesariamente los pronombres personales sujeto en casi todos los contextos imaginables. No obstante, a juzgar por los resultados obtenidos en el análisis tanto de las redacciones de aprendices checos como de las de los estadounidenses, podríamos concluir que el nivel de competencia en español juega un papel fundamental en el dominio del uso/omisión de pronombres personales sujeto. Es cierto que en el caso de los checos se aprecia la asimilación de este fenómeno en etapas más tempranas del

proceso de aprendizaje que en el caso de los estudiantes anglófonos por motivos de analogía en cuanto a la flexión verbal, como ya hemos explicado. Ahora bien, resulta evidente también que este nivel de control no es del todo estable hasta que los aprendices checos no han alcanzado un cierto estadio en su desarrollo lingüístico y proceso de adquisición; de hecho, la mayoría de errores por presencia indebida de las composiciones checas se hallan en aquellas pertenecientes a niveles iniciales, aunque también se produzcan omisiones correctas, mientras que en niveles intermedios o avanzados hay un mayor grado de coherencia y dominio en cuanto al uso o elisión de los pronombres sujeto. Por tanto, en nuestra opinión, el grado de competencia en español explica el volumen de fallos y aciertos respecto al uso explícito o tácito de los pronombres sujeto, puesto que se aprecia una clara tendencia hacia la mejora a medida que se avanza en el conocimiento y control del idioma.

Por otro lado, cabe recordar que, en la tarea confeccionada específicamente para nuestra investigación, que se difundió entre la comunidad de estudiantes de español de la Universidad de Bohemia del Sur, pero de la que desgraciadamente solo obtuvimos cuatro respuestas, incluimos una pregunta vinculada a los materiales que cada informante habría utilizado a lo largo del proceso de redacción. Esta consulta tenía un doble propósito: por una parte, indagar en el desarrollo expresivo de los estudiantes para descubrir si empleaban en algún momento corpus de aprendices o corpus en general como fuente de información y consulta, y, por otra, averiguar si alguno de los materiales de apoyo utilizados por los alumnos podía resultar de especial ayuda para facilitar el correcto uso/omisión de los pronombres personales sujeto. Lo ideal hubiera sido recibir más respuestas a esta pregunta, pero dadas las circunstancias y el escaso número de participantes que se prestaron a hacer nuestra tarea, hubimos de conformarnos con cuatro, que exponemos a continuación.

1) Materiales que he utilizado:
   - https://dle.rae.es/
   - https://www.linguee.com/
2) Para este texto utilicé el diccionario checo Seznam Slovník (https://slovnik.seznam.cz/)
3) Cuando estaba escribiendo esta reacción, no utilicé ningunos materiales de escuela (libros). Ha utilizado solamente diccionarios online (https://slovnik.seznam.cz/ y https://www.linguee.com/) para asegurarme que algunas desidencias de los verbos o construcciones preposicionales están en forma correcta

4) He utilizado el diccionario online Lingea y he tomado datos importantes de Wikipedia.

Como se puede comprobar, los recursos empleados principalmente por los informantes checos que realizaron nuestra tarea de redacción concreta son diccionarios españoles y checo en línea, así como traductores en línea. Ninguno de ellos se sirvió de corpus como base de datos de consulta lingüística ni tampoco de materiales específicos que pudieran ser de particular utilidad en lo relativo al correcto empleo o elisión de los pronombres sujeto. Este hecho invita a la reflexión y a la toma de conciencia sobre la realidad del aula de idiomas: los aprendices checos rara vez son expuestos a los corpus generales o de aprendices como herramienta de aprendizaje, al menos de manera directa. A nuestro juicio, este "estado de cosas" en la didáctica de lenguas extranjeras debe cambiar para fomentar el progreso de la disciplina y los avances significativos y beneficiosos respecto a los materiales, técnicas y enfoques que se emplean y aplican en el proceso de aprendizaje de un idioma. Para ello, es fundamental que tanto docentes como alumnos se familiaricen, en sus respectivos contextos formativos, con recursos como los corpus, cuyo auténtico potencial pedagógico está aún por explotar.

Por último, para finalizar con nuestras conclusiones, nos gustaría aclarar que somos conscientes de las limitaciones de nuestro estudio en términos metodológicos, especialmente en lo relativo a la desigualdad cuantitativa de los corpus analizados, a la disparidad temática de las composiciones y a las diferencias en el proceso de recopilación de datos de uno y otro corpus (longitudinal por un lado, transversal por otro). Es cierto que la naturaleza transversal de la compilación textual en el caso de los checoparlantes nos impiden realizar un estudio longitudinal, que resultaría mucho más provechoso para la investigación en el ámbito de ELE. No obstante, consideramos que un estudio de tales características requeriría no solo de dos corpus realmente equiparables en términos tanto cualitativos como cuantitativos, sino también de mucha más dedicación y tiempo. Asimismo, hemos de recordar que no contamos con información adicional o metadatos relativos a los informantes, ni estadounidenses ni checos, por lo que no podemos saber a ciencia cierta si, a lo largo de nuestro proceso de corrección y etiquetado de textos para el COWS-L2H, hemos analizado varias composiciones procedentes de un mismo aprendiz o si todas ellas pertenecen a estudiantes diferentes, ya que no hemos revisado el conjunto total de producciones escritas compiladas en este corpus estadounidense. En consecuencia, podríamos deducir que ambos análisis textuales, el de checoparlantes y el

de angloparlantes, presentan un carácter en principio transversal que, aunque menos provechoso, puede igualmente aportar beneficios y material provechoso para la enseñanza de español, y, sobre todo, una fuente de inspiración y una base considerablemente sólida para futuras investigaciones (quizá longitudinales, quizá sucesoras de la que hemos llevado a cabo) que profundicen y desarrollen en más detalle el tema del uso de los pronombres personales sujeto por parte de aprendices de español de diversas nacionalidades. Asimismo, a partir de este estudio contrastivo demostramos el potencial y utilidad que pueden tener los corpus de aprendices de español en el progreso de la didáctica de lenguas, que constituye el fin último y principal del presente trabajo.

5. Consideraciones finales

A lo largo del presente trabajo, hemos trazado la línea cronológica de la evolución de la lingüística de corpus y su aplicación al estudio de las lenguas. Pese a tratarse de una metodología considerablemente reciente, lo cierto es que sigue una tendencia de progreso indudablemente ascendente e imparable, con una prospectiva de futuro muy prometedora. Ahora bien, para poder aprovechar realmente el inmenso potencial que la lingüística de corpus ofrece, hemos de compilar y diseñar corpus útiles, pertinentes, válidos y de calidad. Para ello, deben cumplir con los requisitos expuestos por Sinclair (2005), relativos al carácter comunicativo del contenido, la representatividad, el contraste y la verificación de datos, la estructura intuitiva y coherente, el etiquetado de datos lingüísticos, el carácter completo de las muestras, la documentación del proceso de compilación, el equilibrio, la naturaleza general de la temática, y la homogeneidad del contenido. Ha sido precisamente la aplicación de todos estos criterios a la hora de recopilar y construir un corpus la que ha conducido a la situación actual: un amplio abanico de corpus de distintas características y con un brillante futuro de posibilidades de desarrollo y aplicación.

Por tanto, en la actualidad, tenemos acceso a una inmensa diversidad tipológica en lo que a corpus se refiere gracias a estos procesos de "estandarización" o "perfeccionamiento" en la elaboración de bases de datos lingüísticos: desde divisiones tan simples como corpus orales o escritos hasta dicotomías entre corpus sincrónicos o

diacrónicos, pasando por corpus generales o especializados, representativos e inclusivos en términos de variedades lingüísticas o no, monolingües o multilingües, de nativos o de aprendices, etc. Esta realidad resulta cuando menos alentadora, ya que solo con compendios textuales que respondan a criterios tan variados podremos conseguir que la metodología de la lingüística de corpus avance y se desarrolle de manera realmente provechosa y favorable en lo que respecta a los diferentes ámbitos del estudio de las lenguas que se pueden beneficiar de todas las posibilidades de explotación que ofrecen estos recursos.

Uno de esos campos de aplicación es, sin duda, la didáctica. En este sentido, cualquier tipo de corpus puede resultar de gran ayuda a la hora de estudiar la lengua que se pretende enseñar, por lo que la utilidad de estas herramientas es evidente tanto para elaborar los programas y planes curriculares como para confeccionar materiales didácticos (actividades, apuntes, ejemplos…) y para basar las correcciones en datos sólidos y fiables. No obstante, de entre el conjunto de corpus existentes en la actualidad que pueden tener una aplicación en contextos pedagógicos, hemos analizado el potencial y utilidades de un tipo de base de datos lingüísticos muy concreta: los corpus de aprendices, aquellos elaborados a partir de tareas de redacción realizadas de manera espontánea o guiada, con un límite de tiempo más o menos restringido y con acceso a materiales de consulta o no, por estudiantes de una lengua extranjera (a menudo con diferentes niveles de dominio del idioma). Estos corpus resultan fundamentales en la enseñanza de segundas lenguas por una razón esencial: permiten analizar la interlengua de los aprendices. Teniendo en cuenta el ambiente multicultural característico de un aula de idiomas, el trabajo con corpus de aprendices puede facilitar mucho la tarea del docente, pues le permitirá conocer las dificultades de cada grupo concreto de hablantes, establecidos en función de determinados criterios (habitualmente, la lengua materna), para poder enfocar su metodología pedagógica a las necesidades de los alumnos. Por otro lado, desde una perspectiva más específica, un corpus de aprendices también permite conocer la influencia que tienen en el nivel de competencia de un estudiante el contexto formativo, su contacto con otras lenguas, su condición de bilingüe… En suma, a la hora de afrontar la enseñanza de un idioma, debemos saber cómo funciona la lengua meta, por supuesto, pero también sería recomendable y beneficioso que fuéramos conscientes del uso que nuestros alumnos hacen de ella, pues no siempre (de hecho, pocas veces) es directamente equiparable al uso nativo. En este sentido, un corpus de aprendices nos permite analizar cómo es la lengua de los estudiantes en un estadio concreto de su

aprendizaje, por lo que constituye una herramienta imprescindible para todo profesor de idiomas.

Desde el punto de vista de la lengua española, cuya didáctica es el tema que nos ocupa en el presente trabajo, el número de corpus de aprendices existente hasta la fecha se ha incrementado considerablemente en las últimas décadas, pero sigue tratándose de una cifra abarcable[17]. Asimismo, pese a la satisfactoria diversidad tipológica que presenta el conjunto de corpus de aprendices con el que contamos en español, aún queda mucho camino por recorrer, tanto en términos cuantitativos como cualitativos, estos últimos vinculados especialmente a la variedad (es necesario realizar más compilaciones de carácter longitudinal, aplicar una perspectiva inclusiva/panhispánica más a menudo, analizar las dificultades en función de un número mayor de lenguas maternas…), al diseño (interfaces más intuitivas, más anotación lingüística apropiada y útil, más metadatos sobre los informantes…) y a la difusión (mayor exposición de docentes y aprendices a estos recursos). Para fomentar el correcto y próspero desarrollo de los corpus de aprendices de español e incentivar la familiarización de la comunidad académica con este tipo de herramientas, hemos de incidir en sus múltiples utilidades. En este sentido, las posibilidades que nos ofrecen los corpus de aprendices de estudiar la interlengua de los aprendices para determinar sus ventajas y dificultades de cara a aprender un nuevo idioma como el español multiplica su potencial de explotación, puesto que podríamos realizar numerosos estudios contrastivos entre el idioma que se pretende enseñar y la(s) lengua(s) materna(s) que incluya el corpus, o entre diversas lenguas maternas, en términos de facilidades/obstáculos de aprendizaje. Por tanto, este tipo de herramientas promueven no solo el estudio y conocimiento de la interlengua de los estudiantes, sino también el desarrollo de un enfoque pedagógico basado en la lingüística contrastiva, una disciplina que puede revelarse muy útil desde el punto de vista del estudio y/o enseñanza de lenguas.

Para demostrar la utilidad de los corpus de aprendices en el ámbito de la didáctica de idiomas, y como muestra de una de las principales aplicaciones de estos recursos en los ámbitos de la adquisición y la enseñanza de lenguas, hemos desarrollado en este trabajo un pequeño estudio contrastivo a partir de la comparación entre las interlenguas de aprendices de dos idiomas maternos diferentes (inglés y checo) entre sí y con respecto a una misma lengua extranjera, en cuyo proceso de aprendizaje aún están inmersos (español). El fenómeno examinado y contrastado en estos tres idiomas ha sido el uso o,

[17] Véase *Anexo I: Tabla de clasificación de corpus de aprendices de español* (páginas 113-115)

más bien, la omisión del pronombre personal sujeto, una práctica muy habitual en español que suele suponer un problema para determinados grupos de alumnos. Nuestra participación como revisoras de una fracción de un corpus de aprendices de español desarrollado por la Universidad de California Davis llamado COWS-L2H, a través de un proyecto de colaboración con la Universidad de Salamanca, nos dios acceso a una gran cantidad de datos extraídos de redacciones de estudiantes angloparlantes de español. Asimismo, esta experiencia nos inspiró para seleccionar el fenómeno gramatical que estudiaríamos de manera contrastiva con el checo y el español, respectivamente, ya que la presencia/ausencia indebida de pronombres personales sujeto era uno de los errores que, como revisoras, debíamos corregir y etiquetar en las composiciones de los estadounidenses. Por consiguiente, tomando el modelo del corpus de aprendices COWS-L2H, compilamos uno "semejante" con informantes checos (de tamaño mucho menor, pero cualitativamente equiparable) que nos permitiera contraponer las estrategias propias de la interlengua de aprendices angloparlantes y checoparlantes en cuanto al uso/omisión del pronombre sujeto. Tras establecer las comparaciones y contrastes necesarios inglés-checo e inglés-checo-español, llegamos a la conclusión de que este fenómeno característico de la gramática española supone un mayor obstáculo para estudiantes de lengua materna inglés que para los checoparlantes, puesto que estos últimos presentan tendencias y prácticas similares en su idioma a raíz de la riqueza de su flexión verbal, de la que el inglés carece.

El mayor conocimiento de la interlengua de diversos aprendices a través de estudios como el que hemos llevado a cabo en este trabajo facilita el tratamiento de distintos temas gramaticales "conflictivos" en el aula de idiomas, ya que el profesor podrá adaptar su metodología docente de manera informada, justificada y efectiva a las necesidades y características particulares del grupo meta (en nuestro caso, formado por angloparlantes y checoparlantes). En este sentido, gracias a investigaciones de corte contrastivo, a partir de datos compilados en corpus de aprendices, conocerá de primera mano los obstáculos a los que se tendrá que enfrentar él como instructor, por un lado, y sus alumnos como aprendices de una lengua extranjera, por otro. Asimismo, dado que será más consciente de los puntos fuertes y débiles de cada grupo de estudiantes de su clase en términos de asimilación lingüística, podrá fomentar una influencia positiva de unos estudiantes sobre otros, lo que le permitirá crear un ambiente de estudio y motivación favorables, estimulando la empatía intercultural en el aula, el trabajo en equipo y el aprendizaje no solo autónomo, sino también y sobre todo en grupo.

En definitiva, como bien afirma Guillermo Rojo Sánchez (2016), "los corpus son recursos básicos que, además de la explotación directa, constituyen el punto de partida de numerosas aplicaciones mediante las adaptaciones y refinamientos necesarios" (p. 295). No obstante, para que tanto los lingüistas como los profesores y aprendices puedan aprovechar todo su potencial, es necesario fomentar la exposición a estos recursos tanto en la formación de los docentes como en el contexto pedagógico de un aula de idiomas. Los procesos de instrucción del profesorado habituales y convencionales tienden a excluir estas herramientas del plan de estudios. La razón fundamental es la escasez de materiales disponibles que analicen y expliquen su uso, y, en consecuencia, la carencia de conocimientos necesarios en los formadores para abordar la instrucción técnica, lingüística y metodológica relativa al proceso de compilación, diseño y empleo de corpus de aprendices. Por tanto, si los propios docentes desconocen sus utilidades, resulta evidente que la exposición de los estudiantes a este tipo de recursos lingüísticos será mínima. Esto supone un serio impedimento no solo en el desarrollo de corpus de aprendices de español y de otros idiomas, sino en la evolución de la lingüística de corpus, de la didáctica de idiomas y del estudio de las lenguas en general. En esta línea, la falta o ausencia de competencia, de experiencia y de formación no resultan una excusa convincente, puesto que la desconfianza y el temor al desconocimiento constituye precisamente el principal obstáculo para el progreso. Y para conseguir avanzar, hemos precisamente de atrevernos a conocer y explorar lo desconocido.

Nos gustaría abogar en este espacio, por tanto, por la osadía. La osadía de utilizar un recurso totalmente novedoso con el que no estamos familiarizados, de analizar sus posibles usos y aplicaciones, de explotar todo su potencial creando materiales a partir de su contenido, utilizando sus datos como ejemplos en explicaciones teóricas, y exponiendo directamente al alumno a sus propios aciertos y errores, o a los aciertos y errores de otros estudiantes semejantes a él o totalmente diferentes. La osadía de estudiar el funcionamiento y utilidad de los corpus de aprendices como herramientas lingüísticas que pueden enriquecer nuestra formación como profesores y la de las futuras generaciones de docentes. La osadía de descubrir, de progresar, de educarnos, de formarnos… de enseñar nuestra propia lengua y sorprendernos al percatarnos de que precisamente el docente es el que más tiene que aprender.

6. Referencias bibliográficas

Abad Castelló, M. (2019). Uso de corpus lingüísticos por y para profesores de español como lengua extranjera. *RedELE. Revista Electrónica de Didáctica Del Español Lengua Extranjera (ELE)*, 31, 148–167. Retrieved January 2021, from https://dialnet.unirioja.es/servlet/articulo?codigo=7484417

Alonso Pérez-Ávila, E. (2007). El corpus lingüístico en la didáctica del léxico del español como LE. *Boletín de La Asociación Para La Enseñanza Del Español Como Lengua Extranjera*, 37, 11–27. Retrieved January 2021, from http://www.cervantesvirtual.com/obra/num-37-noviembre-de-2007/

Alonso-Ramos, M. (2016). Spanish learner corpus research: Achievements and challenges. In M. Alonso-Ramos (Ed.), *Spanish Learner Corpus Research: Current trends and future perspectives* (pp. 3–31). Retrieved January 2021, from https://books.google.es/books?hl=es&lr=&id=h_GiDQAAQBAJ&oi=fnd&pg=PR1&dq=spanish+learner+corpora&ots=0s_O0ZX6cO&sig=0oREkAU4PMc0dnA6sB5Y7FlOj2k#v=onepage&q=spanish learner corpora&f=false

Barroso Jiménez, M. F. (2011). Desarrollar la competencia léxica a través de la lingüística del corpus. *Mediterráneo. Revista de La Consejería de Educación En Italia, Grecia y Albania*, 3, 8–15. Retrieved February 2021, from https://dialnet.unirioja.es/servlet/articulo?codigo=5247722

Biber, D., Conrad, S., & Reppen, R. (1998). Corpus Linguistics: Investigating Language Structure and Use. In *TESOL Quarterly* (Vol. 32). Retrieved January 2021, from https://books.google.cz/books?id=vMfLCgAAQBAJ&printsec=frontcover&hl=es#v=onepage&q&f=false

Buyse, K., & González Melón, E. (2013). El corpus de aprendices Aprescrilov y su utilidad para la didáctica de ELE en la Bélgica multilingüe. In S. Borrell, B. Blecua Falgueras, B. Crous, & F. Sierra (Eds.), *Actas del XXIII congreso internacional de ASELE. Plurilingüismo y enseñanza de ELE en contextos multiculturales* (pp. 247–252). Retrieved January 2021, from http://cvc.cervantes.es/ensenanza/biblioteca_ele/asele/pdf/23/23_0025.pdf

Buyse, K., Fernández Pereda, L., & Verveckken, K. (2015). The reference to L1 and L2 in SFL: proposals based on the Aprescrilov learner corpus. *Procedia: Social and Behavioral Sciences*, *173*, 274–278. Retrieved January 2021, from https://www.sciencedirect.com/science/article/pii/S1877042815013749

Buyse, K., Fernández Pereda, L., & Verveckken, K. (2016). The Aprescrilov corpus, or broadening the horizon of Spanish laguage learning in Flanders. In M. Alonso-Ramos (Ed.), *Spanish learner corpus research: current trends and future perspectives* (pp. 143–168). Retrieved January 2021, from https://books.google.cz/books?id=h_GiDQAAQBAJ&pg=PA143&lpg=PA143&dq=The+Aprescrilov+corpus,+or+broadening+the+horizon+of+Spanish+language+learning+in+Flanders&source=bl&ots=0s-LX1199O&sig=ACfU3U34qH4Yk-VI21_xN8SR2OSPyX0QxQ&hl=es&sa=X&ved=2ahUKEwilo76gubrwAhVSi8MKHVqND7QQ6AEwB3oECAUQAw#v=onepage&q=The Aprescrilov corpus%2C or broadening the horizon of Spanish language learning in Flanders&f=false

Calero Fernández, M. Á., Serrano Zapata, M., & Gómez-Devís, M. B. (2020). Codificación y etiquetado en los corpus de aprendices y su aplicación didáctica: la propuesta del corpus de interlengua española de aprendices sinohablantes (CINEAS). *E-Aesla. Revista Digital de Lingüística Aplicada.*, 6, 206–222. Retrieved January 2021, from https://cvc.cervantes.es/lengua/eaesla/pdf/06/15.pdf

Campillos Llanos, L. (2014). A Spanish Oral Learner Corpus for Computer-Aided Error Analysis. *Corpora*, *9*(2), 207–238. Retrieved January 2021, from https://www.researchgate.net/publication/266477960_A_Spanish_learner_oral_corpus_for_computer-aided_error_analysis

Centro Virtual Cervantes. (n.d.). Diccionario de término clave de ELE. Retrieved January 2021, from Instituto Cervantes website: https://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/indice.htm

Corpus Oral de Español como Lengua Extranjera. (n.d.). Retrieved January 2021, from Laboratorio de Lingüística Informática website: http://www.lllf.uam.es/ESP/CORELE.html

Davidson, S., & Sagae, K. (n.d.). COWSL2H. The UC Davis Corpus of Written Spanish, L2 and Heritage Speakers. Retrieved November 2020, from GitHub website: https://github.com/ucdaviscl/cowsl2h

Ferreira Cabrera, A. (2018). Errores sistemáticos en un corpus de aprendices de ELE con implicaciones para el feedback correctivo escrito enfocado. In M. Bargalló Escrivá, E. Forgas Berdet, & A. Nomdedeu Rull (Eds.), *Actas del XXVIII congreso internacional de ASELE. Léxico y cultura en LE/L2: corpus y diccionarios* (pp. 217–227). Retrieved February 2021, from https://cvc.cervantes.es/ensenanza/biblioteca_ele/asele/pdf/28/28_0019.pdf

Ferreira Cabrera, A., & Elejalde Gómez, J. (2017). Análisis de errores recurrentes en el Corpus de Aprendices de Español como Lengua Extranjera, CAELE. *Revista Brasileira de Lingüustica Aplicada (RBLA)*, *17*(3), 509–537. Retrieved February 2021, from https://www.scielo.br/pdf/rbla/v17n3/1984-6398-rbla-201710927.pdf

Granger, S. (2002). A Bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 3–33). Retrieved January 2021, from https://books.google.cz/books?id=SgEtnIOdC5kC&printsec=frontcover&hl=es#v=onepage&q&f=false

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching. A critical evaluation. In K. Aijmer (Ed.), *Corpora and Language Teaching* (pp. 13–32). Retrieved January 2021, from https://books.google.es/books?hl=es&lr=&id=3yfOJSMTRhoC&oi=fnd&pg=PA13&dq=spanish+learner+corpora&ots=wubK2mT0sE&sig=vjie7SgpHZBJUhlcx7CcbQUskZ0#v=onepage&q=spanish learner corpora&f=false

Hincapié, D. (2018). Corpus de aprendientes de español como lengua extranjera y segunda lengua (CAELE/2): el componente escrito. *Forma y Función*, *31*(2), 129–143. Retrieved February 2021, from https://rael.aesla.org.es/index.php/RAEL/article/view/274

Instituto Cervantes. (2020). Corpus de aprendices de español (CAES). Retrieved January 2021, from Instituto Cervantes website: https://galvan.usc.es/caes

Laboratoire Lattice. Langues, Textes, Traitements informatiques, Cognition. (n.d.). Retrieved February 2021, from https://www.lattice.cnrs.fr/

Labrador de la Cruz, M. B. (1997). La lingüística de corpus en un estudio contrastivo inglés-español. *Interlingüística*, 6, 63–66. Retrieved February 2021, from https://dialnet.unirioja.es/servlet/articulo?codigo=900645

Leech, G. (1993). Corpus Annotation Schemes. *Literary and Linguistic Computing*, *8*(4), 275–281.

Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 1–23). Retrieved January 2021, from https://books.google.cz/books?id=UAHKAwAAQBAJ&printsec=frontcover&hl=es#v=onepage&q&f=false

Lozano, C. (2009). CEDEL2: Corpus Escrito del Español como L2. In C. M. Bretones Callejas, J. F. Fernán Sánchez, J. R. Ibáñez Ibáñez, M. E. García Sánchez, M. E. Ríos Cortés de los, S. Salaberri Ramiro, … B. Cantizano Márquez (Eds.), *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada actual: Comprendiendo el lenguaje y la mente* (pp. 197–212). Retrieved January 2021, from https://www.researchgate.net/publication/292981987_CEDEL2_Corpus_Escrito_del_Espanol_como_L2

Luján, M. (1999). Expresión y omisión del pronombre personal. In I. Bosque Muñoz & V. Demonte Barreto (Eds.), *Gramática descriptiva de la lengua española. Sintaxis básica de las clases de palabras* (pp. 1275–1315). Madrid: Espasa.

Martín Sánchez, T., Pascual Escagedo, C., & Puigdevall Bafaluy, N. (2017). CORINÉI: una herramienta innovadora para el análisis contrastivo español-italiano. In M. V. Calvi, B. Hernán-Gómez Prieto, & E. Landone (Eds.), *El español y su dinamismo: redes, irradiaciones y confluencias* (pp. 237–257). Retrieved February 2021, from https://cvc.cervantes.es/literatura/aispi/pdf/bib_01/01_237.pdf

Mas Álvarez, I., & Gil Martínez, A. (2018). Los corpus de aprendices: un terreno en expansión para la enseñanza de español. In M. Ellison, M. Pazos Anido, P. Nicolás Martínez, & S. Valente Rodriguez (Eds.), *As línguas estrangeiras no ensino superior: propostas didáticas e casos em estudo* (pp. 35–55). Retrieved January 2021, from https://ler.letras.up.pt/uploads/ficheiros/18192.pdf

Mendikoetxea, A. (2014). Corpus-Based research in Second Language Spanish. In K. L. Geeslin (Ed.), *The Handbook of Spanish Second Language Acquisition* (pp. 11–29). Retrieved January 2021, from https://books.google.cz/books?id=7chcDwAAQBAJ&printsec=frontcover&hl=es#v=onepage&q&f=false

Núñez Nogueroles, E. E. (2020). Pasado, presente y futuro de los corpus de aprendices de ELE. Una revisión bibliográfica. *ReiDoCrea: Revista Electrónica de Investigación Docencia Creativa*, *8*(3), 170–190. Retrieved January 2021, from https://www.ugr.es/~reidocrea/8.3-10.pdf

Palacios Martínez, I. M., & Sampedro Mella, M. (2018). *Los corpus de aprendices y sus aplicaciones. El Corpus de Aprendices del Español (CAES)*. Retrieved January 2021, from https://gramatica.usc.es/~raquel.rivas/corpustextuais/CAES curso verano USC.pdf

Parodi, C. (2015). Reseña del Corpus de aprendices de español (CAES). *Journal of Spanish Language Teaching*, *2*(2), 194–200. Retrieved January 2021, from https://www.tandfonline.com/doi/pdf/10.1080/23247797.2015.1084685

Publicaciones. (n.d.). Retrieved January 2021, from Teletándem - CORINÉI website: https://dti.ua.es/es/teletandem-corinei/publicaciones/publicaciones.html

Real Academia Española, & ASALE. (2009-2011). *Nueva Gramática de la Lengua Española*. Retrieved February 2021, from http://aplica.rae.es/grweb/cgi-bin/buscar.cgi

Rodríguez Muñoz, F. J., & Ruiz Domínguez, M. del M. (2017). Los operadores discursivos de concreción o especificación y de refuerzo argumentativo en el Corpus de aprendices de español como lengua extranjera. *RAEL: Revista*

Electrónica de Lingüística Aplicada, *15*(1), 53–69. Retrieved February 2021, from https://rael.aesla.org.es/index.php/RAEL/article/view/274

Rojo Sánchez, G. (2016). Corpus textuales del español. In J. Gutiérrez-Rexach (Ed.), *Enciclopedia de Lingüística Hispánica* (pp. 285–296). Retrieved January 2021, from https://dialnet.unirioja.es/servlet/extart?codigo=5589798

Rojo, G., & Palacios, I. M. (2016). Learner Spanish on Computer. The CAES "Corpus de Aprendices de Español" project. In M. Alonso-Ramos (Ed.), *Spanish Learner Corpus Research: Current trends and future perspectives* (pp. 55–87). Retrieved January 2021, from https://books.google.es/books?hl=es&lr=&id=h_GiDQAAQBAJ&oi=fnd&pg=PA55&dq=spanish+learner+corpora&ots=0s_O0ZX96U&sig=ib_Dgvqi3PUsgxI0vb5NkVvsZBQ#v=onepage&q=spanish learner corpora&f=false

Sierra, G., Bel, G., & Lázaro Hernández, J. A. (2018). Tipología y clasificación de corpus. Retrieved February 2021, from Lingüística de corpus - Procesamiento de corpus textuales y orales website: http://www.corpus.unam.mx/cursocorpus/1_2_Clasificacion.html#:~:text=Se puede hablar de dos,orales sonoros y orales transcritos

Sinclair, J. (1996). *EAGLES: Preliminary recommendations on Corpus Typology*. Retrieved January 2021, from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.28.1988&rep=rep1&type=pdf

Sinclair, J. (2005). Corpus and text – Basic principles. In M. Wynne (Ed.), *Developing Linguistic Copora: A guide to good practice* (pp. 1–16). Retrieved January 2021, from http://users.ox.ac.uk/~martinw/dlc/chapter1.htm

SPLLOC Conference Presentations & Publications. (n.d.). Retrieved January 2021, from Spanish Learner Language Oral Corpora. Linguistic development in L2 Spanish website: http://www.splloc.soton.ac.uk/publication.html

Studies that have used CEDEL2. (n.d.). Retrieved January 2021, from CEDEL2: Corpus Escrito del Español L2 (version 2) website: http://cedel2.learnercorpora.com/about/studies

Tognini Bonelli, E. (2010). Theoretical overview of the evolution of corpus linguistics. In A. O'Keefe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 14–27). Retrieved January 2021, from https://books.google.co.in/books?hl=en&lr=&id=giaMAgAAQBAJ&oi=fnd&pg=PA14&dq=Theoretical+Overview+of+the+Evolution+Of+Corpus+Linguistics&ots=xr6yrr7SfZ&sig=ABdVPyk3Ae0GFR4rdAWt6FiLih8#v=onepage&q=Theoretical Overview of the Evolution Of Corpus Linguistics&f=false

Tono, Y. (2003). Learner corpora: design, development and applications. *Proceedings of the Corpus Linguistics 2003 Conference*, 16, 800–809. Retrieved January 2021, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.115.6849&rep=rep1&type=pdf

Valverde Ibáñez, M. del P. (2018). Un corpus de blogs de aprendices japoneses de español. In M. Bargalló Escrivá, E. Forgas Berdet, & A. Nomdedeu Rull (Eds.), *Actas del XXVIII congreso internacional de ASELE. Léxico y cultura en LE/L2: corpus y diccionarios* (pp. 845–857). Retrieved February 2021, from https://cvc.cervantes.es/ensenanza/biblioteca_ele/asele/pdf/28/28_0078.pdf

Valverde Ibáñez, M. del P. (2020). Diseño y creación de un corpus de aprendices de español en Japón (CELEN). *E-Aesla. Revista Digital de Lingüística Aplicada.*, (6), 223–240. Retrieved February 2021, from https://cvc.cervantes.es/lengua/eaesla/pdf/06/16.pdf

Vyatkina, N., & Boulton, A. (2017). Corpora in language learning and teaching. *Language Learning & Technology, 21*(3), 1–8. Retrieved January 2021, from https://hal.archives-ouvertes.fr/hal-01237582/document

Yamada, A., Davidson, S., Fernández-Mira, P., Carando, A., Sagae, K., & Sánchez-Gutiérrez, C. (2020). COWS-L2H: A corpus of Spanish learner writing. *RiCL. Research in Corpus Linguistics*, *8*(1), 17–32. Retrieved November 2020, from https://ricl.aelinco.es/index.php/ricl/article/view/109

Anexo I:

Tabla de clasificación de corpus de aprendices de español

Nombre del corpus	Institución y compiladores principales	Modalidad de corpus	L1 de los participantes	Sexo de los participantes	Edad de los participantes	Niveles de competencia de los participantes	Tamaño	Fecha de compilación	Tipos de textos	Objetivos	datos en términos de tiempo	Acceso libre o restringido	Metadatos sobre informantes y muestras	Anotación	Otros datos de interés
Corpus Oral de Español como Lengua Extranjera (CORELE)	Leonardo Campillos Llanos de la Universidad Autónoma de Madrid. Proyecto subvencionado por la Consejería de Madrid y el Fondo Social Europeo	Oral	Portugués, italiano, francés, inglés, neerlandés, alemán, polaco, chino, japonés, coreano, finés, húngaro, turco y español	12 hombres y 28 mujeres	Entre 20 y 25 años	A2-B1	55 567 palabras (13 horas y 36 minutos de grabaciones)	En algún momento de la primera década del siglo XXI	Entrevistas (diálogo semiespontáneo investigador-informante), descripciones de fotografías y relatos.	Análisis de errores fonéticos, léxicos, gramaticales y pragmáticos en la producción oral en español de informantes extranjeros. Análisis de la interlengua y de procesos de etiquetado para la investigación a partir de corpus. Creación de una guía y materiales pedagógicos para explicar a los docentes los errores propios de cada conjunto de aprendices y orientarles sobre su corrección.	Transversal	Acceso libre en http://cartago.lllf.uam.es/corele/index.html	Total de informantes: 40. Estudiantes procedentes del programa Erasmus u otros acuerdos de estudios en el extranjero similares. Por otro lado, estudiantes del Servicio de Idiomas de la Universidad Autónoma de Madrid o de los cursos de español de la Universidad Complutense de Madrid.	Anotación lingüística (etiquetas gramaticales y de errores)	Incluye un corpus de control de hablantes nativos (9 389 palabras y 1 hora y 22 minutos de grabación).
Corpus Oral de Interlengua Bilingüe Español-Italiano (CORINÉI)	Carmen Gonzáez Royo y Stefania Chiapello de la Universidad de Alicante. Colaboración con las Universidades de Salerno y S.Orsola Benincasa de Nápoles en el Proyecto *Teletándem*	Oral	Italiano y español	Hombres y mujeres (predominio de mujeres)	Entre 20 y 35 años	Italianos: B1-C1 / Españoles: A1-B2	268 conversaciones de las universidades italianas, 225 conversaciones de la Universidad de Alicante (total de 123 horas de grabación aprox.)	2014-2015	Grabaciones de interacción real por Skype entre estudiantes españoles e italianos con autotranscripciones	Análisis (contrastivo o no) de las estrategias y fenómenos conversacionales empleadas por los estudiantes italianos y españoles en la interacción nativo/no nativo. Análisis de la evolución de su interlengua, estudio de la autotranscripción de las conversaciones para evaluar errores lingüísticos y pragmáticos, investigación de estructuras lingüísticas y pragmáticas de las conversaciones en estos contextos.	Longitudinal	Acceso restringido	Total de informantes: 76 alumnos de las universidades italianas, 132 de la Universidad de Alicante. Estudiantes de Traducción e Interpretación o de Lenguas y Culturas Extranjeras. Fueron grabados en diferentes momentos a lo largo de un año académico.	Metadatos sobre informantes y muestras	Está en proceso de creación (en fase de publicación en línea en 2017).
Corpus vídeo del español hablado en Texas (SpinTX)	Almeida Jacqueline Toribio y Barbara E. Bullock de la Universidad de Austin (Texas). Proyecto desarrollado en el *Centre for Open Educational Resources and Language Learning* (COERLL)	Oral	Español e inglés	Hombres y mujeres	Entre 18 y 65 años (e incluso mayores)	Diferentes niveles	500 000 palabras	2013	Entrevistas y conversaciones sobre 18 temas diferentes con transcripciones	Análisis y descripción de la variedad del español que se utiliza en el estado de Texas con vistas a configurar herramientas que faciliten el aprendizaje del idioma a estudiantes y profesores.	Transversal	Acceso libre en https://www.coerll.utexas.edu/spintx/	Total de informantes: 97. Bilingües español-inglés.	Anotación lingüística (etiquetas gramaticales)	Aún está en proceso de creación.
Spanish Learner Language Oral Corpus I and II (SPLLOC)	Laura Domínguez de la Universidad de Southampton. Colaboración entre las Universidades de Southampton, York y Newcastle	Oral	Inglés y español	Hombres y mujeres (predominio de mujeres)	Entre 13 y 22 años	A2-C2	50 000 palabras aprox.	2008-2010	Entrevistas, descripciones de imágenes, debates y relatos con transcripciones	Estudio de la adquisición del español como segunda lengua a través de dos proyectos independientes derivados de este corpus: SPLLOC1, que estudia la adquisición de la morfosintaxis española (orden de las palabras, clíticos, etc.) y el desarrollo de la riqueza léxica, y SPLLOC2, que analiza el desarrollo de cuestiones aspectuales y temporales de los verbos españoles.	Transversal	Acceso libre en http://www.splloc.soton.ac.uk/	Total de informantes: 120. Procedentes de escuelas y universidades de Reino Unido. Hablantes de lengua materna inglés, fundamentalmente, y español.	Anotación lingüística (etiquetas morfosintácticas)	Se trata del corpus oral del español abierto al público más grande hasta la fecha. Incluye un corpus de control de hablantes nativos. El proyecto SPLLOC1 ya está terminado, pero SPLLOC2 está en proceso de creación aún.
The Spanish Corpus Proficiency Level Training (SPT)	Dale Koike de la Universidad de Austin (Texas). Proyecto desarrollado en el *Centre for Open Educational Resources and Language Learning* (COERLL)	Oral	Inglés	Hombres y mujeres (predominio de mujeres)	Diversas edades	Todos los niveles de dominio de la lengua (principiante-avanzado)	327 grabaciones	2010-2011	Grabaciones en vídeo con diálogos a partir de una serie de preguntas sobre nueve 9 diferentes y con transcripciones	Formación de profesores y acreditadores de nivel de competencia en español para que sean capaces de valorar el dominio del español. Análisis de diferentes rasgos sobre la interlengua de los informantes y propuestas didácticas (ejercicios) para trabajar sus dificultades.	Transversal	Acceso libre en https://www.laits.utexas.edu/spt/	Total de informantes: 38. Tienen como lengua materna el inglés y la mayoría, el español como lengua de herencia.	Metadatos sobre informantes y muestras	Se trata de un corpus audiovisual. El corpus ofrece ejercicios sobre las muestras con las soluciones.
Fono.ele	Ana Blanco Canales y Mª Ángeles Álvarez Martínez de la Universidad de Alcalá. Equipo Fono.ele	Oral	Polaco, griego, portugués, alemán, taiwanés y egipcio	Hombres y mujeres	Entre 18 y 35 años (divididos en dos grupos de 18 a 24 y de 25 a 35)	A2-C1	34 316 grabaciones	En torno a 2010	Grabaciones de conversaciones breves estructuradas, así como de textos, frases y aplabras leídas en voz alta (029863)	Investigación sobre la adquisición y aprendizaje de la fonética y pronunciación del español como lengua extranjera en relación con cuestiones sociales, culturales y educativas. Desarrollo de materiales de consulta y análisis para mejorar la formación de profesores en este ámbito.	Transversal	Acceso libre en http://www3.uah.es/fonoele/	Total de informantes: 96. De cada informante se proporciona la edad, el sexo, la nacionalidad, el nivel de contacto y de domino del español, así como la experiencia en cuestiones relacionadas con la fonética española.	Anotación lingüística (etiquetas de errores) y metadatos sobre los informantes y sobre las muestras	Contiene datos de producción y de percepción, estos últimos a partir de respuestas a ciertas preguntas de índole fonético-fonológicas.
The Anglia Polytechnic University Spanish Corpus (APU)	Anne Ife y Alicia Peña Calvo de la Universidad Anglia Ruskin (Inglaterra)	Escrito	Alemán, checo, chino, danés, finlandés, griego, inglés, italiano…	Hombres y mujeres	Jóvenes estudiantes (entre 18 y 25 años aprox.)	Todos los niveles de dominio de la lengua (principiante-avanzado)	113 327 palabras	-	Producciones escritas de carácter descritivo y narrativo en tareas de clase y exámenes universitarios	Estudio de diversos aspectos de la interlengua de los aprendices de español.	Transversal	Acceso restringido	-	Sin anotar	-
Aprender a Escribir en Lovaina (Aprescrilov)	Kris Buyse de la Facultad de Letras de la Universidad Católica de Lovaina. Colaboración con la Universidad Lessius Hogeschoool	Escrito	Neerlandés y francés	Hombres y mujeres	Jóvenes estudiantes (entre 18 y 25 años aprox.)	A1-C1	1 000 000 de palabras aprox. en más de 2 700 textos	2005-2011	Tareas de producción escrita y exámenes. Textos expositivos, descriptivos, argumentativos, narrativos y cartas	Estudio de la interlengua de los aprendices y de las interferencias entre su lengua materna o su primera lengua extranjera y el español.	Longitudinal	Acceso restringido (es necesario solicitar una cuenta: https://idp.kuleuven.be/idp/profile/SAML2/POST/SSO?execution=e1s2)	La lengua materna de los aprendices es el neerlandés o, en su defecto, el francés, y su primera lengua extranjera es el neerlandés, el francés o el inglés, lo que convierte al español en su segunda lengua extranjera. Son estudiantes de lengua española matriculados en uno de los tres primeros años de universidad. Cada uno escribió varias redacciones por cuatrimestre académico (tres meses aprox.)	Anotación lingüística (etiquetas gramaticales, léxicas, ortográficas, pragmáticas y de errores)	Es uno de los corpus de aprendices precursores en el ámbito del español. La búsqueda se puede realizar atendiendo a diversos criterios (categoría gramatical, carrera del estudiante, curso, año académico, tarea…). Contiene descripciones cualitativas y cuantitativas de cada elemento para calcular la frecuencia.
Corpus especializado de Aprendientes de Español como Lengua Extranjera (CAELE)	Anita Ferreira Cabrera de la Universidad de Concepción (Chile). Resultado de un proyecto de investigación FON-DECYT denominado *Diseño e implementación de un corpus escrito de aprendientes de Español como Lengua Extranjera (ELE) para el análisis de la interlengua*	Escrito	Alemán, francés, inglés, portugués, sueco, checo, italiano y ruso	Hombres y mujeres (predominio de mujeres)	Promedio de 24 años	A2+ y B2	418 textos	Tres fases entre 2014-2015 (en proceso de ampliación y desarrollo)	Textos de tipo narrativo y descriptivo sobre diversos temas redactados a modo de tareas en clases de ELE	Detección y estudio de los tipos errores más frecuentes y recurrentes en la interlengua de los aprendientes de español a partir de los textos incluidos en el corpus.	Longitudinal	Acceso restringido	Total de informantes : 62. Estudiantes de cursos de español como lengua extranjera en una universidad de Chile. Su nivel de dominio del español se estableció según sus resultado en el examen de Certificación del Español como Lengua Extranjera (CELE). Cada estudiante escribió de cuatro a siete textos en cada fase de recolección. Las producciones escritas se redactaron virtualmente en distintos momentos del semestre sin ningún tipo de apoyo (ni diccionarios ni traductor), con un tiempo máximo de media hora para realizar la tarea.	Anotación lingüística (etiquetas gramaticales, léxicas, ortográficas y de errores)	-
Corpus de aprendices de español (CAES)	Coordinador por Guillermo Rojo e Ignacio Palacios de la Universidad de Santiago de Compostela, subvencionado e impulsado por el Instituto Cervantes	Escrito	Árabe, chino mandarín, francés, inglés, portugués y ruso	Hombres y mujeres (predominio de mujeres)	Entre 15 y 75 años (divididos en cinco grupos de edad de 15 a 21, de 22 a 30, de 31 a 40, de 41 a 60, y de 61 o mayores)	A1-C1	575 000 palabras en 3 878 textos	2011-2013 (accesible en línea desde 2014)	Producciones escritas de tipología muy variada: cartas, postales, emails, reservas de hotel, solicitudes de diversa naturaleza, biografías, narraciones, anécdotas, reclamaciones, descripciones, ensayos y reseñas críticas	Fomentar y facilitar el empleo de datos objetivos y reales en la investigación sobre adquisición y enseñanza-aprendizaje de ELE.	Longitudinal	Acceso libre en https://www.cervantes.es/lengua_y_ensenanza/tecnologia_espanol/caes.htm	Total de informantes: 1 423. Estudiantes matriculados en centros del Instituto Cervantes o en universidades repartidas por todo el planeta. Cada estudiantes escribió dos o tres textos en diferentes momentos de su proceso de aprendizaje a partir de tareas diseñadas siguiendo lo estipulado por el MCER y el Plan Curricular del Instituto Cervantes, y de acuerdo con el modelo de las pruebas de certificación DELE.	Anotación lingüística (etiquetas morfosintácticas)	Es el corpus de aprendices desarrollado más recientemente y con mucho potencial para la investigación en el ámbito de ELE. Pretende convertirse en una referencia de autoridad dentro de los corpus de aprendices de español. La herramienta de búsqueda es muy intuitiva y se pueden realizar consultas atendiendo a diversos parámetros: edad, sexo, lengua materna, nivel de dominio, categoría gramatical, tema...
Corpus de Aprendices Taiwaneses de Español (CATE)	Hui-Chuang Lu de laUniversidad Nacional de Cheng-Kung (Taiwán)	Escrito	Chino	Hombres y mujeres	-	Diferentes niveles	340 000 palabras aprox. en unas 2 000 redacciones	En proceso de creación	Ensayos	Análisis de la interlengua de este grupo concreto de aprendices con el fin de identificar las dificultades principales a las que se enfrentan a la hora de aprender español y desarrollar metodologías y materiales didácticos mejor adaptados a sus necesidades.	Transversal	Acceso restringido (en proceso de creación)	Estudiantes taiwaneses que aprenden español en la universidad.	Anotación lingüística parcial (etiquetas de errores y categorías gramaticales en una parte del corpus)	Está en proceso de creación.
Corpus Escrito del Español L2 (CEDEL2)	Amaya Mendikoetxea de la Universidad Autónoma de Madrid y Cristóbal Lozano de la Universidad de Granada. Resultado del proyecto *Word Order in Second Language Acquisition Corpora* (WOSLAC)	Escrito	Inglés, griego, japonés, portugués, árabe, chino, ruso, italiano, francés, neerlandés, alemán y español	Hombres y mujeres	Diversas edades	A1-C1	1 105 936 palabras	Compilado desde 2006, publicado en 2017 (en proceso de ampliación y desarrollo)	Tareas de producción escrita en línea	Propósito contrastivo inicial: español-inglés y español extranjero (interlengua)-español nativo. Con ello pretendía facilitar el estudio de la hipótesis de la interfaz en el aprendizaje de la gramática española (morfología y la sintaxis) como L2, así como la creación de dos corpus de aprendices (inglés y español) equiparables y útiles para la investigación. Actualmente, constituye una base de datos y de consulta para todo tipo de investigaciones en el ámbito de ELE y es fuente de numerosos ejemplos para la enseñanza.	Transversal	Acceso libre en http://cedel2.learnercorpora.com/	Total de informantes: 4 399. Estudiantes voluntarios de diferentes instituciones (universidades e institutos) y países (EE.UU., Reino Unido, Canadá, Grecia, etc.) que han redactado textos en formato electrónico a partir de 12 temas diferentes relacionados con diversos aspectos de su contexto, su vida y sus opiniones. Tuvieron que realizar un test de nivel de dominio del español (Universidad de Wisconsin, 1998) antes de aportar sus producciones escritas.	Sin anotación lingüística, pero con metadatos sobre los informantes	Se trata de uno de los corpus de aprendices de español más extensos. Se pueden descargar los datos, no solo consultarlos en línea. Incluye un corpus de control (de tamaño mucho menor) de hablantes nativos de español peninsular e hispanoamericanos con un diseño similar. Los textos que contiene el corpus van acompañados de datos adicionales: lugar de redacción, investigación previa, herramientas linguísticas de apoyo utilizadas... Actualmente, está en proceso de ampliar las lenguas maternas que abarca para diversificar los análisis contrastivos y profundizar en el conocimiento e investigación de la interlengua de diferentes aprendices de español.
Corpus de ELE en Japón (CELEN)	Pilar Valverde de la Universidad Kansai Gaidai	Escrito	Japonés	Hombres y mujeres	Jóvenes estudiantes (entre 18 y 25 años aprox.)	A1-B2	140 000 palabras aprox. En 1 840 textos	Compilado desde 2017 (en proceso de ampliación y desarrollo)	Tareas de producción escrita guiadas o semi-guiadas en exámenes, actividades de clase o deberes (en ocasiones, había limitación de tiempo y posibilidad de consultar herramientas de apoyo: Internet, diccionarios...)	Objetivo principal pedagógico: facilitar la aplicación de la lingüística de corpus a los profesores de ELE que trabajan en Japón, aunque también se enfoca a la investigación en el ámbito de ELE en general.	Longitudinal	Acceso restringido en Sketch Engine (es necesario registrarse y enviar una solicitud a la autora)	Total de informantes: 459. Estudiantes universitarios de diversas instituciones japonesas sin conocimientos previos de español antes de empezar sus estudios. Lo habitual es que estudien español porque se habla en muchos lugares del mundo y/o porque tienen interés en la cultura hispana. Cada estudiante redactó varias tareas de producción escrita (entre uno y ocho textos cada uno) a lo largo de un año académico.	Anotación lingüística (etiquetas gramaticales) y metadatos sobre los informantes, los textos y la tarea	Constituye el primer corpus de aprendices dedicado exclusivamente al análisis de la expresión escrita de los aprendices japoneses de español. Se pueden descargar los datos. Aún está en proceso de creación y pretende ampliar sus horizontes hacia otros tipos de textos (blogs, redes sociales...) y una anotación morfosintáctica y de errores más precisa para convertirse en un corpus representativo y de referencia en Japón.
Corpus de Interlengua Española de Aprendices Sinohablantes (CINEAS)	Mª Ángeles Calero Fernández y Maribel Serrano Zapata de la Universidad de Lleida. Resultado del proyecto *Elaboración y catalogación de un corpus de textos escritor en ELE producidos por estudiantes sinohablantes*	Escrito	Chino	Hombres y mujeres	Jóvenes estudiantes (entre 18 y 25 años aprox.)	A1-C1	400 000 palabras aprox. en 4 339 textos	Dos fases de compilación: 2016-2017 y 2019-2020	Producciones escritas de índole descriptiva, narrativa, expositiva y argumentativa redactadas a mano en el aula a partir de una serie de instrucciones	Creación de una fuente de información y consulta fidedigna, precisa y completa que permita desarrollar estudios sobre el español de aprendices sinohablantes y sus problemas en el aprendizaje de este idioma.	Longitudinal	Acceso libre en https://cineas.udl.cat/	Estudiantes universitarios de Filología Hispánica en cuatro universidades chinas y estudiantes chinos estudiando en tres universidades españolas a través de algún tipo de beca de movilidad. Cada uno escribió entre dos y once textos a lo largo de un año académico.	Anotación lingüística (etiquetas de errores) y metadatos sobre los informantes y sobre la tipología textual	Constituye una herramienta que completa el compendio de corpus de aprendices de español al enfocarse en los de origen sinohablante, un vacío que quedaba por cubrir. La búsqueda se puede realizar según criterios lingüísticos y extralingüísticos.

Corpus para el análisis de errores de aprendices de E/LE (CORANE)	Ana Mª Cestero Mancera e Inmacualda Penadés Martínez de la Universidad de Alcalá	Escrito	Alemán, árabe, checo, chino, coreano, danés, esloveno, fang, finés, flamenco, francés, griego, holandés, húngaro, inglés, italiano, japonés, pakistaní, polaco, portugués, ruso, sueco y turco	Hombres y mujeres	Entre 16 y 55 años (divdos en diferentes franjas de edad)	A1-C1	1 091 producciones escritas	2000 (publicado en CD-ROM en 2009)	Textos ensayísticos guiados y redactados dentro y fuera del aula como ejercicio de clase	Análisis de errores y de la interlengua de los aprendientes para favorecer la investigación en la adquisición y la enseñanza del español, así como la creación de materiales didácticos.	Longitudinal	Acceso restringido (se puede adquirir el CD-ROM en que fue publicado)	Total de informantes: 321. Estudiantes de ELE en los Cursos de Lengua y Cultura Españolas para Extranjeros de la Universidad de Alcalá que tuvieron que redactar varios ensayos en diferentes momentos de su proceso de aprendizaje.	Anotación lingüística (etiquetas de errores) y metadatos sobre los informantes y sobre la tarea	Fue el primer corpus de aprendices de español que se creó.
Corpus del español de los italianos (CORESPI)	Sonia Bailini de la Universidad Católica del Sagrado Corazón (Milán)	Escrito	Italiano	Hombres y mujeres	Entre 18 y 40 años	A1-B2	124 648 palabras aprox. en 474 textos	Desde 2008-2009 (en proceso de ampliación y desarrollo)	Interacciones por correo electrónico entre aprendientes y hablantes nativos de español	Análisis de la interlengua en español de los italianos y del interlengua en italiano de los hispanohablantes con vistas a realizar estudios comparativos de lenguas afines.	Longitudinal	Acceso libre en https://corespiycorite.altervista.org/?doing_wp_cron=1581605929.9318449497222900390625	Total de informantes: 90. Escribieron de cinco a veinticinco textos cada uno en un período de 7-8 meses.	Anotación lingüística (etiquetas morfosintácticas)	La búsqueda se puede realizar atendiendo a diversos parámetros: sexo, edad, nivel de dominio del idioma, conocimiento de otras lenguas extranjeras y tipo textual.
Corpus of Written Spanish of L2 and Heritage Speakers (COWS-L2H)	Claudia H. Sánchez Gutiérrez de la Universidad de California en Davis. Proyecto en colaboración con otras universidades	Escrito	Inglés y chino mandarín	Hombres y mujeres	Jóvenes estudiantes (entre 18 y 25 años aprox.)	Diferentes niveles (principiante, intermedio, hablantes de herencia)	1 138 097 palabras aprox. en 3 498 textos	Dos fases de compilación: 2017-2018 y 2018-presente (en proceso de ampliación y desarrollo durante otros cinco años al menos)	Producciones escritas guiadas y redactadas online en torno a cuatro temas bastante amplios/generales (entre 150 y 500 palabras)	Creación de un vasto corpus escrito de aprendices de español que sea longitudinal y representativo de la competencia lingüística de aprendientes de español dentro del contexto específico de las universidades estadounidenses. Compilación de datos procedentes no solo de aprendices de español como lengua extranjera, sino también de hablantes de herencia, para poder analizar también las dificultades y características del aprendizaje de español en este grupo concreto de alumnos.	Longitudinal	Acceso libre en https://github.com/ucdaviscl/cowsl2h	Total de informantes: 1 370. Estudiantes matriculados en asignaturas enfocadas a la enseñanza del español. Cada estudiante tiene que escribir dos redacciones sobre el mismo tema en dos momentos diferentes de cada cuatrimestre (un mes de diferencia entre ambas) y pueden participar en el proyecto en más de un cuatrimestre. Los textos tienen que escribirse sin ningún tipo de material de apoyo y van acompañados de un formulario para extraer datos sobre el contexto lingüístico de cada informante y de una encuesta en la que se autoevalúan.	Anotación lingüística (etiquetas morfosintácticas) y metadatos sobre los informantes	Esta compilación se limita a una única institución universitaria para obtener datos más homogéneos sobre el desarrollo, la metodología y los resultados de un contexto de formación concreto.
Corpus para el estudio de la adquisición del español como lengua extranjera (ESPALEX)	José M. Bustos Gisbert y Jorge J. Sánchez Iglesias de la Universidad de Salamanca en colaboración con el Instituto Cervantes	Escrito	Griego, portugués, italiano, francés, alemán, coreano, chino, japonés, inglés, rumano, árabe, polaco, ruso, húngaro, neerlandés, checo, búlgaro, eslovaco, sueco, turco, ucraniano, albanés, kazajo, moldavo, letón, farsi, vietnamita, serbio, georgiano, azreí, croata, lituano, esloveno, macedonio, noruego, danés, armenio, estonio, hebreo, islandés, napelí, tailandés	Hombres y mujeres	Diversas edades	B2	2 200 000 palabras aprox. en 12 576 textos	2009-2010	Producciones escritas correspondientes a cinco tipologías textuales: carta formal, carta informal, texto narrativo, texto descriptivo y texto expositivo (promedio de 175 palabras por texto)	Compilación de un conjunto amplio de datos lingüísticos objetivos y auténticos que permitan crear una base de consulta sólida y desarrollar un procedimiento científico apropiado y pertinente para el análisis de la adquisición del español como lengua extranjera y para el estudio de la interlengua de los aprendices.	Transversal	Acceso libre en https://diarium.usal.es/espalex/recursos/corpusele/	Total de informantes: 6 288. Personas que han realizado una prueba de certificación de idioma (DELE) en uno de los 360 centros del Instituto Cervantes que abarca el corpus, distribuidos en 85 países diferentes.	Sin anotación lingüística, pero con metadatos sobre los informantes y sobre la tarea	Este corpus se ha creado a partir del contenido de los exámenes de certificación de idioma (nivel B2) del Instituto Cervantes, el Diploma de Español como Lengua Extranjera (DELE).
Corpus de Suecos Aprendices de ELE (SAELE)	Aymé Pino Rodríguez de la Universidad Jönköping de Suecia. Proyecto de colaboración con otra universidad sueca.	Escrito	Sueco	Hombres y mujeres	Jóvenes estudiantes (entre 18 y 25 años aprox.)	A2 y B1	7 000 palabras aprox. en 135 textos	Dos fases de compilación: 2008-2009 y 2009-2010	Producciones escritas de índole argumentativa	Descripción de la interlengua de los aprendices suecos de español de acuerdo con los parámetros del MCER.	Transversal	Acceso restringido	Total de informantes: 45. Estudiantes universitarios de español en Suecia.	Anotación lingüística y metadatos sobre informantes	-
Spanish Corpus of Italian Learners (SCIL)	Sonia Bailini de la Universidad Católica del Sagrado Corazón (Milán)	Escrito	Italiano	Hombres y mujeres (predominio de mujeres)	Jóvenes estudiantes (promedio de edad de 20 años)	A1-B2	124 186 textos	Dos fases de compilación: 2008-2009 y 2012-2013	Cartas y correos electrónicos informales de tipo narrativo, expositivo, descriptivo, instructivo y argumentativo	Compilación de datos objetivos, auténticos y fidedignos para el estudio de la interlengua de estudiantes italianos de español como lengua extranjera y para la creación de materiales didácticos.	Longitudinal	Acceso restringido	Total de informantes: 43. Estudiantes de español como lengua extranjera en un contexto de formación universitario. Cada informante escribió varias redacciones a lo largo de siete meses en cada una de las fases de compilación. Todos los texto fueron recopilados en el contexto de grupos de e-tándem entre estudiantes nativos y aprendientes de español.	Anotación lingüística y metadatos sobre informantes	-
Language and Social Networks Abroad Project (LANGSNAP)	Miembros de los equipos de SPLLOC y FLLOC. Director del proyecto: Rosamond Mitchell	Escrito y oral	Inglés	Hombres y mujeres	Jóvenes estudiantes (entre 18 y 25 años aprox.)	Nivel avanzado	300 000 palabras aprox. de muestras orales y escritas	2011-2013 (23 meses en total)	Producciones orales: entrevistas semiguiadas y relatos a partir de imágenes (con transcripciones). Producciones escritas: argumentaciones breves en respuesta a una pregunta	Documentación del desarrollo de la competencia y del uso de una lengua extranjera.	Longitudinal	Acceso libre en http://www.flloc.soton.ac.uk/search.php	Total de informantes: 27. Estudiantes universitarios. Cada informante escribió seis textos distintos en diferentes momentos a lo largo de un período de veintiún meses. En ese tiempo, se incluía una estancia en el extranjero de nueve meses en un contexto formativo de inmersión (España o México).	Metadatos sobre informantes y muestras	La búsqueda se puede realizar atendiendo a distintos parámetros: tipo de tarea, momento de recogida del texto, informante concreto.