Quintú Quimün Nº 5, Q051, 2021. ISSN: 2591-541X
Dossier: *Habitar la interlengua: escenarios de la enseñanza y la investigación de ELSE en Argentina*

Avances en el tratamiento computacional en corpus de aprendientes de español como lengua segunda y extranjera*

Advances in the Computational Processing in Corpora of Learners of Spanish as a Second Language and Foreign Language

Carolina Paola Tramallino
Consejo Nacional de Investigaciones Científicas y Técnicas
Centro de Estudios de Adquisición del Lenguaje, Universidad Nacional de Rosario
carolinatramallino@gmail.com
ORCID: 0000-0003-4760-7005

Recibido: 5 de julio de 2021
Aceptado: 21 de septiembre de 2021

RESUMEN

El presente trabajo se propone explicitar las aplicaciones y los alcances de la lingüística computacional con el objeto de revisar, particularmente, las investigaciones actuales sobre análisis y tratamiento automático en corpus de aprendientes de español como lengua segunda y extranjera. Por lo tanto, será necesario exponer qué entendemos por lingüística de corpus; qué requisitos debe cumplir la conformación de una muestra de producciones orales o escritas en el ámbito de la adquisición de lenguas; y finalmente, cuáles son los corpus de español como lengua extranjera y/o segunda lengua que se han constituido y cuentan con acceso libre. Luego de realizar un breve recorrido por la evolución del concepto de "error" en este campo investigativo, se hará referencia a los criterios que se utilizan para realizar diferentes clasificaciones de las evidencias halladas en las muestras. A continuación, se expondrán algunos emprendimientos surgidos en los últimos años, cuya finalidad es detectar y contabilizar

* Quiero agradecer a la coordinadora del dossier, Ana Brown, por invitarme a formar parte de este número tan relevante y especialmente a los/las evaluadores/as que con sus comentarios y observaciones genuinas contribuyeron a mejorar el manuscrito. Además, agradezco a Natalia Ricciardi por haber sido la primera lectora del artículo.

las formas y estructuras sintácticas presentes en los corpus, mediante herramientas informáticas.

Palabras clave: lingüística computacional, tratamiento computacional, corpus, español como lengua segunda y extranjera.

ABSTRACT

The present work aims to explain the uses and scope of computational linguistics in order to review mainly the current research on the analysis and automatic processing of errors in corpora of learners of Spanish as a second and foreign language. Therefore, it will be necessary to explain what is meant by corpus linguistics; what requirements must be met by a sample of oral or written productions in the field of language acquisition; and, finally, which are the corpora of Spanish as a second language or foreign language that have been built and are free to access. After a brief review of the evolution of the concept of "error" in this research field, reference will be made to the criteria used to make different classifications of the evidence found in the samples. Next, some projects from recent years will be introduced, whose purpose is to detect and count the syntactic forms and structures present in the corpus, by means of computer tools.

Keywords: computational linguistics, computational processing, corpora, Spanish as a second and foreign language

1. Introducción

Este estudio se propone exponer los principales antecedentes que existen en el ámbito del análisis de corpus de aprendientes de español a través de *softwares* y que impactan en la metodología de enseñanza del español como segunda lengua. El recorrido por las corrientes teóricas encuadradas en la adquisición de segundas lenguas que mantuvieron diferentes posturas frente a los antiguamente denominados "errores" que cometían los aprendientes en sus producciones orales o escritas indicará un largo camino trazado hasta la actualidad. Este trayecto exhibe una clara evolución del tratamiento del error hasta llegar a la teoría de Interlengua (IL), que lo considera un indicio positivo y una forma idiosincrásica propia de ese sistema particular que posee cada estudiante. Por consiguiente, las formas idiosincrásicas se

convierten en mecanismos de aprendizaje que responden a la generación de hipótesis y a la aplicación de nuevas reglas gramaticales que se aprenden y en efecto, son importantes al igual que la presencia de vocablos y estructuras que pertenecen al sistema lingüístico que se adquiere.

El marco teórico en el que se encuadra es la lingüística computacional, disciplina que cuenta con una trayectoria de varias décadas en países como Estados Unidos, sobre todo desde la segunda mitad del Siglo XX, no equiparable con la que poseen los países de América Latina. En efecto, veremos que la mayor disponibilidad de softwares necesarios para desarrollar investigaciones en castellano se ha realizado en algunos centros académicos europeos y en países no hispanohablantes; por lo tanto, se hallan en lenguas diferentes al español, tal como plantea Parodi (2004).

La lingüística computacional surge en respuesta a nuevas necesidades de abordar el conocimiento de las lenguas y de sus usos a partir del auge de las comunicaciones y de los contenidos digitales que tuvo lugar en las últimas décadas del siglo pasado. Se constituye como un área interdisciplinaria que toma saberes de la Lingüística, la Informática y la Estadística y está orientada al estudio del conocimiento lingüístico que se obtiene a partir de una formalización[1]. Como afirma Lavid (2005), emerge como ciencia del lenguaje que contribuye al conocimiento de los procesos cognitivos de comprensión y producción del lenguaje, combinando teorías referidas a esos procesos con diversas técnicas. Se propone, como tarea principal, efectuar el procesamiento del lenguaje natural mediante sistemas computacionales que sean capaces de emular[2] la capacidad lingüística humana. Moreno (1998) manifiesta que el objetivo de la disciplina es el diseño de herramientas informáticas que habiliten la comprensión y generación del lenguaje para que este pueda analizarse de manera confiable y automáticamente. Entre sus aplicaciones, se encuentran la comunicación entre el hombre y las máquinas; la comunicación entre personas que hablan diferentes lenguas; el análisis de textos y la creación de diccionarios electrónicos o de correctores. Como afirma Bonino (2009), la utilidad práctica de este tipo de investigaciones excede el interés de la lingüística teórica. Sin embargo, abren la posibilidad de utilizar la Informática tanto en lo que respecta a la obtención de datos, como a la formulación de hipótesis formales y procedimientos metodológicos. La

[1] El concepto de formalización lingüística alude a la exigencia en la formulación de reglas que contienen el conocimiento lingüístico derivado de una gramática.
[2]Con el término *emular* nos referimos a la tarea de intentar construir sistemas que comprendan y produzcan el lenguaje de manera similar a un humano.

ventaja de utilizar modelos computacionales reside en que habilitan la comprobación de teorías lingüísticas ya que los procesos pueden ser inspeccionados y experimentados a través de la elaboración de programas.

Dentro de las líneas investigativas generadas a partir de desarrollos computacionales, este artículo se centra en la lingüística de corpus, cuyo carácter teórico o metodológico, de acuerdo a sus amplias y productivas aplicaciones, genera discusiones que se expondrán en el siguiente apartado.

La metodología de trabajo consiste en describir y comparar las características que presentan las muestras informatizadas de lenguaje en uso del español que se encuentran disponibles en internet y los recursos que ofrecen. Por lo tanto, en las páginas que continúan, se realizará un recorrido por los diferentes corpus digitales de acceso abierto. Estas muestras de producciones orales o escritas cobran interés para investigadores y docentes que se ocupan de la enseñanza y adquisición del español como Lengua Segunda y Extranjera (ELSE). Asimismo, los lingüistas computacionales se sirven de los mencionados corpus para efectuar diversos análisis automáticos que posibiliten hallar patrones del lenguaje y extraer conclusiones en cuanto a la cantidad y distribución de determinadas formas y estructuras.

De estas cuestiones surge el presente estudio, que tiene por objeto brindar un estado del arte acerca de las investigaciones sobre tratamiento informático en corpus de estudiantes de español.

2. La lingüística de corpus y su relación con la Lingüística computacional

En las décadas de 1940 y 1950 nace la Lingüística aplicada como una disciplina independiente en centros académicos de Estados Unidos interesados por la enseñanza y el aprendizaje de idiomas. Esta puede definirse como una orientación de la investigación lingüística que parte de marcos interdisciplinarios y tiene como objetivo la resolución de problemas que derivan de la praxis lingüística (Payrató 2011).

Rojo (2002) señala el comienzo de la lingüística de corpus con la aparición en 1964 del Brown University Standard Corpus of Present Day American English, de Francis y Kučera, que fue el primer corpus construido para permanecer en una computadora y poder explotarse desde programación informática. Con respecto al carácter de la lingüística de corpus, encontramos diferentes posturas que plantean una discusión en relación con si debe considerarse como una metodología o bien, como una teoría. Por un lado, Parodi (2008) afirma

que la lingüística de corpus se ha convertido en una metodología que se apoya en técnicas estadísticas y computacionales para estudiar datos reales de la lengua. Agrega, además, que la información recabada posibilita extraer resultados certeros que justifican el desarrollo del conocimiento científico. Por lo tanto, el empleo de los corpus como fuente de evidencias es compatible con cualquier tipo de teoría. En este mismo sentido, Altenberg (2011) la define como un componente indispensable del aparato metodológico de la lingüística desde la década de 1990. Hace hincapié en que se trata de una metodología más que de una teoría del lenguaje, ya que ha permitido acceder al estudio de la lengua de forma más fácil en lo que respecta a su naturaleza y la manera en que la procesamos; y además, ha posibilitado vincular la teoría y los datos de una nueva forma (Altenberg 2011).

Por otra parte, Bolaños Cuellar (2015) distingue la Lingüística computacional del análisis de corpus debido a que esta última tiene por objetivo emplear las muestras disponibles o crearlas, mediante las herramientas informáticas que fueron diseñadas en el terreno de la lingüística computacional. Asimismo, la sitúa en el mismo paradigma de las disciplinas que se dedican a estudiar las diferentes manifestaciones del uso del lenguaje en contextos reales de interacción comunicativa, lo cual correspondería a un enfoque lingüístico funcionalista. Este autor subraya que la lingüística de corpus implica, además de una nueva mirada metodológica, una perspectiva conceptual diferenciada de los fenómenos lingüísticos, debido no solo a su transversalidad disciplinaria, sino también a la perspectiva conjunta entre análisis cuantitativo y cualitativo (Bolaños Cuellar 2015).

En los inicios del Siglo XXI, surgen diversos traductores automáticos, diccionarios electrónicos monolingües y multilingües que responden a las llamadas tecnologías del habla. Para la enseñanza de la lengua española, disponemos de variadas herramientas informáticas que han sido creadas por lingüistas computacionales: flexionadores léxicos de sustantivos y adjetivos[3] como el Banco de datos SENSEM (Alonso *et. al* 2005) o desambiguadores como el del Grupo de Estructuras de Datos y Lingüística computacional (Santana Suárez *et al* 2009)[4]. Asimismo, contamos con analizadores sintácticos como *Freeling*, analizador multilingüe[5] que

[3]Disponible en: https://www.cs.upc.edu/~nlp/papers/padro11.pdf
[4]Disponible en: http://www.gedlc.ulpgc.es/investigacion/desambigua/morfosintactico.htm
[5] Disponible en: http://nlp.lsi.upc.edu/freeling/index.php/node/1fuedesarrollado por el Centre de Tecnologies i Aplicacions de Llenguatge i la Parla (TALP) de la Universidad Politécnica de Catalunya (UPC).

comprende un conjunto de herramientas de lenguaje de código abierto (Padró 2011). Es importante mencionar a *WordNet* (Miller et. al 1993), base de datos léxico-conceptual del inglés, creada por la Universidad de Princeton[6] que organiza los sustantivos, verbos, adjetivos y adverbios en conjuntos de sinónimos. Para el español se creó WordNets 1.5 (Atserias *et al*., 1997). Además, Oliver & Climent (2011) construyeron WordNets 3.0 en castellano y en vasco para el que utilizaron sistemas de traducción automática de corpus anotados semánticamente.

3. Corpus digitalizados para el español

Los corpus informatizados representan una herramienta sumamente importante en el ámbito de la investigación lingüística ya que, no solo permiten ahorrar tiempo sino también, trabajar de una manera más ordenada y exhaustiva. En este mismo sentido, Parodi (2008) declara que gracias a los progresos tecnológicos ocurridos en un tiempo muy breve se avanzó en la construcción y el almacenamiento de estas bases de datos computarizadas. Dicho método de trabajo es muy productivo, no solo en el ámbito de la lexicografía y de la terminología, sino también en el de la estadística lingüística. En esta última, las muestras se emplean para constituir índices de frecuencias; para establecer combinaciones léxicas de distinta naturaleza y frecuencias de aparición de vocablos; datos que interesan a nivel sociolingüístico y estilístico (Torruella & Listerri 1999). En síntesis, la diferencia metodológica entre la lingüística descriptiva y la de corpus no se trata solo del crecimiento en el número de textos. Por el contrario, el aumento cuantitativo produce un salto cualitativo, como explica Rojo (2002), ya que las tecnologías permiten trabajar con todos los ejemplos que se pueden detectar en un corpus que contiene cientos de millones de formas.

Torruella & Listerri (1999) manifiestan que la diferencia entre una recopilación de textos y un corpus es que éste último responde a un criterio interno que refiere a patrones lingüísticos presentes en los textos. Parodi (2008) agrega que este conjunto de datos debe ser accesible desde entornos computacionales y, además, poseer visibilidad, de forma tal que pueda utilizarse en otras investigaciones y se integre ese conocimiento para el estudio de una lengua en particular, o bien, para la comparación con otros sistemas lingüísticos. Al mismo tiempo, debe contar con datos precisos acerca de la recolección de la muestra y de su procedencia.

[6] Disponible en http://globalwordnet.org/

Entre los desarrollos más antiguos, cabe mencionar CHILDES (Child Language Data Exchange System)[7], creado en 1984 (Mac Whinney & Snow 1990), que es una base de datos que contiene 44 millones de palabras en 28 lenguas diferentes, entre ellas, el español. Los textos corresponden a transcripciones de conversaciones entre niños/as y sus madres, padres e investigadores. Además, presenta interacciones entre estudiantes y profesores en el marco de la adquisición de segundas lenguas. Este corpus de acceso libre proporciona herramientas acerca de los métodos de codificación lingüística y los sistemas para conectar las transcripciones a las grabaciones digitales de audio y video. Sus archivos poseen formato CHAT y se analizan desde el programa CLAN (*Computerized Language Analysis*), incluyen el cálculo de frecuencias, la búsqueda de palabras y el análisis morfosintáctico, entre otros recursos (MacWhinney & Wagner 2010).

En 2001, Mark David publica el *Corpus del español*[8]; este consta de una muestra original y más pequeña con la que se pueden rastrear cambios históricos y variación de géneros; además del nuevo corpus, que permite localizar variaciones dialectales y cuenta con datos para el español contemporáneo. Este último, denominado corpus *NOW*[9] fue elaborado en 2018, brinda al usuario la posibilidad de localizar palabras, frases, lemas, categorías gramaticales, colocaciones y frecuencias. Cuenta con más de 7.200 millones de palabras extraídas de textos periodísticos digitales de 21 países diferentes de habla hispana que corresponden al período 2012- 2019. Es interesante reparar en el hecho de que todos los meses se actualiza a partir de la incorporación de 150 millones de nuevos datos extraídos de la Web. Al mismo tiempo, posibilita el acceso a gráficos de frecuencias de términos.

Asimismo, contamos con *Corpus Diacrónico y Diatópico del Español de América* (CORDIAM)[10], creado por la Academia Mexicana de la Lengua y la Academia Española de Lenguas. Este es un corpus de especialidad que reúne tres subcorpus con diferentes tipos de datos: documentos, literatura y prensa; contiene solo documentos de América *y* cuenta con textos desde el año 1494 a 1905. Abarca, así, el período fundacional, el virreinal y el primer siglo de la mayoría de las independencias de los países americanos. Expone en su sitio web

[7]Disponible en https://childes.talkbank.org/. Fue creado por Brian MacWhinney & Catherine Snow.
[8] Disponible en: https://www.corpusdelespanol.org/
[9] Disponible en: https://www.corpusdelespanol.org/now/
[10] Disponible en: http://www.cordiam.org/

varias características innovadoras, como por ejemplo, permitir ver y guardar el documento completo o guardar las concordancias seleccionadas en una base de datos.

Además, existe *El GRIAL*[11], desarrollado por la Escuela de Lingüística de Valparaíso, Chile, es una interfaz computacional que permite tanto la realización de anotaciones morfosintácticas en textos planos de español como la interrogación o consulta en forma de base de datos de las muestras. Las búsquedas se aplican sobre tres categorías: formas, lemas y partes de la oración, a las que se puede agregar información sobre género y número. Los resultados se despliegan según dos modalidades: por frecuencia y en contexto. También, puede efectuar localizaciones de acuerdo a la categoría de la palabra. Además, contiene el "Manchador de Textos" (EMT), una herramienta computacional que muestra la frecuencia de aparición de secuencias a través del coloreado de las palabras o estructuras lingüísticas que han sido rastreadas.

Respecto de los corpus de referencia creados por la Real Academia Española (RAE), contamos con el *Corpus de Referencia del Español Actual* (CREA)[12] y el *Corpus Diacrónico del Español* (CORDE)[13], iniciados a mediados de la década del noventa del siglo pasado. Actualmente, disponemos del *Corpus del Español del Siglo XXI* (CORPES XXI)[14], perteneciente a la RAE y a ASALE (Asociación de Academias de la Lengua Española), el cual se encuentra en construcción y en continuo crecimiento. Está formado por textos escritos y orales; procedentes de España, América, Filipinas y Guinea Ecuatorial; con una distribución de 25 millones de formas por cada uno de los años correspondientes al siglo XXI. La nueva versión de febrero de 2021 reúne más de 316.000 documentos que superan los 333 millones de formas ortográficas, procedentes de textos provenientes de España en un 70 % y de América en un 30 %. Con respecto al eje temporal, aumenta el número de datos producidos entre 2016 y 2020, sumando más de 42 millones de formas**.**

4. La evolución del concepto de "error" en la adquisición de lenguas

En la década de 1940, en el ámbito de la lingüística aplicada, aparecen las primeras investigaciones que tienen como objeto de estudio la lengua del aprendiente. A partir de un

[11] Disponible en: http://www.elgrial.cl/
[12] Disponible en http://corpus.rae.es/creanet.html
[13]Disponible en: http://corpus.rae.es/cordenet.html
[14] Disponible en https://www.rae.es/banco-de-datos/corpes-xxi

análisis contrastivo entre la lengua nativa del estudiante y la lengua meta, aquella que se busca aprender, se intentan predecir los errores que este cometerá con la finalidad de evitarlos. De esta forma, se realiza una exploración descriptiva con respecto a las zonas de contacto y también de divergencia entre ambos sistemas en todos los niveles lingüísticos: el fonológico, morfológico, sintáctico y léxico para así, detectar y pronosticar los posibles errores. No obstante, al avanzar en las investigaciones y advertir que la causa de estos no respondía únicamente a la interferencia con la lengua materna (entendida esta como el efecto de la lengua nativa sobre la lengua meta) surge un nuevo paradigma que se denomina análisis de errores. El primer referente de esta línea investigativa es Corder (1967), quien les atribuye a los errores un valor positivo, en lugar de considerarlos como un hecho negativo, los toma como indicadores del proceso de aprendizaje llevado a cabo por el aprendiz. En este sentido, el análisis de errores, desde una perspectiva más empírica, se ocupa de clasificarlos, pero no de predecirlos, tarea que ejercía un lugar central en el análisis contrastivo.

La corriente de interlengua (IL) significa un giro metodológico ya que adopta el estudio y análisis tanto de las formas propias o idiosincrásicas como de las que coinciden con el sistema lingüístico que se adquiere (Alexopolou 2010). El término es empleado, inicialmente, por Selinker (1972), quien define la IL desde un enfoque psicolingüístico, como un sistema que se halla en un lugar intermedio entre la lengua materna y la meta, hecho que la hace disponer de elementos que son comunes con ambas lenguas. De esta forma, explica el tránsito por etapas o niveles de competencia que modifican ese sistema a medida que los estudiantes adquieren léxico y nuevas estructuras de la lengua que aprenden. El referido atributo de variabilidad responde, por lo tanto, a su carácter transitorio.

Otra particularidad que posee la IL es la sistematicidad establecida por su coherencia interna en un determinado momento de su desarrollo. Esto se percibe en la aplicación de reglas lingüísticas que responden a estrategias y procesos que activan los aprendientes. Lo interesante de esta perspectiva se halla en la hipótesis que sostiene: aunque esta lengua idiosincrásica difiera en cada alumno en particular, presenta zonas de intersección en aprendientes que comparten un mismo nivel de instrucción. Corder (1971), la había denominado dialecto idiosincrásico o transitorio al que también le atribuía como característica distintiva la de poseer una gramática propia que se encuentra en continuo cambio. Nemser (1971), se refería al mismo

concepto como un sistema aproximativo al que le adjudicaba no solo una gramática particular sino también un vocabulario específico.

Es importante detenernos en el concepto de *fosilización* que propone Selinker (1972) para justificar la presencia de elementos de la IL que son ajenos a la lengua meta pero que, sin embargo, persisten con el paso del tiempo e incluso se trasladan a niveles de aprendizaje más avanzados. Las causas contemplan la interferencia lingüística con la L1 o con otras lenguas que ha aprendido el estudiante pero, además, responden a las estrategias de comunicación y de aprendizaje que activa el sujeto. Es interesante reparar en la observación que realizan Santos Gargallo & Alexopolou (2021), quienes afirman que este fenómeno es más frecuente en hablantes no nativos cuyo objetivo es obtener una competencia comunicativa funcional en situaciones de la vida cotidiana y, también, en estudiantes que acceden al aprendizaje de una L2/LE en la edad adulta y sin experiencia previa en el aprendizaje de otros idiomas.

En esta línea de investigación, Alexopolou (2005, 2010) se basa en un corpus de producciones escritas de estudiantes griegos de español y propone el criterio etiológico como el más indicado para desentrañar los mecanismos subyacentes para determinado comportamiento lingüístico. Fernández López (1995), analiza las desviaciones de IL según los estadios del proceso de aprendizaje en los que aparecen, a partir de producciones escritas. Concluye que en el nivel inicial predominan las estrategias interlinguales, cuando se recurre a la LM o a otra lengua y las intralinguales (al repetir frases hechas, neutralizar oposiciones o evadir ciertas estructuras). En el nivel intermedio, abundan las hipergeneralizaciones de reglas de la lengua que se adquiere, hay más riesgo y por lo tanto, mayor cantidad de formas propias de IL y a su vez, persisten las estrategias empleadas en la etapa anterior. Aclara que la fosilización o no avance en este camino, se debe a una falta de motivación que se alcanza al obtener un nivel suficiente de comunicación En efecto, tomar la IL como marco de estudio permite estudiar el avance en la adquisición de la nueva lengua al determinar la ocurrencia y persistencia de los "errores" transitorios, fosilizables y fosilizados. Tal como explica Fernández López (1995), los primeros, también denominados de desarrollo, son los que pertenecen al primer estadio y se superan en etapas posteriores, por ejemplo, la regularización de verbos irregulares; los fosilizables son los más persistentes, por ejemplo, el uso o la omisión de artículo y la oposición ser/estar más adjetivo. Por último, los fosilizados son aquellos que pueden aparecer en estadios avanzados pero que se autocorrigen.

En referencia a la perspectiva que puede aplicarse en la taxonomía de las desviaciones existen numerosos criterios, sin embargo, el aspecto más trabajado por los investigadores es el

criterio lingüístico, que considera los niveles del lenguaje en el que estas se producen. Así lo indica el relevamiento realizado por Santos Gargallo & Alexopolou (2021) acerca de las tesis doctorales sobre análisis de la IL de español de universidades españolas publicadas en los últimos 30 años. Los resultados de este estudio indican que los más indagados son los fenómenos morfosintácticos seguidos de los léxico-semánticos, los gráfico-ortográficos, los fonético-fonológicos y, por último, los pragmático-discursivos. Como explica Bustos Gisbert (1998), se prioriza el estudio formal en detrimento del carácter funcional. El criterio comunicativo, pragmático y cultural no está debidamente investigado. Dicha perspectiva que valoriza la competencia lingüística por sobre las competencias pragmáticas y socio- discursivas está anclada en una visión reduccionista acerca de la concepción de la enseñanza y el estudio de la lengua. En esta misma dirección, Arispe *et al*. (2020) manifiestan la carencia de investigaciones de línea pragmática dentro de los estudios en ELSE: "los manuales para estudiantes de español no suelen incluir explícitamente tareas o dimensiones pragmáticas, especialmente durante los primeros niveles" (Arispe *et al*. 2020: 2).

La ventaja de incorporar los aspectos extralingüísticos radicaría en analizar los fenómenos desde una perspectiva global para atender tanto al proceso de adquisición como a las hipótesis que realiza el aprendiente en cada etapa de aprendizaje. Para ello es necesario complementar estudios que abarquen distintos factores, como por ejemplo, las causas de aquellos indicios junto con el carácter pedagógico y el estudio sociolingüístico para no agotar el análisis en la mera corrección. En coincidencia con Camargo Angelucci & Pozzo (2020), es desde la perspectiva comunicacional que el alumno se convierte en una figura central, que se considera capaz de generar, probar y reformular hipótesis sobre lo que aprende.

Actualmente, aquella primera concepción sobre el "error" derivada del análisis contrastivo que lo entendía como un hábito lingüístico negativo que debía ser evitado, ha evolucionado hasta convertirse en una evidencia positiva, un rasgo particular que exhibe todo hablante en el proceso de adquisición de una nueva lengua. Asimismo, desde la Socio-lingüística se desmitifica el lugar de la norma en la enseñanza de idiomas y lo que se discute es la concepción de lo que se considera o no "error". A raíz de esto, se establece una distinción entre la eficacia del intercambio comunicativo y el uso correcto de la norma (Camargo Angelucci & Pozzo 2020).

4.1. Corpus informatizados del español

Los corpus informatizados de aprendices surgen a comienzos de la década del 90 del siglo pasado pero cobran relevancia a partir de 2002, cuando se publica *International Corpus of Learner English* (ICLE) llevado a cabo por Sylviane Granger[15]. La conformación de corpus de aprendientes de español es mucho más restringida y de reciente aparición. Palacios Martinez et. al (2019), distinguen tres grandes líneas de trabajo este ámbito: 1) estudios comparativos entre el uso del español por parte de hablantes nativos y no nativos en la línea del Análisis contrastivo de interlengua (Granger 2002); 2) investigaciones que tienen como objeto confirmar o rechazar hipótesis acerca de la existencia de ciertas áreas de la gramática española que resultan dificultosas para el aprendizaje y 3) estudios sobre la adquisición de morfemas gramaticales de español lengua extranjera, en paralelo a los que se realizaron sobre el español como lengua materna. Ferreira Cabrera & Elejalde Gómez (2017), sostienen que en el campo de la enseñanza, el empleo de corpus resuelve la necesidad de observar cómo los sujetos usan de manera real la lengua meta y, a partir de sus resultados, averiguar qué áreas presentan mayores problemas, cuáles persisten o responden a diferentes fenómenos en lo que respecta al desarrollo de la IL. Torijano (2008), expone que la necesidad de identificar la recurrencia de errores responde al interés por averiguar cuáles son los que se mantienen a lo largo del tiempo en un determinado proceso de aprendizaje.

Por consiguiente, este tipo de muestras permite, además, proporcionar datos de interés para la elaboración de gramáticas, diccionarios, glosarios y libros de texto específicos para la enseñanza del español como lengua extranjera. A continuación, se detallan los corpus disponibles para uso y consulta de estudiantes, docentes e investigadores.

Dentro de los corpus orales, es preciso hacer alusión al Corpus oral de español como lengua extranjera (ELE)[16] que recoge muestras de 40 estudiantes que se encuentran en los niveles A2, correspondiente al segundo estrato dentro del nivel inicial y B1, que refiere al primer estrato dentro del nivel intermedio, según el Marco Común de Referencia Europeo (MCER[17]). Los informantes de estas muestras son alumnos de programas de idiomas de la Universidad Autónoma y de la Universidad Complutense de Madrid. En cuanto a la búsqueda de errores, estos pueden localizarse según las siguientes etiquetas: nivel lingüístico, categoría,

[15]Fundadora del *Centro de Lingüística de Corpus del Inglés* (CECL), disponible en: www.uclouvain.be/encecl.htm

[16] Disponible en: http://cartago.lllf.uam.es/exist/rest//db/corpus/home_es.html

[17] Disponible en http://cvc.cervantes.es/ensenanza/biblioteca_ele/marco/cvc_mer.pdf.

lengua materna, mecanismo de cambio, tipo de error y nivel de competencia. Además, distingue entre errores ambiguos y no ambiguos y proporciona acceso al vocabulario frecuente, a la lista de lemas, a los fenómenos de pronunciación más prominentes de cada lengua nativa y a actividades para profesores, entre otros recursos.

También, existe SPLLOC (*Spanish Learner Language Oral Corpus*)[18], que es una recopilación de textos orales de español LE surgido a partir de una investigación conjunta entre tres universidades del Reino Unido. Fundamentan su creación por el hecho de disminuir de alguna forma el "vacío metodológico" existente en lo que respecta a recursos tecnológicos para la lengua española. Cuenta con datos recopilados a través de ejercicios de juicios de hablantes, además de datos orales correspondientes a distintos géneros textuales; los informantes pertenecen a diferentes niveles de instrucción: principiante, intermedio y avanzado. Asimismo, posee vinculación con el sistema CHILDES y acceso libre a los materiales, como archivos de sonido y transcripciones.

Es pertinente mencionar el proyecto *AACFELE*[19] (Adquisición y aprendizaje del componente fónico del español como lengua extranjera) el cual está financiado por el Ministerio de Ciencia e Innovación de España. Reúne a investigadores de siete universidades de diferentes países cuyas nacionalidades corresponden a las lenguas de origen de los estudiantes: alemanes, griegos, taiwaneses, polacos, portugueses y egipcios. Se puede acceder al audio de las muestras de aprendices de los niveles A2, B1, B2 y C1 (MCRE). Contiene, además, una plataforma de código abierto llamada *fono e-learning* que permite al docente crear cursos específicos donde puede incorporar contenidos teóricos y recursos disponibles como audios, instrumentos y fichas. Al mismo tiempo, otorga herramientas de evaluación y de comunicación (foros, chat, wikis), que posibilitan un entorno de aprendizaje interactivo.

En lo que refiere a corpus de textos escritos, contamos con CEDEL2[20] (Corpus escrito del Español como L2), creado por Lozano Cristóbal (2009), de la Universidad de Granada. Asciende, actualmente, a una muestra de 1.105.936 palabras provenientes de 4.399 participantes con diferentes lenguas maternas como el inglés, francés, portugués, alemán, italiano, griego, holandés, ruso, japonés, chino y árabe. Incluye, además, textos de hablantes

[18] Disponible en: http://www.splloc.soton.ac.uk/
[19] Disponible en: http://www3.uah.es/fonoele/proyecto.php
[20] Disponible en: http://cedel2.learnercorpora.com/

nativos de español y la posibilidad de participar, a través de un formulario estandarizado de *Google*, como informante del corpus.

Además, disponemos del *Corpus de Aprendices de Español como Lengua Extranjera* (CAES)[21] del Instituto Cervantes y la Universidad de Santiago de Compostela, que reúne 600000 elementos lingüísticos etiquetados y lematizados. Estos pertenecen a producciones escritas con consignas pautadas de estudiantes de niveles: A1, A2, B1, B2 y C1 de seis L1 diferentes (árabe, chino mandarín, francés, inglés, portugués y ruso). Palacio Martinez *et al*. (2020) explican cómo se constituyó esta muestra: primero se utilizó *Freeling* para el análisis automático y luego se empleó una herramienta que se encargó se desambiguar manualmente el resultado de la etiquetación automática.

Con respecto a corpus recopilados en América Latina, podemos mencionar el CAELE (*Corpus de Aprendientes de Español como Lengua Extranjera*) que se conformó en Chile. Esta muestra se compone de más de 400 textos escritos, procesados en formato digital, que corresponden al nivel A1, A2, B1 y B2 (MCRE), tiene como referente a Ferreira Cabrera de la Universidad de Concepción.

Otro proyecto que merece referenciarse es (CAELE/2), *Corpus de Aprendientes de Español como Lengua Extranjera y Segunda Lengua* en su componente escrito, que fue recopilado en Colombia por el Instituto Caro y Cuervo. Reúne 166 muestras sistematizadas, tomadas de 83 participantes que respondieron a pruebas diseñadas (Hincapié, 2018). Lo interesante del emprendimiento es que recopila muestras de segunda lengua en respuesta a la diversidad lingüística del país, para de esta forma contemplar las necesidades educativas de colombianos indígenas y sordos, quienes no tienen como L1 el español.

Además, podemos mencionar otros corpus, un tanto más pequeño y no disponible en internet como es el caso de CORANE [22] (Cestero Mancera *et al*. 2001) de la Universidad de Alcalá de Henares. Cuenta con 1091 composiciones de 213 informantes cuyas lenguas nativas son más de 20. Los ejercicios corresponden a consignas dadas en las clases de español y los niveles de referencia según el MCRE son A2, B1, B2 y C1. Por último, el corpus SAELE (suecos aprendices de español lengua extranjera), está constituido por una colección digital de textos argumentativos escritos por aprendices de nivel A2 y B1 (Pino Rodríguez 2009).

[21] Disponible en: http://galvan.usc.es/caes/

[22] Se encuentra disponible en CD ROM, el siguiente sitio web reúne informaciones sobre el proyecto: http://repositorios.fdi.ucm.es/corpus_aprendices_espa%C3%B1ol/view/cm_view_virtual_object.php?idov=320&seleccion=1

4.1.1. Analizadores de corpus de español lengua extranjera

Entre los desarrollos correspondientes a los últimos años elaborados para el análisis y corrección de textos de aprendientes de español, podemos referenciar ELE Tutor (Ferreira et al. 2012). Este tutor inteligente cuenta con la posibilidad de analizar gramaticalmente una entrada en lenguaje natural y luego, generar un enunciado correspondiente a una estrategia de *feedback* correctivo adecuado para la falla localizada. Dentro de este proyecto, se adscribe la investigación de Kotz Gabrole & Ferreira Cabrera (2013), quienes proponen la construcción de un analizador sintáctico computacional. Las autoras establecen una jerarquía de evidencias según dos criterios: la gravedad y la frecuencia. Entre las más graves se encuentran no dar una respuesta acorde al ejercicio propuesto, mientras que, entre las más frecuentes se hallan las fallas de concordancia y de empleo de verbos (Kotz Grabole & Ferreira Cabrera 2013).

Por otra parte, Elejalde & Vine (2014) presentan un análisis de Errores Asistido por Computadora a partir del CAELE, que se compone de resúmenes realizados por estudiantes de nivel B1. Para ello, diseñan un sistema de etiquetas de anotación de errores y los analizan en diversos niveles de categorización a partir del programa *Nvivo 10.* Los resultados arrojan que la mayoría de los errores son lingüísticos y entre estos, los más representativos son los gramaticales. Con el mismo corpus, Ferreira Cabrera & Elejalde Gómez (2017) determinan los errores más recurrentes y obtienen la frecuencia por cada sujeto acorde con el nivel de competencia. Para ello, emplean el programa *Uam Corpus Tool*. Los resultados del análisis, efectuado con estudiantes de variadas lenguas origen y pertenecientes a los niveles A2 y B1, según MCRE, indican que los problemas más frecuentes son los de falsa selección de género gramatical en el nivel intermedio y la omisión de la tilde ortográfica en el nivel inicial. Además, Ferreira Cabrera *et al*. (2020) exploran la concordancia gramatical del verbo ser en estudiantes pertenecientes a dos subcorpus del CAELE: estudiantes anglosajones y francófonos. Los resultados evidencian que el primer grupo presenta mayores dificultades en la selección de categorías gramaticales, especialmente en la concordancia de género y número de los adjetivos, con respecto al grupo francófono.

Por otra parte, Campillo Llanos (2014), analiza los errores léxicos de hablantes no nativos a través de computadora en estudiantes pertenecientes a nivel intermedio-bajo para extraer frecuencias de uso de categorías y de unidades léxicas. Los resultados muestran que los

errores formales son más frecuentes en el nivel A2, pero persisten aumentando ligeramente en el nivel B1.

García Salido (2017), utiliza el corpus CEDEL 2 e indaga la relación existente entre la corrección de las colocaciones producidas por aprendices de español en textos escritos y la asociación que sus miembros presentan en un corpus representativo de esta lengua. Obtiene como resultado, que en los niveles iniciales no hay colocaciones infrecuentes pero en niveles avanzados podría ser razonable tratar colocaciones no frecuentes, pero con una información mutua alta, lo cual impactaría en la selección del léxico en programas de enseñanza de ELSE (García Salido 2017:42).

En la Universidad Nacional de Rosario, Argentina, se crea en 2004 el proyecto de investigación INFOSUR desde el Centro de Estudios de Adquisición del Lenguaje, radicado en la Facultad de Humanidades y Artes. El equipo de investigadores se propone confeccionar el módulo español Argentina (variante rioplatense) perteneciente al *Sistema NooJ*, diseñado por Max Silberztein en 2002 desde la Universidad de Framche- Comté para el análisis automático de lenguas naturales y de acceso abierto[23] (Bonino 2011; Solana *et al*. 2013; Tramallino). Este software cuenta con numerosas posibilidades para efectuar búsquedas (de estructuras, palabras, terminaciones) en grandes muestras textuales, además de efectuar análisis morfosintácticos y semánticos. Dentro del proyecto, surge una línea de investigación que propicia la generación de recursos didácticos para la enseñanza de la lengua española a partir del empleo de la mencionada herramienta informática. (Bonino & Rodrigo 2020; Tramallino & Arnal 2021).

Con el objeto de analizar automáticamente la IL de aprendientes de español, a partir del 2010, el grupo elabora un corpus de producciones escritas de estudiantes de español que pertenecen a instituciones educativas de la UNR. Para ello se establece una muestra aleatoria de aprendices extranjeros de español, cuyos textos son digitalizados. La muestra se encuentra dividida en dos niveles según el grado de aprendizaje de los sujetos: inicial e intermedio, de acuerdo al Marco Común Europeo de Referencia para las lenguas (MCERL). A su vez, cada corpus está dividido en grupos según la lengua de origen de los aprendices que son variadas; por un lado, románicas (portugués y francés) y por el otro, germánicas (alemán, holandés e inglés). Estos sujetos son estudiantes jóvenes, el promedio de edad es de 31 años y se encuentran habitando en Rosario. En el nivel inicial, estrato A1, se ubica el grupo de lengua

[23] Ver *Sistema NooJ* disponible en http://www.nooj-association.org y NooJ Manual disponible *on line* en www.nooj4nlp.net

portuguesa, francesa e inglesa, en total se compone de 95 sujetos. Los estudiantes del conjunto alemán pertenecen al nivel B1, mientras que los de holandés y portugués corresponden al B2 y reúnen producciones de 52 aprendices. Asimismo, cada grupo está compuesto por una cantidad determinada de textos que han sido numerados y corresponden a producciones de diferentes sujetos.

La investigación más reciente realizada con esas muestras (Tramallino *et al*. 2021) se ocupó de localizar y contabilizar los términos que poseían sufijos nominales a través de los diccionarios y archivos pertenecientes al sistema NooJ. Para ello se analizó automáticamente cada grupo de textos y se extrajeron conclusiones en cuanto al porcentaje de a) sufijos nominales coincidentes con el español estándar, (*-dad, -ción*), - b) sufijos inexistentes (*-tion, -dade, -iño*) c) sufijos nominales propios del español pero no combinables con las bases (*tipical, identitad, segurancia*). A continuación, se emplearon dos técnicas estadísticas para la medición de los datos: el Test No paramétrico de Wilcoxon para muestras independientes y la Prueba no paramétrica de Kruskal- Wallis. Los resultados expusieron que existen diferencias significativas respecto del uso y la distribución de sufijos nominales pertenecientes al español entre los corpus que no comparten un mismo nivel de instrucción, debido al carácter transitorio de la IL. Se determinó una mayor presencia de sufijos nominales en el nivel intermedio con respecto a lo que sucede en el nivel inicial. Esto se manifestó en las medias: 4,8 para el nivel A frente a 9,3 para el B. Es factible que el hecho de que los estudiantes empleen y produzcan nuevos términos se deba al conocimiento del léxico que va aumentando a medida que se avanza en el proceso de adquisición. Además, se observó un aumento pronunciado en la cantidad de sufijos coincidentes (SC) con la lengua meta en el paso de un nivel a otro. En respuesta a esto, dentro de los grupos del nivel B no se observan discrepancias con relación al holandés (10 %) y portugués (12 %) que pertenecen al estrato B2 pero sí las hay con relación al alemán que corresponde a un B1 y, por lo tanto, presenta una media inferior de sufijos coincidentes con el español (4,5 %). En cuanto a los grupos que integran el nivel inicial, el portugués es el que difiere del resto en tanto manifiesta no solo, una mayor cantidad de sufijos coincidentes sino también, un aumento de sufijos inexistentes y no combinables. Esto puede explicarse por la mayor proximidad de la lengua portuguesa con la española lo cual permite favorecer el proceso productivo en sujetos luso hablantes, pero también interferir negativamente en las primeras instancias de aprendizaje.

5. Consideraciones finales y discusiones

Este estudio tuvo por objeto realizar una panorámica de los desarrollos encuadrados en la lingüística computacional para el idioma español y asimismo, dar cuenta de herramientas informáticas que permiten realizar análisis morfosintácticos, semánticos y desambiguar oraciones, entre otras funciones. Puso de manifiesto que todavía no se ha alcanzado para el idioma español el mismo nivel de impulso global del que gozan el inglés y otras lenguas, en lo que respecta a las tecnologías informáticas, en coincidencia con Llisterri (2007). En la misma dirección, Camacho Caballero & Zevallos Salazar (2020) manifiestan que la complejidad y el alto costo que conlleva conformar muestras para el procesamiento natural del lenguaje, trae aparejado los escasos intentos por conformar corpus para las lenguas de países en vías de desarrollo. Estos autores sostienen que solo una pequeña fracción de las más de 6900 lenguas del mundo posee los recursos económicos suficientes para implementar las tecnologías del lenguaje humano. (Camacho Caballero & Zevallos Salazar 2020:189). Sin embargo, emprendimientos como Deepl Traductor y Google Translate son diseñados para traducir textos en cuantiosas lenguas, el último cuenta, actualmente, con traductores de 109 lenguas que son utilizadas por más de 75 millones de hablantes. Cabe hacer alusión al módulo quechua disponible en el sistema NooJ, de código abierto y acceso libre y al reciente trabajo colaborativo de investigadores argentinos, quienes han publicado el primer diccionario virtual wichí-castellano (DIWICA 2021), de acceso gratuito y construcción permanente.

En el ámbito de la adquisición de segundas lenguas, se hizo referencia a las distintas hipótesis de investigación; desde el análisis contrastivo, surgido a mediados del Siglo XX que dio paso al análisis de errores; hasta la corriente de IL. Este recorrido tuvo como foco las diferentes miradas acerca del concepto de "error" y las diversas acciones que suscitó la presencia de este en las interacciones de los estudiantes; desde los intentos para predecirlo, evitarlo y corregirlo hasta considerarlo un indicio positivo de las estrategias cognitivas activadas por los aprendientes. En los últimos años, los esfuerzos se concentran en generar emprendimientos computacionales, específicamente tutores inteligentes, como ELE Tutor, que se proponen imitar la dinámica de los docentes en las clases de español y otorgar un feedback correctivo.

Se evidenció, además, que, aunque existen numerosos criterios para clasificar y contabilizar esas formas idiosincrásicas, características de los diferentes estadios en el

aprendizaje del español como segunda lengua, aún hoy, se elige mayoritariamente investigar el aspecto descriptivo que involucra solo la perspectiva lingüística y hace referencia a los niveles gramaticales en los que se advierten estos indicios, sobre todo en lo que respecta a la morfología y la sintaxis. Como consecuencia, al no explorar el enfoque comunicativo de la lengua podríamos caer en el riesgo de creer que predomina una visión reducida en las prácticas de enseñanza y aprendizaje del español como segunda lengua, ligada a los aspectos más formales en detrimento de la dimensión socio-pragmática y/o cultural.

Asimismo, se revisitaron los corpus que se han elaborado para el estudio de la lengua española en uso, particularmente, los corpus de hablantes de español no nativos. Estas muestras representativas contribuyen al ámbito del aprendizaje de lenguas extranjeras y, especialmente, al de la enseñanza de español ELSE. En efecto, los resultados de las búsquedas y del análisis de esos textos pueden aplicarse en las clases y también, posibilitan basarse en ellos a la hora de diseñar material didáctico específico para grupos de estudiantes, considerando su nivel de aprendizaje y/ o sus lenguas maternas.

Con el propósito de dar a conocer los avances realizados en materia de análisis computacional de corpus de aprendices de español, se mencionaron las últimas investigaciones que describen y contabilizan diferentes fenómenos lingüísticos de la IL. Entre ellas, se expuso, brevemente, la labor realizada en las últimas dos décadas en la Universidad Nacional de Rosario, Argentina.

No obstante, a pesar de contar con varios emprendimientos para la lengua española, la mayoría de ellos provenientes de España, y algunos, de países de Latinoamérica, los esfuerzos siguen resultando escasos en comparación con desarrollos para lenguas provenientes de países con mayor potencial económico para financiar las investigaciones.

Por todo lo dicho anteriormente, este estudio pretendió no solo visibilizar los avances en lo que respecta a la conformación de corpus de aprendices y al tratamiento computacional de estas muestras de lenguaje, sino también incentivar a que se ejecuten nuevos proyectos en conjunto con centros de estudios europeos y americanos para revalorizar el idioma español y otorgarle el lugar central que amerita.

Referencias

Alexopoulou, Angélica (2005). Aproximación al tratamiento del error en la clase de E/LE desde la perspectiva del análisis de errores. *Estudios de Lingüística Aplicada*, 23(41).

Alexopoulou, Angélica (2010). La función de la interlengua en el aprendizaje de lenguas extranjeras. *Revista Nebrija de Lingüística aplicada*, 9.

Alonso, Laura *et al*. (2005). The Sensem Project: Syntactico-Semantic Annotation of Sentences in Spanish. Proceedings of the International Conference RANLP, 39-46. Borovets, Bulgaria.

Altenberg, Bengt (2011). "Preface". En F. Meunier *et al*. (eds.) *A Taste for Corpora. In honour of Sylvianne Granger*. Amsterdam: John Benjamins, 13-15.

Arispe, Agustín *et al*. (2020). Comprensión de los significados pragmáticos en hablantes de español de Argentina como lengua segunda y extranjera. *Quintú Quimün. Revista de lingüística* 4, 1-16.

Atserias, Jordi *et al*. (1997). Combining multiple methods for the automatic construction of multi-lingual WordNets. En Recent Advances in Natural Language Processing II. Selected papers from RANLP, volumen 97, página 327–338. Recuperado a partir de http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.97.5872 el 7 de septiembre de 2021.

Blanco Canales, Ana & Marta Nogueroles López (2013). Descripción Y Categorización De Errores Fónicos En Estudiantes De Español/L2. Validación De La Taxonomía De Errores AACFELE. *Logos: Revista De Lingüística, Filosofía Y Literatura*, *23*(2), 196–225.

Bolaños Cuéllar, Sergio (2015). La lingüística de corpus: perspectivas para la investigación lingüística contemporánea. *Forma y Función*, 28 (1), 31-54.

Bonino, Rodolfo (2011). Una propuesta para la implantación de la morfología verbal del español. *Revista INFOSUR*, (5), 79-86.

Bonino, Rodolfo (2009). Presentación de la lingüística computacional en Solana, Zulema (ed). *La interlengua de los aprendientes del español como L2,* Rosario, Centro de Estudios de Adquisición del Lenguaje, Facultad de Humanidades y Artes, UNR, pp. 7-18

Bonino, Rodolfo & Andrea Rodrigo (2020). El análisis automático de la sílaba escrita del español mediante la herramienta NooJ en Tramallino, Carolina (Ed.) *Homenaje a Zulema Solana: estudios de lingüística computacional, adquisición y enseñanza de lenguas*, pp. 64-80. Recuperado a partir de: http://hdl.handle.net/2133/19502

Bustos Gisbert, José M. (1998). Análisis de errores: problemas de categorización. Revista *DICENDA. Cuadernos de Filología Hispánica,* (16), 11-40.

Cabrera, Anita *et. al*. (2014). Análisis de errores asistido por computador basado en un Corpus de Aprendientes de Español como Lengua Extranjera. *Revista Signos*, 47(86), 385–411.

Camacho Caballero, Luis & Rodolfo Zevallos Salazar (2020). Lingüística computacional para la revitalización y el poliglotismo. *Letras (Lima)*, *91*(134), 184-198.

Camargo Angelucci, Thalita & María I. Pozzo (2020). Ang Errors and Mistakes in Foreign Language Learning: Drawing Boundaries from the Discourse of Argentine Teachers en E. Vanderheiden y C. H. Mayer (Eds.) *Mistakes, Errors and Failures across Cultures. Navigating Potentials*. pp. 383-398. Cham: Springer Nature. Recuperado de: https://doi.org/10.1007/978-3-030-35574-6

Campillos Llanos, Leonardo (2014). Errores léxicos en el español oral no nativo: análisis de la interlengua basado en corpus. *ELUA*, 0*(28), 85-124*.

Cembreros Castaños, Diana (2014). Lingüística computacional aplicada a la investigación educativa: un enfoque matemático de la enseñanza de vocabulario en lengua inglesa para hispanohablantes. Tesis Doctoral UAM (Universidad Autónoma de Madrid). Recuperado a partir de <http://hdl.handle.net/10486/660767> el 13 de septiembre de 2021.

Cesto Mancera, Ana *et al*. (2001). Corpus para el análisis de errores de aprendices de E/LE (CORANE). En *Actas del XII Congreso Internacional de ASELE: tecnologías de la información y de las comunicaciones en la enseñanza de la E/LE*, Valencia, 2001, 527-534.

Corder, Pit (1971). Idiosincratic Dialects and Error Analysis. *International Review of applied Linguistic* 9, 149-159.

[DIWICA] *Wichi-siwelelhayhilh / Diccionario wichí-castellano*. Buenos Aires: INILSyT, Universidad Nacional de Formosa & IFLH, Universidad de Buenos Aires & DILA, CAICyT-CONICET, 2021. Recuperado de:<www.diccionariowichi.com.ar> el 7 de septiembre de 2021.

Fellbaum, Christiane (1998). WordNet: Anelectronic lexical database. The MIT press.

Fernández López, Sonsoles (1995). Errores e interlengua en el aprendizaje del español como Lengua extranjera. *Didáctica*, 7, 203-216.

Fernández López, Sonsoles (2000). Corrección de errores en la expresión oral. *Carabela* 47: 133-150.

Ferreira Cabrera, Anita *et al*. (2012). La Arquitectura de ELE-TUTOR: Un Sistema Tutorial Inteligente para el Español como Lengua Extranjera. *Revista signos*, *45*(79), 102-131.

Ferreira Cabrera, Anita *et al.* (2014). Análisis de errores asistido por computador basado en un corpus de aprendientes de español como lengua extranjera. *Revista signos*, Valparaíso, v. 47 n. 86, p. 385-411, 2014.

Ferreira Cabrera, A. & Jéssica Elejalde Gómez (2017). Análisis de errores recurrentes en el Corpus de Aprendices de Español como Lengua Extranjera (CAELE). *Revista Brasileira de Lingüística Aplicada* (RBLA) 17 (3), 509-537.

Ferreira Cabrera, Anita *et al*. (2020). Análisis de Errores en el Corpus CAELE: estudio sobre la concordancia gramatical en el verbo SER en aprendientes francófonos y anglosajones. *Revista Nebrija De Lingüística Aplicada a La Enseñanza De Lenguas*, *14* (29), 76 - 99.

García Salido, Marcos (2017). Frecuencia y corrección colocacional en la producción escrita de aprendices de español *ONOMÁZEIN* 38, 22 – 46.

Granger, Sylviane (2002). A Bird's-eye view of learner corpus research. In Granger, S., Hung, J., &Petch-Tyson, S. (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* Philadelphia: John Benjamins Publishing Company, 3–33.

Hincapié, Diana (2018). Corpus de Aprendientes de Español como Lengua Extranjera y Segunda Lengua (CAELE/2): el componente escrito. *Forma y Función*, 31(2), 129-143.

Kotz Grabole, Gabriela & Anita Ferreira Cabrera. (2013). La precisión gramatical mediada por la tecnología: el análisis y tratamiento automático de errores. *Literatura y lingüística*, (27), 219-242.

Lavid, Julia (2005) *Lenguaje y nuevas tecnologías. Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI*. Cátedra. Madrid.

Llisterri, Joaquín (2007). El español y las nuevas tecnologías. En M. Lacorte (Ed.), *Lingüística aplicada del español.* Madrid: Arco/Libros, pp.483-520.

Lozano, Cristóbal (2009). CEDEL2: Corpus Escrito del Español L2. En Bretones Callejas, Carmen M. et al. (eds) *Applied Linguistics Now: Understanding Language and Mind.* Almería: Universidad de Almería, 197-212.

MacWhinney, Brian, & Johannes Wagner (2010). Transcribing, searching and data sharing: The CLAN software and the Talk Bank data repository. *Gesprächsforschung, 11*, 154-173. Recuperado a partir de <https://psyling.talkbank.org/years/2010/macw-wagner.pdf> el 8 de septiembre de 2021.

MacWhinney, Brian & Catherine Snow (1985). The Child Language Data Exchange System. Journalof Child Language, 12, 271-295. Recuperado a partir de https://psyling.talkbank.org/years/1985/jcl-childes.pd el 2 de septiembre de 2021.

Miller, George A.*et al.* (1993). A semantic concordance. En Proceedings of the workshop on Human Language Technology, HLT '93, página 303–308, Stroudsburg, PA, USA. Association for Computational Linguistics. ACM ID: 1075742.

Moreno Sandoval, Antonio (1998). *La lingüística computacional. Introducción a los modelos simbólicos, estadísticos y biológicos.* Madrid: Editorial Síntesis.

Muñoz Liceras, Juana (2009). La interlengua del español en el siglo XXI. *Revista Nebrija de Lingüística Aplicada* 5, 36-49.

Nemser, William (1971). Aproximative systems of foreign language learners. International Review of Applied Linguistics 9 (2):115-123. (Traducción al español: "Los sistemas aproximados de los que aprenden lenguas segundas" en J. Muñoz Liceras (comp.) (1991): *La adquisición de las lenguas extranjeras.* Madrid: Visor: 51-61.

Oliver González Antoni & Salvador Climent (2011). Construcción de los WordNets 3.0 para castellano y catalán mediante traducción automática de corpus anotados semánticamente. *Revista Procesamiento del Lenguaje Natural*, (47), 293-300.

Padró, Lluis (2011). Analizadores Multilingües en FreeLing. *Linguamatica* (2), 13–20. Recuperado a partir de: <http://linguamatica.com/index.php/linguamatica/article/view/115> el 11 de septiembre de 2021.

Palacios Martínez, Ignacio *et al*. (2019). El Corpus de Aprendices de Español (CAES) y sus aplicaciones para la enseñanza/aprendizaje del español como lengua extranjera en Blanco, Marta, Hella Olbertz y Victoria Vázquez Rozas (eds.): *Corpus y construcciones. Perspectivas hispánicas.* Anejo 79 de Verba, 2019, 273-303.

Parodi, Giovanni (2004). Textos de especialidad y comunidades discursivas técnico-profesionales: una aproximación basada en corpus computarizado**.** *Revista Estudios filológicos*, (39), 7-36.

Parodi, Giovanni (2008). Lingüística de corpus: una introducción al ámbito. *RLA. Revista de lingüística teórica y aplicada*, *46* (1), 93-119.

Payrató, Luis (2011) *De profesión, lingüista. Panorama de la lingüística aplicada*. Barcelona: Ariel Letras.

Pino Rodríguez, Aymé (2009). Palabras en interacción: un corpus de aprendices suecos de E/LE. A survey of corpus-based research, 2009. p. 470-487. Recuperado a partir de <https://goo.gl/XAHPR3> el 5 de abril de 2020.

Rojo, Guillermo (2002). "Sobre la lingüística basada en el análisis de corpus". Ponencia plenaria en el XV Congreso de la ALFAL (Montevideo, 18-21de agosto de 2008). Recuperado a partir de <https://gramatica.usc.es/~grojo/Publicaciones/Lgca_corpus_lgca_espanol.pdf> el 8 de junio de 2020.

Santana Suárez, *et al.* (2009).Functional Disambiguation Using the Syntactic Structures Algorithm for each Functional Interpretation for Spanish Language. *Lecture Notes in Computer Science. Theoretical Computer Science and General Issues.* Springer. v. 5717, 1611-3349.

Santos Gargallo, Isabel (2004). El análisis de los errores en la interlengua del hablante no nativo, en J. Sánchez Lobato e I. Santos Gargallo (eds.) (2004): *Vademécum para la formación de profesores. Enseñar español como segunda lengua (L2) / lengua extranjera (LE)*. Madrid: SGEL, 391-411.

Santos Gargallos, Isabel & Angélica Alexopolou (2021). Metaanálisis de las tesis doctorales de análisis de errores en la interlengua española a los largo de tres décadas (1991-2019). *Marco ELE. Revista de Didáctica Español Lengua Extranjera,* núm. 32, 2021. Recuperado de <https://www.redalyc.org/journal/921/92165031009/html/> el 14 de junio de 2020.

Selinker, Larry (1972). Interlanguage. *International Review of Applied Linguistics*, 10/3, 209-231, 1972.

Solana, Zulema *et al.* (2013). Análisis automático morfológico con las herramientas SMORPH Y NOOJ. *Revista de epistemología y ciencias humanas* 5, 230-256.

Torijano, Agustín (2008). El estudio de los determinantes en aprendices luso hablantes de español, DICENDA. *Cuadernos de Filología Hispánica* 26, 235-257.

Torruella, Joan & Joaquim Llisterri (1999). Diseño de corpus textuales y orales. En: Blecua J. M., Clavería G. Sáncez C. Toruella, J. (editores). *Filología e informática*, Univ. Autónoma de Barcelona, Editorial Milenio, 45-77.

Tramallino, Carolina (2013). Análisis morfológico con herramientas informáticas. Reconocimientos de nombres en textos de español con el sistema NooJ. *Revista Lingüística y Literatura,* (63), 33-48.

Tramallino, Carolina & Romina P. Arnal (2021). Reconocimiento de sintagmas nominales construidos con indefinidos a través del sistema NooJ en corpus de español como segunda lengua. *Revista IRICE*, (38), 129-162. Recuperado a partir de: https://ojs.rosario-conicet.gov.ar/index.php/revistairice/article/view/1310

Tramallino, Carolina; Beltrán, Celina & Natalia Ricciardi (2021). Localización y contabilización de sufijos nominales en corpus de aprendientes de español como segunda lengua. *Entrepalavras, 11* (10esp), 412-436. doi: http://dx.doi.org/10.22168/2237-6321-10esp2116